

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

Data Mining for Molecular Karyotyping: Linked Analysis of Array-CGH Data and Biomedical Text

Promotoren:

Prof. dr. ir. B. De Moor
Prof. dr. ir. Y. Moreau

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Steven VAN VOOREN



KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

Data Mining for Molecular Karyotyping: Linked Analysis of Array-CGH Data and Biomedical Text

Jury:

Prof. dr. ir. Etienne Aernoudt, voorzitter
Prof. dr. ir. B. De Moor, promotor
Prof. dr. ir. Y. Moreau, promotor
Prof. dr. ir. J. Vermeesch
Prof. dr. ir. S. Van Huffel
Prof. Dr. K. Devriendt
Prof. Dr. F. Speleman
Dr. D. Rebholz - Schuhmann

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Steven VAN VOOREN

ISBN 978-94-6018-112-2

© Katholieke Universiteit Leuven – Faculteit Ingenieurswetenschappen
Arenbergkasteel, Kasteelpark Arenberg 10, B-3001 Heverlee (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

D/2009/7515/95

ISBN 978-94-6018-112-2

Voorwoord

Vele mensen hebben direct of indirect bijgedragen tot het tot stand komen van dit werk, dat bij uitstek het resultaat is van verschillende samenwerkingen. Mensen zowel binnen en buiten het departement Electrotechniek van de faculteit Ingenieurswetenschappen, binnen en buiten onze universiteit, en binnen en buiten Vlaanderen hebben ertoe bijgedragen, en ik dank hen van harte.

Enkele mensen wil ik expliciet vermelden. Vooreerst wil ik mijn promotoren Prof. dr. ir. Yves Moreau en Prof. dr. ir. Bart De Moor bedanken om me de gelegenheid te geven de overstap te maken naar wetenschappelijk onderzoek, en me de kans te geven aan de slag te gaan in een stimulerende en professionele onderzoeksomgeving. Bart weet als geen ander zijn studenten en researchers enthousiast te maken terwijl hij hen tegelijk de vrijheid geeft die noodzakelijk is in wetenschappelijk onderzoek, en opent deuren naar projecten en samenwerking voor zijn onderzoekers binnen de wetenschap en daarbuiten. Yves legt zijn hart en ziel maar ook een indrukwekkend pak visie in de bioinformaticagroep, waardoor we internationaal op de kaart staan. Hem dank ik verder ook hartelijk voor de wetenschappelijke discussies, de nodige inzichten in statistiek, en zijn enthousiast promotorschap.

Ik heb het geluk gehad ingebed te kunnen zijn in een onderzoeksgroep waar mensen met heel diverse achtergronden en kennisdomeinen aan de slag zijn. Zo is er op ESAT steeds een expert in de buurt die kan helpen wanneer je vastloopt in een materie waar je nog niet in thuis bent. De afgelopen jaren heb dan ook ik met veel plezier samengewerkt met Bert, Frizo, Dries en Shi Yu rond onze infrastructuur voor tekstontginning. Joke, Olivier en Peter dank ik voor hun statistisch inzicht, Kristiaan en Tijl voor hun kernelkennis, Stein voor zijn hypergedrevenheid en onze samenwerking rond *cis*-regulatoire modules, Janick voor de biologische validaties, en Patrick Glenisson voor zijn mentorschap tijdens mijn afstudeerwerk en de eerste stappen van mijn onderzoek. Verder verdient vriend en vijand Bert Coessens een expliciete vermelding. Het is een plezier om met jou aan hetzelde zeel te trekken in onderzoek, het maken van publicaties, in softwareontwikkeling, en het samen laten ontkiemen van onze spin-off. Verder dank ik Tom, Raf, Bert P, Thomas, Karen, Nathalie, Cynthia, Gert, Frank, Ruth, Kathleen, Kristof, Geert, Gert, Pieter, Anneleen, Frank, Niels, Ernesto, Sonia, Daniela, Lieven, Wout, Tim, Liesbeth, Tunde en alle anderen voor een fijne samenwerking op ESAT.

Roland et Léo, je vous remercie pour tout qu'on a fait ensemble sur les Wikis,

la prioritisation des gènes, l'intégration des bases de données genetiques, et les sessions de formation qu'on a donné ensemble.

Vervolgens verdient Prof. dr. ir. Joris Vermeesch absoluut een eigen paragraaf in dit dankwoord. Hem wil ik bedanken voor de waardevolle samenwerking die mede dankzij zijn visie tot stand is kunnen komen tussen de bioinformatici van ESAT en zijn groep binnen het CME, voor de onderdompeling die ik vanaf het begin heb gekregen in de wereld van de Cytogenetica en de interessante onderzoeksvraagstukken die daaruit zijn voortgekomen, voor de kansen die hij heeft gegeven om mijn werk te tonen aan andere onderzoekers en laboratoria via zijn contacten en door co-publicaties, en voor de waardevolle kennis binnen de klinische genetica en daarbuiten die hij en de onderzoekers in zijn groep me de afgelopen jaren hebben bijgebracht. Ik heb veel opgestoken tijdens de *group meetings* waar iedereen wekelijks zijn of haar lopend onderzoek voorstelt, en ik heb er met plezier aan deelgenomen. Joris, bedankt.

De algorithmen en methoden die door bioinformatici uitgewerkt worden, komen pas tot leven wanneer ze toegepast worden op reële biomedische problemen en data uit echte lab-experimenten. Daarom kan de waarde van de samenwerking met het Centrum Menselijke erfelijkheid niet onderschat worden. Ik bedank Nicole voor het aanreiken van een eerste dataset, Irina voor de waardevolle discussies rond congenitale oogafwijkingen, Bernard en Koen voor de vele suggesties voor verdere uitdieping en nieuwe toepassingen van onze algoritmen, Cindy, Natalie, Reinhilde en Anneleen voor het grondig toetsen van de toepassingen aan de dagelijkse routine en diagnostiek, Guy voor de samenwerking rond primerdatabanken, Julio voor de opening naar toepassingen voor oncologie, en Jeroen, Femke, Evelyne, Hilde, Caroline, Ilse, Sigrun, Elyes, Thierry, An en Ivo voor de vele nuttige commentaren, discussies, concrete vragen en analyseproblemen rond Bench, LOOP, Clone Wars en tal van andere tools. Een speciale vermelding gaat uit naar Thomy. Jouw niet aflatend enthousiasme, jouw zin voor detail, en je doordachte commentaren waarmee we ons werk hebben kunnen afstemmen op de klinische work flow waren van onschatbare waarde, en zijn dat nog steeds.

Verder dank ik Ida, Péla, Ilse P., het departementaal secretariaat en Rita voor de administratieve ondersteuning. Ida, het was met veel plezier dat ik mijn informatietechnologische knobbel ter beschikking kon stellen in ruil voor de vruchten van je culinair talent. Ook Edwin en Maarten ben ik veel verschuldigd voor de snelle en kwalitatieve technische ondersteuning.

Ook enkele mensen van de Universiteit Gent, mijn oorspronkelijke Alma Mater, dank ik voor de goede samenwerking, de waardevolle discussies, en de gezamenlijke publicaties. Hier verdienen dr. Björn Menten en Prof. Dr. Frank Speleman, die bovendien bereid was op te treden als lid van mijn doctoraatsjury, een speciale vermelding.

A very warm thank you goes to Dr. Nigel Carter, head of the Wellcome Trust Sanger Institute Cytogenetics Team, who gave me the opportunity to put my scientific work into the international context of DECIPHER, a world wide patient case consortium effort to advance medical and scientific knowledge about submicroscopic chromosomal imbalances and their relation to human develop-

ment and health. The couple of months I've spent at the Sanger institute have been inspiring. I'd like to thank Alan, Heike, Jodie, Paul, Shola and Manuel at Sanger and Helen Firth at the Addenbrooks hospital for the fruitful and currently still ongoing collaboration. Special thanks go to Roger Pettett for the mentorship, the PERL-fu, the friendly banter, the geeky discussions, and the warm welcome in his home. Give my warmest regards to Josie, Michael, Francis and Whiskey.

Equally, a warm thank you goes to dr. Dietrich Rebholz-Schuhmann, for taking me up in his text mining group at the European Bioinformatics Institute while I was at the Hinxton genome campus, and for agreeing to be on my PhD jury. The work of your group is renowned in the field of biomedical text mining and was instrumental in linking my research to the DECIPHER initiative. The fact that your group is publishing its research results not just to the scientific community but also as publicly available and well supported web services and resources is visionary. I'm indebted to Miguel, Pjotr, and Sylvain, who are smart and hard-working members of your research group.

Verder bedank ik *friends, family and fools* voor hun steun en voor het vakmanschap dat ze aan de dag legden bij het nalezen van dit proefschrift. Bedankt aan mijn ouders, mijn broer Bart, en bedankt aan de andere proofreaders: Frederik (Kezel), Frizo, Gunther, Feike, Boris – die ik bovendien bedank voor de directe communicatielijn naar zijn brein waarlangs ik steeds zijn diepgaand technisch Vernu(n)ft kon aftappen, voor de met Rotwein overgoten discussies over architectuur (Spring Framework of Tadao Ando?), en voor de vele interacties rond software design en ontwerpstijl – en bovenal bedank ik Ilse, niet alleen als biomedisch klankbord, maar voor haar geduld en haar steun, voor haar bezorgdheid, voor haar geduld (nogmaals), en voor haar mooie zelf. Ilse, ik hou je graag.

Verder ben ik tal van mensen dankbaar voor hun rol in het spin-offavontuur dat we samen hebben opgestart rond dit werk. Vooreerst Yves en Bart, die over de jaren heen binnen het departement de omgeving hebben gecreëerd waarbinnen zo'n project mogelijk wordt, en Joris voor zijn inzicht in de wereld van de cytogenetische diagnostiek en research, zijn sturing en zijn waardevol netwerk. Dank ook aan Peter, Wim, Rudi, Paul en de andere mensen van LRD die met professionele raad en daad hebben begeleid naar de opstart. Dank aan Herman voor zijn mentorschap en onschatbare meerwaarde langs de zakelijke kant van het verhaal, en Kenneth, Tom en Johan voor hun geloof in ons project.

Tenslotte wens ik ook de voorzitter van mijn doctoraatsjury, Prof. dr. ir. Aernoudt, en de assessoren en leden van het begeleidings- en leescomité, Prof. dr. Bruynooghe, Prof. dr. ir. Van Huffel, Prof. Dr. Devriendt, Prof. Dr. Speleman en Dr. Rebholz-Schuhmann te bedanken voor het advies dat ze mij gegeven hebben tijdens mijn doctoraat en de waardevolle feedback tijdens het opstellen van de tekst die nu voor u ligt.

Allen, bedankt!

Abstract

Defects in the genetic make up of a cell or organism can lead to developmental problems and hereditary disease. In order to be able to discover these aberrations in the genome of patients, certain techniques are used in both clinical routine diagnostics as in genetic research studies to chart the genetic material. Screening of patient tissue by investigating the genetic material in a cell at a molecular level is called Molecular Karyotyping. Recent evolutions in this field allow a significant increase in resolution when determining the chromosomal characteristic of a tissue. In this respect, the determination of deletions and duplications, among other genomic aberrations, happens through a technology called Array CGH (Comparative Genomic Hybridization), in which representative genomic fragments are used as reporters and are spotted on small glass slides in microarrays.

The growing body of array and patient related data bring specific challenges with respect to the identification and pin-pointing of genomic regions linked to inherited or acquired aberrations, and to the automated analysis of all patient related data. Important challenges have emerged regarding the interpretation of the results of this genetic screening technique.

In the context of this doctoral dissertation, one of the main goals has been to find ways to support and enable this interpretation. On the one hand, methods were sought to handle the raw, untreated data that come from array CGH experiments. Where earlier genetic screening methods relied on interpretation through visual inspection, with array CGH, this is no longer possible, and the need of means for analysis and interpretation of large quantities of numeric data has become obvious. On the other hand, methods have been developed – through approaches such as automated text analysis on publicly available biomedical literature and on patient data, among other things – to link phenotype traits to genomic regions, and to give clinical meaning to the results from array CGH assays. Finally, the combination of these methods has lead to a platform for storage, analysis, and management of all data related to patients for whom a genetic screening through Array CGH has been performed, including not only raw data but phenotype descriptions and clinical information as well. This platform supports the interpretation of genetic, clinical and phenotype information and allows for genome annotation in both a diagnostic and a research context.

Korte inhoud

Defecten in het genetische materiaal van een cel of organisme leiden tot problemen bij de ontwikkeling en tot erfelijke aandoeningen. Om deze aandoeningen bij een patiënt in kaart te brengen, wordt zowel in de klinische routine voor diagnose als in genetische onderzoeksstudies gebruik gemaakt van technieken om het genetisch materiaal in kaart te brengen. Het screenen van patiëntweefsel door het onderzoeken van genetisch materiaal in een cel heet Moleculaire Karyotypering. Recente evoluties in dit veld laten toe om de resolutie bij het bepalen van de chromosomale karakteristieken van een weefsel aanzienlijk te verfijnen. Zo gebeurt het bepalen van onder meer deleties en duplicaties aan de hand van rooster-CGH (Comparative Genomic Hybridization), een techniek waarbij representatieve genomische fragmenten op kleine glasplaatjes in microroosters worden aangebracht.

De groeiende hoeveelheid rooster- en patiëntengegevens houdt uitdagingen in op het vlak van identificatie en lokalisatie van genoomregio's gelinkt aan aangeboren afwijkingen, en het automatisch analyseren van patiëntgerelateerde data. Vooral op het vlak van interpreteren van de resultaten uit deze genetische screeningtechniek zijn belangrijke uitdagingen ontstaan.

In het kader van dit doctoraat is dan ook gewerkt aan manieren om deze interpretatie mogelijk te maken. Enerzijds werden methoden gezocht om de ruwe data uit rooster-CGH te verwerken. Waar vroegere screeningmethodes berustten op interpretatie met het blote oog, is rooster-CGH gebaseerd op de analyse en interpretatie van grote hoeveelheden numerieke data. Anderzijds zijn methoden ontwikkeld om, onder meer via tekstontginning op publiek beschikbare biomedische literatuur en op patiëntengegevens, fenotypische kenmerken aan genoomregio's te koppelen, en om betekenis te geven aan de analyseresultaten in een klinisch-diagnostische context. Ten slotte heeft de combinatie van de beschreven methoden geleid tot een platform voor het verwerken en beheren van patiëntengegevens waarbij genetische screening door middel van rooster-CGH is toegepast, en waarin ook gegevens over het fenotype van de patiënt beschikbaar zijn. Dit platform ondersteunt de interpretatie van genetische informatie en laat genoomannotatie toe, in zowel een diagnostische als onderzoekscontext.

Nederlandstalige samenvatting

Situering van het onderzoek

Fouten in het genoom van een mens veroorzaken ontwikkelingsstoornissen en aangeboren genetische aandoeningen. Zo geeft bijvoorbeeld een extra kopij van chromosoom 21 aanleiding tot het syndroom van Down. Andere wijzigingen aan het genetisch materiaal zoals duplicaties en deleties van chromosoomfragmenten, of mutaties en verschuivingen van letters in de genetische code op specifieke plaatsen, bepalen niet alleen de lichamelijke verschillen tussen individuen, maar kunnen eveneens aanleiding geven tot ziekte.

Verschillende methoden zijn ontwikkeld om het genetisch materiaal van een patiënt in kaart te brengen. Door middel van een genetische screen kan men chromosomale aberraties en breekpunten in kaart brengen. Een gekende manier om dit te doen is door middel van G-bandkleuring, waarbij genetisch materiaal gekleurd wordt zodanig dat aberraties op het genoom onder de microscoop zichtbaar worden. Deze techniek is echter beperkt in resolutie.

Een tweede techniek, Fluorescent In Situ Hybridization, werd vanaf de jaren 1980 gebruikt om de aanwezigheid en situering van heel precieze fragmenten na te gaan. Hiermee brak het tijdperk van de moleculaire cytogenetica aan. Nadeel van de techniek was en blijft dat FISH niet kan worden gebruikt om het volledige genoom in één keer in kaart te brengen. Bijgevolg zijn steeds eerst aanwijzingen nodig, zoals een ziektebeeld of andere genetische screen, vooraleer men met FISH gericht op zoek kan gaan naar genetische aberraties in het weefsel van een patiënt.

Een recente doorbraak heeft er echter toe geleid dat het mogelijk werd om het volledige genoom van een patiënt in een enkel experiment na te gaan op kopijvariaties. Dank zij Array CGH heeft het ontdekken van nieuwe genetische syndromen een sprong voorwaarts genomen, en zijn gekende syndromen veel beter in kaart gebracht kunnen worden. Bij deze roostergebaseerde comparatieve genomische hybridizatie (Array CGH) steunt men op op de beschikbaarheid van verzamelingen van fragmenten van genomisch materiaal, kloonfragmenten genaamd, die beschikbaar zijn in de vorm van bibliotheken en waarvan de sequentie en dus fysische locatie ten opzichte van een referentiegenoom gekend

zijn. Dit is mogelijk dankzij het feit dat rond de milleniumwisseling de effectieve sequentie van het genoom van de mens volledig in kaart gebracht is.

Het klinische belang van deze nieuwe technologie is groot. Dankzij Array CGH is het wetenschappelijke veld van genetisch onderzoek alsook de genetische diagnostiek in een stroomversnelling gekomen. Array CGH wordt gebruikt voor genoomwijde analyse in de zoektocht naar aneuploidieën en segmentele aneusomieën. Naast het in kaart brengen van aberraties bij specifieke patiënen, is Array CGH succesvol gebleken om de link tussen ziekten en genen verder in kaart te brengen en bij te leren over de processen die aan de basis liggen van genetische aandoeningen. Verder heeft de techniek ertoe geleid dat onderzoekers hebben ontdekt dat mensen onderling veel feller van elkaar verschillen op genetisch vlak dan aanvankelijk gedacht.

Bij Array CGH wordt genomisch DNA van een patiën gemerkt met een fluorescente kleurstof (bijvoorbeeld rood). Een normaal referentiestaal wordt gemerkt in groen, en beide stalen worden tergelijkertijd gehybridizeerd op een microrooster, waarop een zeer groot aantal gekende genomische doelfragmenten aanwezig is. Met een laserscanner kan vervolgens worden uitgelezen op welke posities van het genoom, de patiën afwijkt van het normale referentiestaal. Deze techniek is intussen gecommercialiseerd door verschillende bedrijven en wordt als pasklare kit aangeboden. De resolutie waarin deze assay kan gebeuren, stijgt met de tijd.

Het belang van deze techniek in de klinische diagnose is niet te onderschatten. Voor dezelfde prijs wordt aan veel hogere resolutie en veel sneller dan bij klassieke experimenten, een genoom in kaart gebracht en onderzocht op kopijvariaties. Als resultaat daarvan worden dubbel zoveel afwijkingen ontdekt als voorheen. Ook hoeft men niet op zoek naar aanwijzingen om op voorhand te bepalen in welke regio van het genoom men bij de patiën moet gaan zoeken: de test bemonstert het volledige genoom.

0.1 Impact van Array-CGH als genetische screeningtechniek

Array CGH leidt nu reeds tot een dieper begrip van de complexiteit van het humaan genoom. Het is mogelijk steeds kleinere afwijkingen in kaart te brengen, zowel in onderzoekscontext als in de kliniek. Het succes van de techniek heeft echter geleid tot een enorme groei aan genetische informatie. In die mate zelfs, dat het op dit ogenblik de analysemogelijkheden en het goede begrip van de betekenis ervan in een diagnostische context overstijgt. Men leert bij over de architectuur van het genoom van de mens, over bestaande syndromen, men ontdekt nieuwe linken tussen genen en ziekten, en men zet de eerste stappen in het in kaart brengen van complexe en multifactoriële genetische aandoeningen.

De enorme groei aan genetische informatie, de nieuwe kennis rond de manier waarop elk van ons verschillen vertoont in kopijaantallen zonder daarbij noodzakelijk een ziektebeeld te vertonen, en de impact van de grote hoeveelheden

numerieke data die voortkomen uit dit soort genetische screening, met alle uitdagingen rond interpretatie, beheer en visualisatie die hiermee gepaard gaan, vormen de grondslag van dit doctoraat.

Deze dramatische nieuwe evoluties leiden namelijk tot de nood aan specifieke benaderingen om deze gegevens te analyseren en interpretatie ervan mogelijk te maken. Hierbij gaat het niet enkel om de ruwe data die voortkomen uit Array CGH experimenten, maar ook om alle metadata daarrond, zoals klinische en fenotypegegevens van de patiën, experimentontwerp, en alle genetische kennis die reeds beschikbaar is, bijvoorbeeld in publieke databanken en biomedische literatuur. Dit werk beschrijft hoe er in de context van een interdisciplinaire samenwerking tussen verschillende onderzoeksgroepen binnen en buiten de K.U.Leuven gewerkt is aan het gemeenschappelijk doel om de beschreven uitdagingen het hoofd te bieden aan de vooravond van een nieuw tijdperk in de genetische geneeskunde.

0.2 Opbouw van dit proefschrift

Nadat we in Hoofdstuk 1 een meer gedetailleerde inleiding hebben doorlopen op de biologische, klinische en technische context waarin dit proefschrift tot stand kwam, wordt in Hoofdstuk 2 gesproken over het beheer en de verwerking van hoofdzakelijk ruwe data die voortkomen uit het uitvoeren van Array CGH experimenten. Eerst en vooral wordt besproken hoe in de beginfase van de introductie van de technologie vanuit het onderzoek naar de klinische praktijk, kwaliteitscriteria zijn opgesteld om de techniek op een robuuste manier voor diagnostisch gebruik te kunnen verantwoorden. Hierbij wordt gesproken over aspecten in verwerking en rudimentaire interpretatie van genetische data. Verder wordt in dit hoofdstuk ArrayCGHBase besproken, in het kader van databanken die analyse van ruwe roostergegevens toelaten. Vervolgens wordt ingegaan op een alternatief experimenteel ontwerp, waarbij niet langer wordt gewerkt met een referentiestaal, maar waarbij patiënenstalen ten opzichte van elkaar worden vergeleken. We bespreken statistische modellen die de gegevensverwerking die hierop volgt mogelijk maken, en kaderen ook de informatietechnologie hierrond. Verder wordt getoond hoe uit het aggregeren van gegevens over duizenden experimenten heen, voordeel kan worden gehaald voor het in kaart brengen van variatie binnen het menselijk genoom.

In Hoofdstuk 3 gaan we dieper in op de functionele interpretatie van genetische informatie. Waar we eerst spraken over ruwe data-analyse, kwaliteitscriteria en experimentontwerp, gaan we een stap dieper in de richting van interpretatie van de gegevens. Om betekenis te kunnen geven aan de grote hoeveelheid informatie gekoppeld aan een patiën en voortkomend uit het verzamelen van gegevens van verschillende experimenten, zijn verschillende tools ontwikkeld, sommige binnen onze onderzoeksgroep, andere daarbuiten. Er wordt een overzicht gegeven van de state of the art in tools en databanken die genomische loci in hun functionele context kunnen plaatsen. Het doel is ten eerste de publiek beschikbare tools te overlopen die toelaten om die kandidaatgenen te

identificeren die een rol spelen in nieuwe syndromen gekenmerkt door submicroscopische chromosomale onevenwichten, en vervolgens enkele van deze tools in detail te bespreken. Meer concreet bekijken we in dit hoofdstuk aBandApart, waarbij biomedische literatuur wordt gebruikt om genomische bandpatronen zoals die zichtbaar zijn bij klassieke karyotypering, te koppelen aan fenotypische kenmerken en ziekten. Verder bespreken we aGeneApart, een gelijkaardig platform dat genen en fenotypes koppelt op basis van een statistisch model dat wordt toegepast op cocitatiegegevens uit literatuur. Ten slotte bekijken we Endeavour, een vlaggenschip van de bioinformaticagroep dat heel wat databronnen gebruikt en door middel van een datafusie-benadering erin slaagt om kandidaatgenen te prioritizeren. Bij het bespreken van Endeavour wordt enkel de nadruk gelegd op de bijdrage in het kader van dit doctoraatswerk, namelijk het gebruik van literatuur als bron van informatie voor genprioritizatie.

Hoofdstuk 4 ligt in het verlengde hiervan, maar heeft een focus op klinische interpretatie en ondersteuning van diagnose. We kijken naar het collaboratieve aspect van biomedische research en diagnostiek, en het databankaspect. We gaan nu vooral op zoek naar de manier waarop tools uit het vorige hoofdstuk, samen met intelligente databanken waarin alle mogelijke patiënengegevens worden opgeladen en bijgehouden, kunnen worden gebruikt binnen een omvattend platform dat voldoet aan de specifieke uitdagingen waarmee diagnostiek en onderzoek in de constitutionele genetica worden geconfronteerd. We bespreken in dit hoofdstuk twee dergelijke platformen. Een eerste is Store+Bench, een platform voor onderzoek, diagnose en genoomannotatie. Vervolgens behandelen we DECIPHER, een gelijkaardig platform dat meer is gericht op onderzoek en dat als doel heeft de leden van een gesloten internationaal consortium toe te laten patiënengegevens uit te wisselen om zo zeldzame syndromen te ontdekken en in kaart te brengen. Deze platformen hebben de annotatie van het humaan genoom met nieuw verworven kennis tot doel. Het concept genoomannotatie wordt dieper onderzocht in het volgende hoofdstuk.

Hiertoe introduceren we eerst enkele technische aspecten van genoomannotatie in Hoofdstuk 5. In dit practische gedeelte bekijken we wat genoomannotatie is, en hoe we diverse en steeds opnieuw geactualiseerde genetische kennis op gedistribueerde wijze kunnen beschikbaar stellen aan de onderzoeksgemeenschap, op een gestandaardiseerde manier, door het gebruik van het Distributed Annotation Server protocol. We illustreren bovendien hoe dit wordt gebruikt in de context van wat in vorige hoofdstukken werd besproken: Bench zowel als DECIPHER ondersteunen op deze manier visualisatie en data-integratie in genoombrowsers. Ook de visualisatie van primers komt aan bod.

In een tweede luik zullen we genoomannotatie niet enkel vanuit puur technisch standpunt bekijken, maar gaan we dieper in op een casus waarbij effectief genoomannotatiegegevens worden gecreëerd, en dit in de context van biomedisch basisonderzoek naar transcriptionele regulatie-elementen. In dit deel tonen we hoe transcriptionele regulatie wordt ontrafeld door een combinatie van manuele curatie door een consortium van onderzoekers, en een work flow gebaseerd op tekstontginning. Het tekstontginningsplatform dat aan bod komt, wordt gevalideerd met een aantal gevallenstudies.

In het afsluitende Hoofdstuk 6, gaan we dieper in op de tekstontginningsaspecten die doorheen de verschillende hoofdstukken aan bod zijn gekomen. We doen dit door te kijken naar een heel specifieke toepassing van datafusie: we bekijken twee gevallenstudies waar de ontginning van biomedische tekst samen wordt gebruikt met gegevens uit een ander type microrooster-experimenten, namelijk die waarbij de expressie van genen onder verschillende omstandigheden in kaart wordt gebracht. Een eerste casus behandelt de manier waarop genen en clusters van genen die typisch voortkomen uit de analyse van expressie-experimenten, kunnen worden geprofileerd ter validatie van deze gengroepen. Een tweede casus toont hoe textprofielen gebruikt worden voor het integreren van microroostergegevens uit expressie-experimenten en tekst. Hier wordt besproken hoe Bayesiaanse netwerken met toepassingen in de oncologie en in genregulatienetwerken worden versterkt met priordata die voortkomen uit biomedische literatuur.

Ten slotte volgt een nabeschouwing waarin verder onderzoek en toekomstperspectieven worden belicht.

0.3 Samenwerkingen

Dit doctoraat was niet tot stand kunnen komen zonder een aantal erg vruchtbare samenwerkingen zowel binnen als buiten de K.U.Leuven. Naast de collega's binnen de bioinformaticagroep zowel als andere groepen binnen het departement Elektrotechniek van onze Alma Mater, moeten zeker en vast het Centrum Menselijke Erfelijkheid van het universitair ziekenhuis Gasthuisberg worden vermeld, het Wellcome Trust Sanger Institute (en daar specifiek de groep Molecular Cytogenetics en het Web Team), het European Bioinformatics Institute (en daar specifiek de Rebholz Group voor Text Mining), het Centrum voor Medische Genetica van de Universiteit Gent, en het Leuven Biostatistics and Statistical Bioinformatics Centre.

Contents

Voorwoord **i**

Abstract **v**

Korte inhoud **vii**

Nederlandstalige samenvatting **ix**
0.1 Impact van Array-CGH als genetische screeningtechniek x
0.2 Opbouw van dit proefschrift xi
0.3 Samenwerkingen xiii

Contents **xiv**

List of acronyms **xxi**

Publication list **xxv**

1 Introduction **1**
1.1 Genetics in Medicine 1
1.1.1 The chromosomes 1
1.1.2 Development 2
1.1.3 Looking at genomes 2
1.2 Enter microarray technology 4
1.3 Microarray technology in the clinic 6
1.4 Impact of Array CGH 7
1.5 Context, novelty, contributions and collaborations 7
1.5.1 molecular cytogenetics: a new field and new challenges . . 7
1.5.2 Key issues addressed in this work 9

2 Array CGH data management and analysis **13**
2.1 Array CGH quality criteria for constitutional genetic diagnosis . 14
2.1.1 Array CGH in genetic diagnosis 14
2.1.2 Parameters influencing Array CGH quality 15
2.1.3 Quality criteria and threshold values 18
2.1.4 Polymorphisms 18

2.1.5 Mosaicisms . . . 19
2.2 Discussion on quality criteria . . . 19
2.2.1 Parameters influencing Array CGH quality . . . 20
2.2.2 Quality criteria . . . 20
2.2.3 Pitfalls for pre- and postnatal diagnosis: polymorphic loci 21
2.2.4 Discussion . . . 22
2.2.5 Back to the future . . . 24
2.3 A loop design analysis for Array CGH data . . . 25
2.3.1 Introduction . . . 25
2.3.2 Acknowledgements and credits . . . 26
2.3.3 A change in experimental setup . . . 26
2.3.4 New statistical models for loop design analysis . . . 28
2.3.5 Methods . . . 30
2.3.6 Statistical models: a discussion . . . 33
2.3.7 Web application . . . 38
2.3.8 Discussion on the LOOP approach . . . 49
2.3.9 Future work . . . 51
2.4 Human copy number variation . . . 52
2.4.1 Importance of copy number variation mapping . . . 52
2.4.2 Charting CNVs through accumulated diagnostic assay data 53
2.4.3 LOOP hybridizations as valuable data source . . . 54
2.4.4 Charting CNVs: conclusions . . . 59
2.5 Array repositories . . . 61
2.5.1 ArrayCGHBase . . . 62
2.5.2 Towards functional interpretation and intelligent databases 63

3 Functional interpretation of genetic information 65
3.1 Candidate genes for novel submicroscopic chromosomal imbalance syndromes . . . 66
3.2 Computational approaches in correlating genotype and phenotype 67
3.2.1 Querying genotype-phenotype correlations in literature databases 67
3.2.2 Finding phenotype-rich genotypical features . . . 68
3.2.3 Pinpointing phenotype-related genes: guilt by association 69
3.3 Challenges for automated genotype-phenotype correlations . . . 70
3.4 Statistically linking genotypes and phenotypes . . . 74
3.4.1 Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations . . . 74
3.4.2 Identification of loci involved in development, dysmorphy and pathology . . . 75
3.4.3 Automatically building chromosomal aberration maps . . 78
3.4.4 Results . . . 85
3.4.5 Discussion . . . 88
3.5 aGeneApart: mapping genes to human dysmorphology phenotypes by literature mining . . . 89
3.5.1 A pipeline to extract relevant concepts from literature . . 90
3.5.2 Validation efforts . . . 94

3.5.3 Availability through a web interface . . . 96
3.5.4 Integration in other tools . . . 96
3.6 Gene prioritization through genomic data fusion . . . 97
3.6.1 Data fusion for candidate gene prioritization . . . 98
3.6.2 Principle of Endeavour . . . 100
3.6.3 A text model for Endeavour . . . 101
3.7 Setting the scene for intelligent databases . . . 103

4 Intelligent databases for constitutional cytogenetics 109
4.1 Biomedical interpretation, collaborative research and diagnostic support . . . 109
4.2 Public databases . . . 110
4.3 Store and Bench . . . 111
4.3.1 Challenges for genomics IT platforms . . . 111
4.4 DECIPHER as an Array-CGH genomics platform . . . 116
4.5 DECIPHER and copy number changes . . . 116
4.5.1 Copy number changes and disease . . . 117
4.5.2 Normal copy number variation . . . 117
4.5.3 The DECIPHER project . . . 117
4.5.4 The DECIPHER consortium . . . 118
4.5.5 Key features of DECIPHER . . . 119
4.5.6 Using DECIPHER . . . 124
4.5.7 Main benefits of DECIPHER . . . 125
4.6 Store+Bench . . . 127
4.6.1 Main feature 1 - managing Array CGH related information 127
4.6.2 Main feature 2 - flexible reporting . . . 130
4.6.3 Main feature 3 - a data store . . . 133
4.6.4 Main feature 4 - integration and interaction with external resources . . . 133
4.6.5 Main feature 5 - gene prioritization . . . 133
4.6.6 Key applications of Store+Bench . . . 135
4.7 Conclusions . . . 137

5 Genome annotation 139
5.1 Determining sequence, and annotating it . . . 139
5.2 A reference sequence for many genomes . . . 140
5.3 Current approach to genome annotation . . . 141
5.4 Genome annotation: technical aspects . . . 141
5.4.1 What is DAS? . . . 141
5.4.2 How does DAS work? . . . 142
5.4.3 Distributed annotation: PrimerDB . . . 142
5.4.4 Distributed annotation: Bench and DECIPHER . . . 144
5.5 Case study: text-mining assisted regulatory annotation . . . 144
5.5.1 Unlocking transcription regulation information in biomedical literature . . . 144
5.5.2 Background . . . 144

5.5.3 A literature management system for community annotation and text mining . . . 147
5.5.4 A vector space model identifies MEDLINE abstracts with high cis-regulatory content . . . 147
5.5.5 Coverage . . . 149
5.5.6 Estimating the size of the cis-regulatory corpus . . . 151
5.5.7 Abstract relevance ranking aids the construction of regulatory networks . . . 153
5.5.8 Text Mining for cis-regulatory information: discussion . . . 153

6 Data fusion on biomedical text and gene expression assays 157
6.1 Case 1: TXTGate – a platform for gene group profiling and expression array cluster validation . . . 158
6.1.1 Introduction - a flexible framework . . . 158
6.1.2 Rationale . . . 158
6.1.3 Related software . . . 160
6.1.4 Representing text data . . . 161
6.1.5 Application overview . . . 162
6.1.6 Program development . . . 164
6.1.7 Construction of domain vocabularies . . . 165
6.1.8 Online clustering . . . 166
6.1.9 Cluster coherence . . . 167
6.2 TXTGate summarizes and identifies sub-clusters . . . 167
6.2.1 Yeast data . . . 167
6.2.2 Human data . . . 168
6.2.3 Textual information through the eyes of different vocabularies . . . 172
6.3 Application of TXTGate to a real-life research problem . . . 174
6.4 Case 2: Integration of microarray and textual data through Bayesian networks . . . 175
6.5 Introduction . . . 175
6.6 Bayesian networks . . . 177
6.6.1 Model building . . . 178
6.6.2 Structure prior . . . 179
6.6.3 Inference . . . 180
6.7 Prior data . . . 181
6.7.1 Gene prior . . . 181
6.7.2 Class variable prior . . . 181
6.8 Data . . . 182
6.8.1 Discretization . . . 182
6.9 Implementation . . . 183
6.10 Results and discussion . . . 183
6.10.1 Veer data . . . 183
6.10.2 Bild data . . . 184
6.11 Conclusions . . . 184
6.11.1 Microarray and text data for gene cluster profiling . . . 184

6.11.2 Microarray and text data for Bayesian Networks 185

7 Conclusions and perspectives **187**
7.1 Accomplishments and conclusions 187
7.2 Future research and perspectives 188
7.2.1 Future proof platforms 188
7.2.2 Collecting and representing the genome and phenome . . 189
7.2.3 Clinical validation 191
7.2.4 Towards personal genomics 191
7.2.5 The future of genome annotation 191

Bibliography **193**

List of acronyms

AUC Area Under the ROC Curve

AJAX Asynchronous JavaScript and XML

BAC Bacterial Artificial Chromosome

BED Browser Extendable Data

BIND Biomolecular Interaction Network Database

BLAST Basic Local Alignment Search Tool

bp base pairs

cDNA Complementary DNA

CGH Comparative Genomic Hybridization

CME Centrum Menselijke Erfelijkheid, Katholieke Universiteit Leuven

CNP Copy Number Polymorphism

CNV Copy Number Variation / Copy Number Variant

CNVR Copy Number Variant Region

CPT Conditional Probability Table

DAG Directed Acyclic Graph

DAS Distributed Annotation System

DAS Distributed Annotation Server/Service

DNA Deoxyribonucleic Acid

DOP-PCR Degenerate Oligonucleotid Primer (DOP) PCR

EBI European Bio-informatics Institute

EMBL European Molecular Biology Laboratory

EmsEMBL a comprehensive genome information system featuring genome databases and annotation

ESAT Department of Electrical Engineering, Katholieke Universiteit Leuven

EST Expressed Sequence Tag

FDA Food and Drug Administration

FISH Fluorescent In Situ Hybridization

Feature physical reporter spotted, printed, or otherwise linked to a substrate at a specific location

GO Gene Ontology

GPR GenePix Results File

GUI Graphical User Interface

HGNC Human Gene Nomenclature Committee

HMM Hidden Markov Model

HPO Human Phenotype Ontology

HUGO Human Genome Organization

ICD-9 International Classification of Diseases Clinical Modification

IDF Inverse Document Frequency

InDel Insertion or Deletion

ISCN International System for human Cytogenetic Nomenclature

JSP Java Server Pages

kb kilobases (1000 base pairs)

KB Kilobytes (1024 bytes)

kD kilo-Dalton (molecular weight of a protein)

KEGG Kyoto Encyclopedia of Genes and Genomes

JSP Java Server Pages

LIMMA Linear Models for Microarray Data

LIMS Laboratory Information Management System

LDDB Oxford Medical Dictionary London Dysmorphology Database

LNDB Oxford Medical Dictionary London Neurology Database

LOESS see LOWESS

LOWESS LOcally WEighted Regression Scatter-plot Smoothing

MA M (intensity ratio) versus A (average intensity) plot to visualize intensity-dependent ratio of raw microarray data

mb megabases (1000 base pairs)

MB Megabytes (1024 kilobytes)

MCA Multiple Congenital Anomalies

MIAME Minimal Information About a Microarray Experiment

MPO Mammalian Phenotype

MR Mental Retardation

MR-MCA Mental Retardation and Multiple Congenital Anomalies

mRNA messenger RNA

MySQL My Structured Query Language

NCBI National Center for Biotechnology Information

NEG negative

NGS Next Generation Sequencing

oligo A short oligonucleotide DNA probe whose hybridization is sensitive to a single base mismatch

OMD Oxford Medical Dictionary

OMIM Online Mendelian Inheritance in Man

ORF Open Reading Frame

PAC P1-plasmid-derived Artificial Chromosome

PCR Polymerase chain reaction

PDF Portable Document Format

PEM Paired End Mapping

PERL Practical Extraction and Report Language

PHP Hypertext PreProcessor (server-side scripting language)

PMID PubMed Identifier

POS positive

PVV Positive Predictive Value

QA Quality Assessment / Assurance

qPCR Quantitative PCR

RAID Redundant Array of Independent Disks

RMI Remote Method Invocation

RNA Ribonucleic Acid

ROC Receiver Operating Characteristic

SaaS Software as a Service

SGD Saccharomyces Genome Database

SNP Single Nucleotide Polymorphism

SN Signal to Noise ratio

SNOMED Systematized Nomenclature of Medicine

SOAP Simple Object Access Protocol

SSL Secure Socket Layer

SVM Support Vector Machine

TFBS Transcription Factor Binding Site

TF Term Frequency

TF Transcription Factor

TFBS Transcription Factor Binding Site

TG Target Gene

TP True Positive Rate

TRANSFAC knowledge base containing data on transcription factors

UCSC University of California Santa Cruz

USCA Catalogue of Unbalanced Structural Chromosome Abnormalities

UMLS Unified Medical Language System

VOC vocabulary

VSM Vector Space Model

WBS Williams-Beuren Syndrome

YAC Yeast Artificial Chromosome

Publication list

International journal papers

- Firth H V, Richards S, Bevan P, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett R M, Carter N P., "DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources" [65]. *Am J Hum Genet* 84(4):524-33. 2009.

- Yu S, Van Vooren S, Tranchevent LC, De Moor B, Moreau Y "Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining" [272] *Bioinformatics* 24(16):i119-125, 2008.

- Tranchevent L C, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y, "ENDEAVOUR update: a web resource for gene prioritization in multiple species [245]" *Nucleic Acids Res.* 36:W377–384, 2008.

- Gevaert O, Van Vooren S, De Moor B, "A framework for elucidating regulatory networks based on prior information and expression data [81]" *Ann N Y Acad Sci.*, 1115:240-8, 2007.

- Gevaert O, Van Vooren S, de Moor B, "Integration of microarray and textual data improves the prognosis prediction of breast, lung and ovarian cancer patients [82]" *Pac Symp Biocomput.*, 2008:279-90, 2008.

- Breckpot J, Takiyama Y, Thienpont B, Van Vooren S, Vermeesch J R, Ortibus E, Devriendt K, "A novel genomic disorder: a deletion of the SACS gene leading to Spastic Ataxia of Charlevoix-Saguenay [33]" *Eur. J. Hum. Genet.*, electronic publication ahead of print, 2008.

- Maas N M C, Van Vooren S, Hannes F, Van Buggenhout G, Mysliwiec M, Moreau Y, Fagan K, Midro A, Engiz Ö, Balcis, Parker M J, Sznajer Y, Devriendt K, Fryns J P, Vermeesch J R., "The t(4;8) is mediated by homologous recombination between olfactory receptor gene clusters, but other 4p16 translocations occur at random [148]" *Genet. Couns.*, 18(4):357-65, 2007.

- Aerts S, Haeussler M, Van Vooren S, Griffith O L, Hulpiau P, Jones S J, Montgomery S B, Bergman C M, The Open Regulatory Annotation Consortium, "Text-mining assisted regulatory annotation [4]" *Genome Biol.* 9(R31):1-13, 2008.

- Balikova I, Martens K, Melotte C, Amyere M, Van Vooren S, Moreau Y, Vetrie D, Fiegler H, Carter N P, Liehr T, Vikkula M , Matthijs G, Fryns J P, Casteels I, Devriendt K, Vermeesch J R., "Autosomal dominant microtia linked to five tandem copies of a copy number variable region at chromosome 4p16 [13]" *Am J Hum Genet.*, 82(1):181-7, 2008.

- Van Vooren S, Coessens B, De Moor B, Moreau Y, Vermeesch J R, "Array CGH and computational genome annotation in constitutional cytogenetics: suggesting candidate genes for novel submicroscopic chromosomal imbalance syndromes [252]" *Genet Med.*, 9(9):642-9, 2007.

- Van Vooren S, Thienpont B, Menten B, Speleman F, De Moor B, Vermeesch J R, Moreau Y, "Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations [253]" *Nucleic Acids Res.*, 35(8):2533-43, 2007.

- Menten B, Maas N, Thienpont B, Buysse K, Vandesompele J, Melotte C, de Ravel T, Van Vooren S, Balikova I, Backx L, Janssens S, De Paepe A, De Moor B, Moreau Y, Marynen P, Fryns JP, Mortier G, Devriendt K, Speleman F, Vermeesch J R, "Emerging patterns of cryptic chromosomal imbalance in patients with idiopathic mental retardation and multiple congenital anomalies: a new series of 140 patients and review of published reports. [158]" *J Med Genet.*, 43(8):625-33, 2006.

- Menten B, Pattyn F, De Preter K, Robbrecht P, Michels E, Buysse K, Mortier G, De Paepe A, van Vooren S, Vermeesch J, Moreau Y, De Moor B, Vermeulen S, Speleman F, Vandesompele J, "arrayCGHbase: an analysis platform for comparative genomic hybridization microarrays [159]" *BMC Bioinformatics.*, 6:124, 2005.

- Vermeesch J R, Melotte C, Froyen G, Van Vooren S, Dutta B, Maas N, Vermeulen S, Menten B, Speleman F, De Moor B, Van Hummelen P, Marynen P, Fryns JP, Devriendt K, "Molecular karyotyping: array CGH quality criteria for constitutional genetic diagnosis [258]" *J Histochem Cytochem.*, 53(3):413-22, 2005.

- Glenisson P, Coessens B, Van Vooren S, Mathys J, Moreau Y, De Moor B, "TXTGate: profiling gene groups with text-based information [86]" *Genome Biol.*, 5(6):R43, 2004.

International conference papers

- Yu S, Van Vooren S, Tranchevent L C, De Moor B, Moreau Y, "Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining", in Proc. of the European Conference of Computational Biology 2008 (ECCB 2008), Cagliari, Italy, Sep. 2008, 7 p.

- Yu S, Van Vooren S, Coessens B, De Moor B, "Interpreting Gene Profiles from Biomedical Literature Mining with Self Organizing Maps." In *Proceedings of International Symposium on Neural Networks (ISNN)*, (2):635-641, Chengdu, China, 2006.

- Glenisson P, Coessens B, Van Vooren S, Moreau Y, De Moor B, "Text-Based Gene Profiling with Domain-Specific Views." In *Proceedings of First International Workshop on Semantic Web and Databases (SWDB)*, (2):15-31, Berlin, Germany, 2003.

- Pelckmans K, Van Vooren S, Coessens B, Suykens J A K, De Moor B, "Mutual Spectral Clustering: Microarray Experiments Versus Text Corpus." In *Proceedings of workshop on Probabilistic Modeling and Machine Learning in Structural and Systems Biology (PMSB)*, 1-4, Helsinki, Finland, 2006.

- Gevaert O, Van Vooren S, De Moor B, "The use of prior distributions for learning genetic networks." In *Proceedings of workshop on Probabilistic Modeling and Machine Learning in Structural and Systems Biology (PMSB)*, 103-107, Tuusula, Finland, 2006.

- Gevaert O, Van Vooren S, de Moor B, "Integration of microarray and textual data improves the prognosis prediction of breast, lung and ovarian cancer patients." *Pac Symp Biocomput.* 279-90, 2008.

- Van Vooren S, Maas N M C, Vermeesch J, Moreau Y, De Moor B, "CGH-Gate: Array-CGH, Case Reports, Phenotypes and Biomedical Literature for Human Genome Annotation" *ISMB 2005*, in Proceedings Thirteenth International Conference on Intelligent Systems for Molecular Biology 2005, Detroit, MI, USA, 25-29 June 2005.

Internal reports

- Barriot R., Breckpot J., Thienpont B., Van Vooren S., Coessens B., Tranchevent L.-C., Van Loo P., Gewillig M., Devriendt K., Moreau Y., "Collaboratively charting the gene-to-phenotype network of human congenital heart defects", Internal Report 08-184, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2008.

- Yu S, Van Vooren S, Janssens F, De Moor B, Moreau Y., "Using high dimensional gene profiles from multiple views of text mining for Gene Prioritization". *Internal report.*, 2007.

- Allemeersch J, Van Vooren S, Hannes F, De Moor B, Vermeesch J, Moreau Y., "An experimental loop design improves the detection of congenital chromosomal aberrations by array CGH" Manuscript in preparation.

- Vermeesch J., Martens K., Maas N., Melotte C., Van Vooren S., De Moor B., Engelen J., Starke H., Vetrie D., Fiegler H., Matthijs G., Fryns JP., Casteels I., Devriendt K., "Familial Inherited Microtia caused by 4pter Olfactory Receptor Gene Cluster Amplification", Internal Report 05-25, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2005.

- Coessens B., Van Vooren S., Glenisson P., Moreau Y., De Moor B., "Term-based literature mining across systems biology and biomedicine", Internal Report 04-196, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2004.

International Talks and Posters

- Van Vooren S, Maas NMC, Melotte C, Devriendt K, Vermeesch JR, De Moor B. "CGHGate: annotating the human genome by linking CGH array case reports, clinical descriptions and biomedical literature." 1st Marie Curie Conference on array CGH and Molecular Cytogenetics, Cambridge, UK, 29/9/2004 - 2/10/2004. Oral presentation.

- Van Vooren S, Maas NMC, Melotte C, Devriendt K, Vermeesch JR, De Moor B. "CGHgate: Array-CGH, Case reports, phenotypes and biomedical literature for human genome annotation." 13th International conference on intelligent Systems for Molecular biology, Michigan, USA, 25-29/6/2005. Oral presentation.

- Menten B, Buysse K, Maas NMC, Thienpont B, Vandesompele J, Melotte C, de Ravel T, Van Vooren S, Balikova I, Backx L, Janssens S, De Paepe A, De Moor B, Moreau Y, Marynen P, Fryns J-P, Mortier G, Devriendt K, Vermeesch JR, Speleman F. "Array CGH findings in a large series of 150 patients with idiopathic mental retardation and congenital anomalies"

- Thienpont B, Maas NMC, Menten B, Buysse K, Vandesompele J, Melotte C, de Ravel T, Van Vooren S, Balikova IG, Backx L, Janssen S, De Paepe A, De Moor B, Moreau Y, Marynen P, Fryns JP, Mortier G, Devriendt K, Speleman F Vermeesch JR. "Detection of as low as 5 % structural low grade mosaics by array CGH in patients with idiopathic mental retardation and multiple congenital aberrations". Oral presentation, 2nd Marie Curie Conference on arrayCGH and Molecular Cytogenetics, 2005, Bari, Italy

- Menten B., Pattyn F., De Preter K., Robbrecht P., Michels E., Buysse K., Mortier G., De Paepe A., Van Vooren S., Vermeesch J., Moreau Y., De Moor B., Vermeulen S., Speleman F., Vandesompele J.: "arrayCGHbase: an analysis platform for comparative genomic hybridization microarrays." BMC Bioinformatics, 6, 124, 2005.

Chapter 1

Introduction

"La vie a une histoire écrite par l'évolution et conservée par hérédité."

Jules Carles (1945)

1.1 Genetics in Medicine

The basic unit of function in an organism is the cell. While most micro-organisms are single-celled, organisms like you – the person reading this PhD booklet – are composed of many cells. The number of cells in your body is estimated to 10 000 000 000 000 [37] (10^{13}) or even 100 000 000 000 000 [211] (10^{14}). Every one of these cells contains the same set of genes: each cell nucleus has your genetic material in the form of 23 different chromosomes. These molecules are long chains of very small molecules or nucleotides: G, A, T and C.

Interestingly, your human body contains ten times more microbes than it does cells [211]. Still, at the very beginning of our lives, we all originate from one single cell. This cell, the fertilized egg cell, is the fusion of an egg and sperm cell and hence has received 23 chromosome molecules from the mother and an equal number from the father.

Cells multiply through cell division. This means a cell grows in size, duplicates its chromosomes and separates them for exact distribution between the two daughter cells. These processes form the *cell cycle*. This mechanism allows organisms to grow. Defects in the genetic make-up of a cell or acquired genomic alterations may lead to problems in the development of an organism.

1.1.1 The chromosomes

Chromosomes are thread-like structures composed of DNA and other proteins that are present in every cell of the body and carry the genetic information needed for that cell to develop. On chromosomes we find genes which are units of information that are encoded in our DNA forming blueprints for proteins which are the working horses of any organism.

Normal human cells have 46 chromosomes, arranged in 23 pairs. Between males and females, 22 out of 23 are always alike (the autosomes), while the 23rd pair are the sex chromosomes ('X' and 'Y'). Each member of a pair of chromosomes carries the same information, and within each pair of chromosomes, the same genes are in the same location along the two chromosomes. The individual genes, however, may vary. These alternative versions of genes are called alleles, and account for many phenotypic differences.

1.1.2 Development

As organisms develop, errors can occur during cell division. For example, chromosome pairs are supposed to divide evenly at cell division. When a sperm and an egg meet, the numbers add up again to the regular 46 chromosomes. When this fails, this can result in cells having a surplus or lack of genetic material. This can happen for entire chromosomes, but also for parts of these molecules. A typical example is Down Syndrome, which is in most cases caused by a fertilized egg ending up with three instead of two copies of the 21th chromosome.

Normal development of a healthy embryo can also be disrupted when errors occur after fertilization. Typically, a number of cells will then end up with a correct genetic make-up, and a fraction won't: this is called mosaicism.

Other mechanisms for genetic disease include mutations in the genetic material at a later stage in life, disrupting the cell cycle mechanism and leading to cancer. Cancer is in essence a genetic disease for which a predisposition can exist in genetic make-up, but is also influenced by mutation events during life itself, e.g. through environmental factors such as smoking.

While acquired genomic alterations leading to cancer are different from congenital aberrations, leading to developmental problems and genetic disease, the common ground is the genetic material of the cells in an organism. In this PhD, we will mainly focus on congenital aberrations, although a number of methods and tools discussed in this work can be applied to the realm of cancer as well.

1.1.3 Looking at genomes

Classical karyotyping

Different techniques have been developed and brought into practice to investigate the genome of an organism - from virus to man, from plant to bacterium or yeast. A specific application domain has always been genetic testing in clinical practice and research. One type of genetic variation are single nucleotide polymorphisms or SNPs. These correspond to single base pairs that deviate from the majority of occurrences at that specific location of the genome. These variations give rise to different alleles, and can alter the function of a gene. The effects can be harmless (e.g. causing the inability to complete the full process of building hair pigment, so that the person in question has blond hair) or harmful (leading to developmental issues or genetic disease). Another form of variation that can have either benign or harmful effects, are chromosomal rearrangements. In the

process of cell division, fragments of genetic material can end up duplicated, deleted, inverted, or otherwise rearranged.

The characterization of chromosomal rearrangements and the identification of corresponding breakpoints is hence an important task in the study of genetic disease. In the field of human cytogenetics, geneticists and researchers strive to find reliable approaches to do this. One technique that is well-established due to its relative ease to perform at acceptable cost, is G-band staining. In this method, chromosomes are painted so that banding patterns emerge that can be discerned under a microscope, revealing copy number or other structural variations. As of today, it is still used as the cytogenetic gold standard.

Fishing for higher resolution

A second technique is FISH. Fluorescent In Situ Hybridization was introduced in the 1980s and marked the birth of molecular cytogenetics, because this technique allows us to look at very precise locations on the genome. Due to resolution limits of looking at chromosome material under microscopes, this is not possible through G-band staining, where the highest level of visible detail lies at 3 to 5 million base pairs, and smaller chromosomal aberrations often remain hidden. Resolution of genomic screening is crucial: over the last decades, improved resolution of cytogenetic techniques has lead to a significant increase in the detection rate of chromosomal aberrations in patients with mental retardation (MR) and/or congenital anomalies. The identification of submicroscopic subtelomeric alterations in 3 - 7% of idiopathic MR patients [67, 123, 227, 193, 197], as well as the sporadic reports of submicroscopic interstitial chromosomal rearrangements, suggests that a substantial portion of idiopathic MR may be caused by smaller chromosomal rearrangements. These observations make it clear that higher resolution screening techniques for the detection of small deletions or duplications at any chromosomal position will drastically increase the elucidation of human genetic diseases.

In contrast to standard chromosome banding techniques, which are based on an enzymatic reaction like G-Banding, FISH banding methods are DNA-specific. With FISH, a single sequence probe is used to recognize its counterpart on the genome, and through fluorescent microscopy, copy number at that specific genomic location can usually be clearly discerned. FISH permits the determination of the number and location of specific DNA sequences, both in metaphase chromosomes and in interphase nuclei.

FISH in a diagnostic setting

While the resolution of working with FISH probes is very high, it is not possible to look at the entire genome at once, despite the existence of processes like multicolor-FISH, where probes of different colors are used simultaneously. Fluorescence In Situ Hybridization (FISH) banding approaches are standard for the exact characterization of simple, complex and even cryptic chromosomal aberrations within the human genome. FISH is a technique that is heavily used

in a clinical setting for genetic screening. It is performed routinely to check for aneuploidy, rearrangements, small deletions and duplications in contiguous gene syndromes, for both constitutional and oncology applications. Although multi-color FISH and chromosome painting probes allow for flexible approaches for genetic screening, still, FISH cannot chart the full genome in a single step. Hence, the researcher can only use FISH if he or she knows what to look for, for example, by looking for features consistent with a clinically recognizable syndrome with a known chromosomal etiology, or by first performing a different type of screening to check for an abnormal karyogram that requires further molecular characterization. Although FISH has dramatically increased the sensitivity of detection of genomic imbalances, this need for prior knowledge of the chromosomal region(s) of interest makes FISH inapplicable for whole genome screening approaches required in a diagnostic setting.

A recent breakthrough in characterizing the structure and copy number in a patient genome has superseded this limitation of FISH. With the advent of Array CGH, in essence, thousands of FISH experiments can be executed simultaneously. In this technique, probes are now fixed to a known position on a small microscope glass slide. Also, for these probes, there is a connection to a mapped position in the human genome sequence.

1.2 Enter microarray technology

The advent of array based comparative genomic hybridization has revolutionized the field over a brief period of time, with applications in many areas of medicine and biomedical research. The impact of the adoption of Array CGH has been great: the discovery of new syndromes has leaped forward, existing syndromes have been characterized with increased detail since. Current applications of high-density array CGH have led to the identification of specific genetic alterations in sporadic and supposedly multifactorial conditions such as cancer, Alzheimer and autism [19]. Moreover, through the rise of Array CGH as a tool in genetic research, a new appreciation of the diversity of the human genome is now emerging.

Array Based Comparative genomic hybridization (Array CGH) is in fact a variation on the FISH theme that allows to reveal imbalances across the whole genome. Array CGH enables the identification of chromosomal copy number changes. The availability of clone sets covering the human genome opens the possibility for the widespread use of array CGH for both research and diagnostic purposes: pre- and postnatal cytogenetic testing is aimed at the genome-wide analysis for the detection of chromosome aneuploidies and segmental aneusomies. In this, Array CGH is taking over the role of karyotyping. In addition to its clinical importance, the identification of a chromosomal aberration in specific patients has proven to be a successful way to identify the implicated genes and to gain insight in the pathogenesis of different genetic conditions.

Array CGH uses mapped DNA sequences in a microarray format as a platform for the detection of chromosomal deletions/duplications. In this tech-

nique, genomic DNA from the patient is labeled with one fluorescent dye while a normal reference sample is labeled with a different dye, and finally these samples are co-hybridized to the array containing the genomic DNA targets. Chromosomal imbalances across the genome can thus be quantified and their positions defined by analyzing the ratio of the fluorescence of the two dyes along the targets. The resolution of array CGH depends on the size of the genomic fragments as well as on their density. Proof of principle was established in 1997 [178, 232] and since then, more and more labs have mastered the technology. Currently, Array CGH is at a global breakthrough. Array CGH detects copy-number aberrations and variations at a high resolution, and on a genome-wide scale [230, 179]. The technique is also called molecular karyotyping [259, 258, 257], and genome-wide array CGH has been applied to detect chromosomal imbalances in patients with congenital anomalies and mental retardation in several studies [224, 196, 209, 260, 158, 73, 51]. An illustrative example is presented in Figure 1.1.

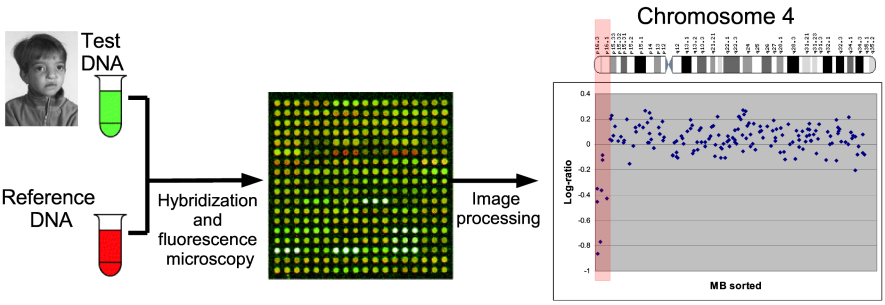


Figure 1.1: BAC-based Array CGH. Wolf-Hirschhorn syndrome was discovered in 1961 by Herbert Cooper and Kurt Hischhorn. The phenotypical features include mental retardation, distinct facial appearance (typical Greek warrior helmet faces, high forehead), and seizures. Wolf-Hirschhorn is characterized by a deletion of the end of the short arm of chromosome 4; in particular, a deletion of the terminal band (4p16.3) is essential for full expression of the phenotype. Wolf-Hischhorn can be detected with array CGH by comparing a genomic DNA sample of the patient (test) with that of a normal individual (reference). DNA extracted from test and reference sample is labeled with different fluorescent dyes (typically Cy3 and Cy5) and hybridized to the microarray. Array CGH probes can be PCR-amplified Bacterial Artificial Chromosomes or BAC clones or spotted long oligos. The microarray is scanned by a two-channel laser scanner and aneuploid chromosomal regions are detected as probes with a deviant log-ratio. This example clearly indicates a deviation of the log-ratios at the end of the short arm of chromosome 4 and allows to confirm the hypothesis of Wolf-Hirschhorn syndrome.

1.3 Microarray technology in the clinic

Array CGH is a highly effective technique that is entering routine clinical use much faster than other microarray technologies. Indeed, compared to for example expression microarrays, array CGH enjoys several technical advantages: (1) genomic DNA samples are less prone to degradation than mRNA samples, (2) genomic DNA samples show much less variation between biological replicates than mRNA samples, and (3) interpretation of chromosomal imbalances is much easier than that of expression fingerprints. These advantages explain why array CGH for the diagnosis of constitutional anomalies is progressing faster towards the clinic than expression microarrays for the prediction of clinical outcome (e.g., in cancer), for which a few applications are now entering clinical practice [30, 38, 225].

As explained earlier, Array CGH mostly competes with and is complementary to conventional karyotyping and fluorescent in situ hybridization (FISH). Compared to conventional karyotyping, it offers a resolution between 10kb and 1Mb, depending on the platform used. With high density oligo arrays, the limit lies even lower. With about half a million reporters on an assay spread out evenly over the human genome, as is the case with currently available commercial slides, and since the human genome is 3 billion base pairs long, one could in theory attain a resolution of 6 kb with a single assay. Multiple assays can be combined and the resolution of available platforms is ever increasing.

Array CGH detects at least twice as many aberrations as conventional karyotyping [51]. Furthermore, it does not require the use of metaphase chromosomes, which makes it faster and less labor intensive. However, current array CGH techniques cannot detect balanced translocations, while this is straightforward (at least, for large enough aberrations) with conventional karyotyping. Compared to FISH, array CGH provides genome-wide coverage, instead of covering only a limited set of probes—so, it does not require prior knowledge of which aberration might be present (*e.g.*, based on the phenotype of the patient).

The most frequent experimental setup for array CGH consists in comparing genomic DNA of a patient (test) with that of a normal individual (reference) using a two-channel microarray consisting of DNA segments spread across the whole genome. In the case of the initial clinical platform that was set up at CME Leuven when the technology was being adopted, the DNA segments consist in PCR-amplified BAC clones. However, the discussion applies equally to spotted long oligo platforms. Hence, we refer to probes as *reporters*. DNA from the test and reference samples is extracted, labeled with different fluorescent dyes (usually Cy3 and Cy5), hybridized to the microarray, and then scanned by a two-channel laser scanner. Aneuploid chromosomal regions are detected as probes with a deviant log ratio of the intensities of the test against reference signal (approximately $\log_2(1/2)$ for a deletion and $\log_2(3/2)$ for a duplication). Possibly, the experiment is repeated in a dye-swap with the fluorescent labeling of test and reference exchanged. The signals are then averaged over the dye-swap replicates to reduce the signal-to-noise ratio.

1.4 Impact of Array CGH

The development of array CGH is the most recent advancement in molecular cytogenetics and has lead to a deeper understanding of the complexity of the human genome. The increase in complexity is manifested by the identification of ever more subtle DNA copy gains and losses, where results of earlier genetic screens would return negative. The success of Array CGH as a technique in both a clinical setting and research has resulted in an exponential increase in genetic information. The body of information that becomes available far exceeds our ability to understand and use it in a clinical setting. We learn more on existing syndromes, the discovery of new syndromes has expanded drastically, and the same will hold for our understanding of complex and multifactorial genetic conditions.

A key consequence of the adoption of this technology is in fact an unprecedented reversal of the usual order of the practice and progress of medicine [19]: while clinical suspicion and medical insights typically suggest specific genetic lesions and hence are the drivers for the actual genetic screen and lab investigation, the advent of Array CGH turns this paradigm around. With the advent of next generation sequencing technologies, that are expected to enable patients to be fully sequenced for under 1000 Euro within the next ten years, this paradigm shift will even be strengthened, expressed by the phrase 'sequence first, ask questions later'. Clinicians are now able to scrutinize the genome for guidance in their clinical practice.

Another consequence of the use of Array CGH is that our understanding of copy number variations (CNVs) and their roles in health and disease is improving. A large effort will be needed to mark their role as susceptibility factors in complex genetic disease, to chart population specific aspects of copy number variations, and to make results of such research amenable to use in clinical practice.

These dramatic changes require the design of specific analysis approaches to deal with this flood of new information in research and in clinical practice. This dissertation does not describe the work of a single PhD student, but marks a fruitful collaboration of different research groups that have set the common goal of addressing past and future challenges of this exciting discipline, marking the advent of the era of genomic medicine.

1.5 Context, novelty, contributions and collaborations

1.5.1 molecular cytogenetics: a new field and new challenges

Genetic screening has gone through a rapid evolution over the past couple of years. The introduction of array CGH as a technique for detecting copy number variation in individuals has been a revolution in this field, and is now commonly

called molecular karyotyping. The use of molecular karyotyping has introduced a need for means to process, manage, visualize and interpret the larger and larger amounts of resulting genotype data and use them in clinical routine en in biomedical research. In this newly arisen niche, a tight collaboration between the bio-informatics group and the center for human genetics at the University Hospital was founded. It is within this collaboration that this PhD work has been conducted. In this section, specific contributions of this work to the state of the art are described. Since this research is of collaborative nature by definition, credit is attributed to people that have played a mayor or even leading role in one or more of the achievements described in this manuscript.

contribution of this dissertation to the state of the art

The key contribution of this work to the state of the art has been the development of methods and approaches as well as production-grade software and database tools to annotate phenotype information to the genotype in a way that clinicians can readily and easily use it for research purposes as well as in routine clinical diagnostics. To this aim, different methods and approaches have been adopted from the realms of statistics, machine learning, search and clustering algorithms, ontologies and graphs, text mining approaches, data mining and data fusion, knowledge visualization, but also practices from software development and user interaction, since robust and stable interfaces had to be built to ensure a level of user interaction and confidence on par with the needs in hospital routine.

In short, new methods and tools have been developed that help answer research and clinical diagnosis questions that have arisen since the introduction and wider adoption of molecular karyotyping in genetic screening, from analysis of raw data to downstream interpretation of genetic and clinical findings.

novelty and personal contributions: chapter specific discussion

Rather than providing a summary of the entire dissertation manuscript, this section highlights specific contributions per chapter. In the initial phases of bringing Array CGH from research into clinical practice, questions arose regarding how to deal with the raw genotype data resulting from DNA copy number assays. Chapter 2 primarily deals with this, and shows how this doctoral work resulted in a specific data analysis approach for loop design experiments, where credit for setting up the statistical methods goes to Joke Allemeersh. Secondly, with a rise in the amount of raw data available from assays on samples from different individuals, that chapter discusses an approach to charting the variation in the human genome: in this work, several steps have been made towards building population based copy number variation maps.

Chapter 3 discusses functional interpretation of genetic information. A key contribution in this context has been a method that automatically annotates functional and phenotype information to the genome at band levels (aBandApart) and later at gene levels (aGeneApart) based on biomedical literature.

Apart from these annotation approaches, another significant contribution to functional genome interpretation has been text-based gene representations based on literature as a data source to Endeavour, a candidate gene prioritization platform developed at the bio-informatics group.

Chapter 4 moves further down from functional analysis into clinical interpretation of genomic and phenomic data. Here, we move into the realm of diagnosis clinical decision support. Two main contributions are to be highlighted: (1) the development of CGHGate (Now Cartagenia Bench), a platform for research and diagnosis support for array CGH based on the use of patient CNV information, phenotype and clinical annotations, knowledge in public databases and biomedical literature, and various phenotype ontologies; and (2) the results of a 5 month internship at the Wellcome Trust Sanger Institute, where the text based gene prioritization approaches developed up until that time in the context of this PhD work have been implemented in DECIPHER, an international database initiative to correlate rare genome events and phenotypes to further the understanding of genetic disease. Both platforms comprise key results of this doctoral work and contribute significantly to addressing challenges in bringing Array CGH and related genomic screens to routine clinical practice as well as research in gene function and heritable disease.

In Chapter 5, genome annotation is introduced, where a fruitful collaboration within the Open Regulatory Annotation Consortium has lead to the application of the text mining pipeline developed during this doctoral work to a specific example of genome annotation, *i.e.* transcription regulatory elements. This work has resulted in an automatic approach to populating a literature queue for speeding up the process of manual curation of cis-regulatory elements. Credit for validation of this same platform goes to Stein Aerts. Finally, this same chapter discusses specific examples of data fusion of text and microarray data. A first contribution in this context is TXTGate, a platform for validation of gene clusters obtained from gene expression assays. A second contribution is the construction of text based priors for learning structure and transition probabilities for Bayesian Networks. Credit for training and validation of the networks goes to Olivier Gevaert.

1.5.2 Key issues addressed in this work

Building a technical and generic framework for high throughput text mining

At the basis of many of the tools and methods that are discussed in this work lies the building of a generic text mining framework. This work was started at the bio-informatics research group by Peter Antal and continued by Patrick Glenisson. A complete overhaul of this framework led to most if not all of the text mining applications at our research group, and was not just the work of a single person: Frizo Janssens, Bert Coessens, Dries Vandromme and Yu Shi have also contributed significantly to this framework. Named TextPack, it has become an extensive and extensible pipeline that addresses all text mining needs

of the research group and serves as a starting point for many text based data sources used by aGeneApart, aBandApart, TXTGate, and Endeavour, as well as work on Support Vector Machines, clustering, data fusion, Bayesian Networks, and more.

describing the patient phenotype

We have worked on ways to to describe the patient phenotype and find relevant genotypes based on these descriptions. This has included adopting ontologies and vocabularies to represent knowledge on phenotypes of individual patients and of disease, and developing and adopting algorithms and statistics to traverse them, use them to compare patients and their representations, as well as other objects such as text documents and disease descriptions.

comprehensive phenotype annotation

So as to be able to effectively work with clinical descriptions and patient phenotypes, phenotype annotations have to be amenable for use in computer algorithms. The ontology and vocabulary approach in this work has been comprehensive, in that it extends from representing patient phenotype profiles over clusters of patients, profiles of diseases, genes, groups of genes, cytogenetic bands, to more complex entities such as disease outcome predictive networks.

extracting genotype and phenotype information from literature

Another key point in building aforementioned tools is the extraction of genotype and phenotype information from the literature, whether mined from unstructured data sources, extracted from PubMed, or detected in large bodies of patient hospital discharge reports. Features that enable this have been to a large extent built into the TextPack framework.

having a universal method to identify phenotype information from the literature

The methods used in this works are designed to be universally applicable in the sense that they are not disease specific or built to represent one particular phenotype, rather, they can readily be used after substituting one ontology with another, one domain vocabulary with another, one research question with another.

different levels of phenotype information

Phenotype information is considered at different levels. The approaches work for phenotypes at the physiological and disease level (when using representations of complex traits such as dysmorphy, disease, and medical subjects) up to the molecular level (illustrated through use of the Gene Ontology).

bringing together patient phenotypes and patient genotypes

Only now is awareness growing with practitioners involved in genetic diagnosis that phenotype information is crucial for adequate interpretation of the results of genetic screening assays. It is this insight that is embedded in many of the approaches that have been elaborated in this dissertation, and that have been developed within the fruitful collaboration of the bioinformatics departement and the university hospital.

Chapter 2

Array CGH data management and analysis

"Though this be madness, yet there is method in't."

William Shakespeare
Hamlet Act 2, scene 2, 193-206

In this chapter, we start out with a discussion on quality criteria to bring Array CGH as a technology into every day clinical practice. The subject of Array CGH Quality Criteria for Constitutional Genetic Diagnosis is discussed in further detail by Vermeesch *et al* [258]; the main points will be covered here. Next, we elaborate on how this is brought into practice in the Center for Human Genetics at the University Hospital of Leuven. For this, we go into deeper detail on statistical aspects of one specific type of diagnostic work-flow, where loop designs are used. We further illustrate how the data that comes forth from a diagnostics mission of a genetics center, can also be of use in a research setting. We show how patient data, when stored in a database and annotated with design information, can be used for charting copy number variations, an important domain that recently took an important place in clinical genetics research.

This approach has not yet been published, but there currently exists a prototype of both the analysis method as well as the web based interface application. This application, which has Clone Wars as a working title, is illustrated with a number of specific examples. Then, a list of necessary improvements and directions for future work are set out. Next, we continue with a discussion on array repositories themselves, and how they can prove an added value both in a research and a routine diagnostics work flow.

We discuss a number of existing systems, and as a specific case, we describe ArrayCGHBase, a system that was developed by B. Menten of the Medical Genetics department at the University of Ghent within the frame of a research collaboration between Ghent en Leuven, where both groups co-published.

Finally, we briefly discuss how these systems can evolve into intelligent databases that move closer towards expert systems that are able to support diagnosis, as well as support more fundamental genetics research; a discussion that will serve as an introduction to the next chapter, which will elaborate on this theme in much more detail.

2.1 Array CGH quality criteria for constitutional genetic diagnosis

Molecular Karyotyping through Array CGH (comparative genomic hybridization) enables the identification of chromosomal copy number changes. The availability of clone sets covering the human genome opens the possibility for the widespread use of array CGH for both research and diagnostic purposes. This section reports on joint work with the Cytogenetics group at the Center for Human Genetics, Leuven University Hospital [258], identifying parameters that have proven critical for successful implementation of the technology, assessing quality criteria, and discussing the potential benefits and pitfalls of the technology for improved pre- and postnatal constitutional genetic diagnosis. We discuss the main Array CGH quality criteria and some essential caveats regarding basic data handling and analysis. For all lab protocols involved, as well as an in-depth discussion on how these parameters were optimized, the reader is referred to the actual publication [258].

2.1.1 Array CGH in genetic diagnosis

Pre- and postnatal cytogenetic testing is aimed at the genome-wide analysis for the detection of chromosome aneuploidies and segmental aneusomies. Traditionally, karyotyping has performed this role. In addition to its clinical importance, the identification of a chromosomal aberration in specific patients has proven to be a successful way to identify the implicated genes and to gain insight in the pathogenesis of different genetic conditions.

Over the last decades, improved resolution of cytogenetic techniques has lead to a significant increase in the detection rate of chromosomal aberrations in patients with mental retardation (MR) and/or congenital anomalies. However, resolution of traditional cytogenetic techniques is limited to at best 5 megabases (mb), and smaller chromosomal aberrations often remain hidden. The identification of submicroscopic sub-telomeric alterations in 3 - 7% of idiopathic MR patients [67, 123, 227, 193, 197], as well as the sporadic reports of submicroscopic interstitial chromosomal rearrangements, suggests that a substantial portion of idiopathic MR may be caused by smaller chromosomal rearrangements. These observations make it clear that higher resolution screening techniques for the detection of small deletions or duplications at any chromosomal position will drastically increase the elucidation of human genetic diseases.

Although fluorescence in situ hybridization (FISH) has dramatically increased the sensitivity of detection of genomic imbalances, this approach re-

quires prior knowledge of the chromosomal region(s) of interest and therefore is not applicable for whole genome screening approaches required in a diagnostic setting.

More recently, so-called array or matrix CGH (comparative genomic hybridization) utilizes mapped DNA sequences in a microarray format as a platform for the detection of chromosomal deletions/duplications. In this technique, genomic DNA from the patient is labeled with one fluorescent dye while a normal reference sample is labeled with a different dye, and these samples are co-hybridized to the array containing the genomic DNA targets. Chromosomal imbalances across the genome can thus be quantified and their positions defined by analyzing the ratio of the fluorescence of the two dyes along the targets. The resolution of array CGH depends on the size of the genomic fragments as well as on their density. For a more detailed historical review, see Section 1.2. However, due to technical challenges, the technique had not made the transition to clinical practice as of 2004. This section describes the initial exploration of some of the technical aspects essential for optimal results in array CGH, and quality criteria are defined that an array experiment should reach to be reliable in a clinical setting.

2.1.2 Parameters influencing Array CGH quality

The two main technical challenges when performing array CGH are obtaining adequate signal-to-noise ratios and low standard deviations (SD) of the intensity ratios. In contrast to expression arrays, genomic array CGH deals with a vastly more complex probe mixture, and the copy number variations are much lower. In pre- and postnatal diagnostic samples, copy-number ratios vary mainly between 1/2 for deletions and 3/2 for duplications. To obtain optimal signal-to-noise ratios, controlling the fluorescent dye incorporation during labeling, assaying probe quantity, type of slides used, and adequate hybridization skills proved essential.

Because usually only a small subset of the genome is expected to be aberrant in pre- and postnatal constitutional abnormalities, the intensity ratios of patient vs control DNA will be equal for most spots on the array. This offers a unique possibility to measure the quality of an array experiment by determining the SD of all intensity ratios. The log_2 values of these intensity ratios behave in a linear fashion because the relation of the intensities of both dyes inversely influences the value of the intensity ratio at a spot. The normalized log_2 of the intensity ratio values will typically fluctuate around 0. The lower this fluctuation, the better the array CGH experiment. To obtain the lowest SD of the intensity ratios, adequate signal-to-noise ratios are essential as well as an optimized *Cot-1*/probe ratio (See Figure 2.1). Both insufficient Cot-1 and an excess lead to an increase of the SD. The SD of all pooled clones can be further reduced by averaging several independent hybridization experiments (See Figure 2.2).

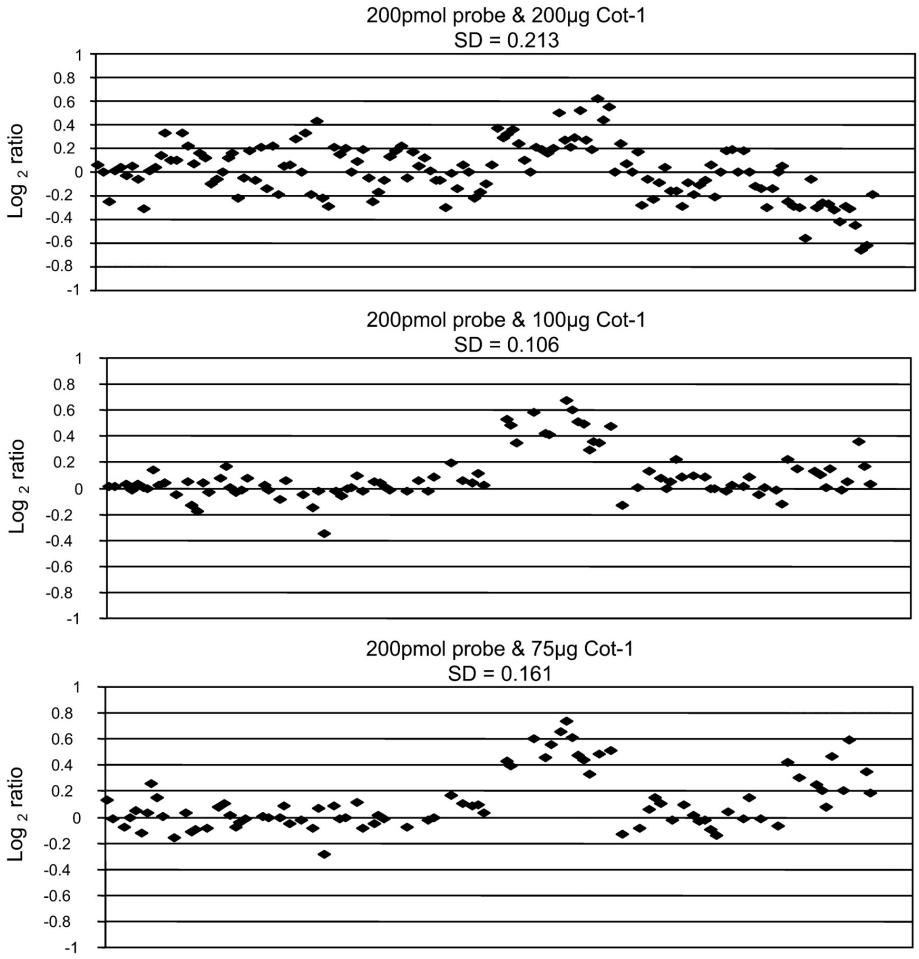


Figure 2.1: Chromosome 9 array CGH ratio profiles using DNA from a patient with a duplication $9q22.33 \rightarrow 31.1$ using different ratios DNA/Cot-1 (indicated at each panel). For each panel, the x-axis represents the clones ordered from the 9p telomere to the 9q telomere according to their clone position in the February 2004 Ensembl freeze. The y-axis marks log_2 transformed intensity ratios at each locus.

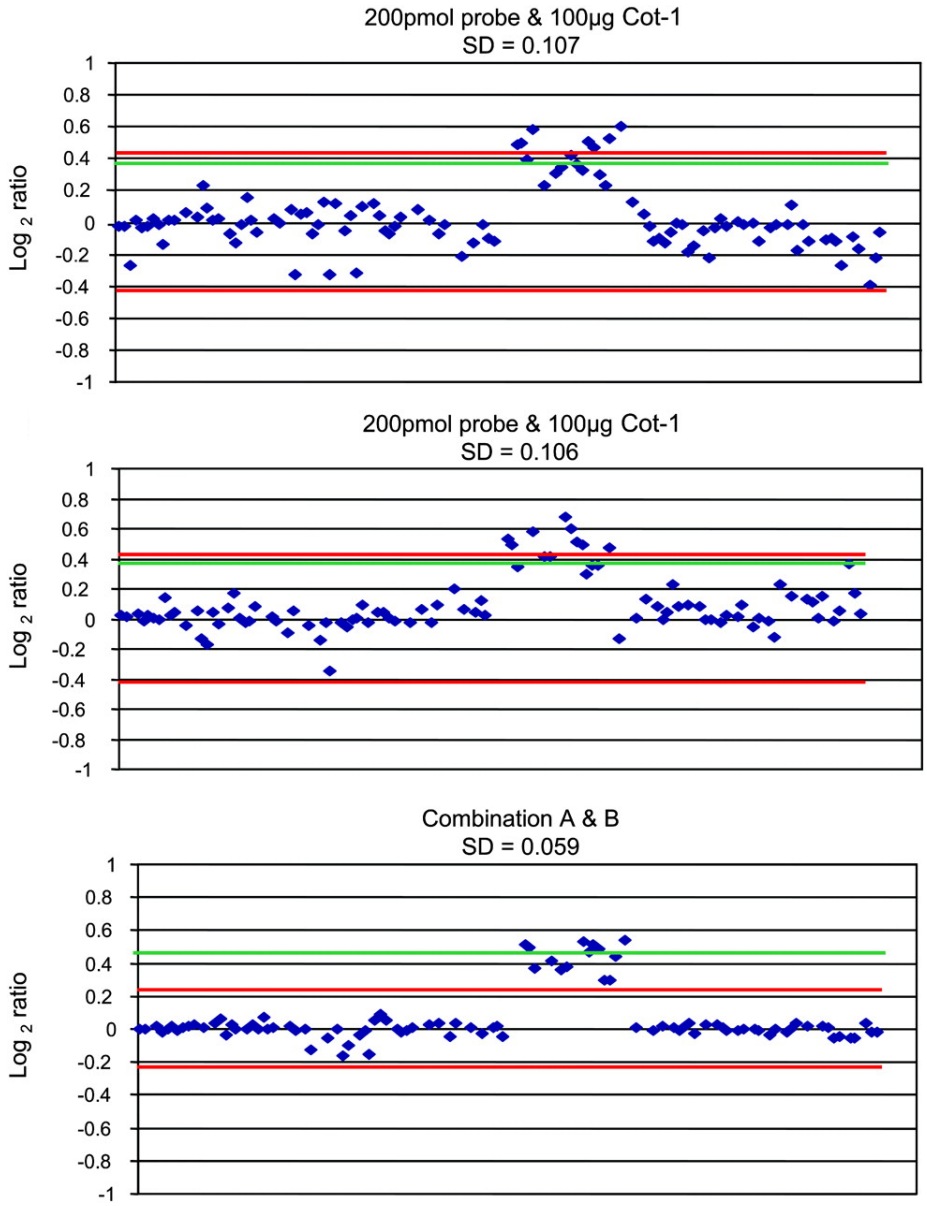


Figure 2.2: Chromosome 9 array CGH ratio profiles using DNA from a patient with a duplication $9q22.33 \rightarrow 31.1$. For each panel, the x-axis represents the clones ordered from the 9p telomere to the 9q telomere according to their clone position in the February 2004 Ensembl freeze. The y-axis marks the hybridization ratio plotted on a log_2 scale. The top two panels represent the analysis of the same sample hybridization but with inverted dye labeling. The bottom panel represents the combined results of both. The red lines indicate the lower thresholds for clone deletion or duplication ($\pm 4 \times SD$) and the green line the upper threshold ($log_2 \left(\frac{3}{2}\right) 2 \times SD$). The y-axis shows the log_2 transformed intensity ratios at each locus; 97.84% of the spots were included in the analysis.

2.1.3 Quality criteria and threshold values

When performing a classical karyotype, the resolution of an experiment is defined by the number of bands that can be discerned. When performing a genome-wide array CGH experiment, in theory, all clones will provide a value. Hence, the resolution of the experiment is defined by the number of genomic fragments on the array as well as by their size. However, in practice, a number of spots fail due to bad quality, *e.g.*, bad spotting or hybridization efficiency. Therefore, it is essential to report the percentage of spots that were amenable to analysis. In the study described in the publication under discussion [258], typically more than 97.5% of the spots provided adequate intensity ratio values, i.e., signal-to-noise over two and the SD among the duplicates less than 10%.

When performing array CGH, the main challenge is to define a threshold value above which no false positives are retained without eliminating true positives (i.e., avoiding false negatives). Since the log_2 transformed normalized intensity ratios fluctuate in a Gaussian fashion around 0, the SD can be used to define thresholds. A threshold level is described as mean plus or minus three or four times the SD [210, 224]. Using three times the SD as a cut-off level, 99.7% of the fragments will fall within the normal range. This would result in about 10 false-positive clones on an array with 3500 different loci. Using four times the SD as a cut-off level, 99.9936% of the fragments fall within the normal range, resulting in one false positive for every four analyses. Therefore, for an array containing 3500 loci, four times the SD is the threshold value of choice.

In addition, using four times the SD as a threshold value defines an important array quality value: the value of the SD of the log_2 transformed normalized ratios. To be applicable in constitutional diagnosis, the value of four times the SD needs to be below the detection limit of an autosomal deletion or duplication. Because the ratio of a duplication (3/2) is closer to a normal ratio than the ratio of a deletion (1/2), four times the SD needs to be below the detection limit of a duplication. This threshold can be defined as the log_2 transformed mean intensity ratio of the duplicated loci minus two times the SD, or:

$$4 \cdot SD \leq log_2 \left(\frac{3}{2} \right) - 2 \cdot SD \tag{2.1}$$

or

$$SD \leq 0.096 \tag{2.2}$$

2.1.4 Polymorphisms

While four times the SD seems to be an adequate cut-off level based on statistical grounds, this reasoning assumes a perfect Gaussian distribution of intensity ratios and neglects the biological and experimental causes of non-Gaussian behavior of a (small) subset of genomic fragments.

Polymorphisms can be identified on arrays as those clones for which the intensity ratios of independent experiments repeatedly fall outside the above-

defined cut-off level. Identifying and reporting these polymorphic clones is thus an essential first step toward proper interpretation of array CGH data. When abnormal values are obtained in a number of separate experiments either using DNA from individuals without an obvious clinical phenotype or in different experiments with DNA from individuals with very different clinical phenotypes, it can be assumed that these clones are polymorphic. For a detailed discussion on the analysis of aberrant signal intensity ratios that point to putative polymorphic loci, see the manuscript [258].

Polymorphic loci will not only be detected by intensity ratios repeatedly surpassing the four times SD threshold value, but when the polymorphisms are embedded within a single clone on the array, they will be evidenced as smaller deviations from the average intensity ratio. Indirect evidence for such polymorphisms comes from experiments with a low SD. A low SD results in a low threshold level which, in turn, results contradictory in many more DNA fragments that are (false) positive for a deletion or duplication than in a less optimal experiment with a higher SD.

2.1.5 Mosaicisms

When using DNA from patients carrying mosaic segmental or chromosomal aneuploidies using array CGH, none of the data points derived from the aneuploid loci may reach the threshold levels defined above. However, a series of intensity ratios will deviate from the mean value in a similar direction. Visual inspection of array CGH data can rapidly recognize the presence of non-random spreading of a series of flanking data points. DNA from a 100% trisomy-13 cell line DNA and a mixture containing only 20% of the trisomy-13 cell line DNA was hybridized vs DNA from a normal cell line. In Figure 2.3, the presence of a trisomy 13 in 20% of all cells can be identified by viewing the intensity ratio values of all the chromosome 13 loci, which are all unidirectionally above the mean intensity ratio value. Molecular karyotyping can thus, in analogy with conventional karyotyping, readily be used to identify low-grade mosaicisms of at least 8% [158]. However, smaller-sized mosaic segmental duplications or deletions that span only a few clones on an array will inevitably not be detected.

2.2 Discussion on quality criteria

Because both genome-wide array CGH analysis and conventional karyotyping, based on staining chromosomes, aim to identify chromosomal aberrations by screening the genome, it is proposed to call this technology molecular karyotyping. Molecular karyotyping is likely to become part of the routine genetic diagnosis and to replace in part current karyotyping technologies. Advantages over conventional karyotyping include a higher resolution, direct mapping of aberrations to the genome sequence, amenability to automatization and quality control procedures and, probably, higher throughput and shorter reporting times. Essential for the successful introduction of the technology in current

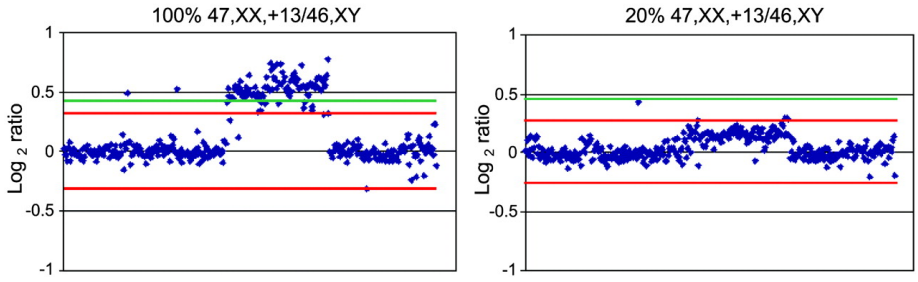


Figure 2.3: Trisomy mosaicism Array CGH ratio profiles using DNA from a mixture containing 20% DNA from a trisomy 13 cell line and 80% DNA from a normal female cell line vs DNA from a normal male cell line. The figure shows a partial karyotype enlarging the ratio profiles obtained for chromosomes 12, 13, and 14. Red lines indicate the lower thresholds for clone deletion or duplication ($\pm 4 \times SD$) and the green line the upper threshold ($log_2\left(\frac{3}{2}\right) - 2 \times SD$). The y-axis shows the log_2 transformed intensity ratios at each locus; 97.84% of the spots were included in the analysis.

cytogenetic laboratories are (a) robust protocols and a good understanding of the critical technical factors, (b) quality criteria which define a successful array experiment, and (c) reporting guidelines to enable correct interpretation of the results obtained by different laboratories that includes a good knowledge of polymorphic loci in the human population.

2.2.1 Parameters influencing Array CGH quality

While array formats may differ in that they may use BACs [232, 178], cDNAs [179], oligos [141] or PCR fragments [150] as targets, the main steps involved in an array experiment are similar and include array production, probe preparation, hybridization and washing, and data analysis. When comparing different described protocols, a large variation in the details of the method can be discerned [40]. Whereas all aspects of the protocol are important for optimal array CGH data, the DNA/Cot-1 ratio as well as the efficiency of fluorescent dye incorporation were identified as the main causes of inter-experimental variation. Batch-to-batch variability of commercial Cot-1 preparations has been previously pinpointed as the cause of variability among different experiments [40]. The fluorophore incorporation efficiency (specific activity) has a direct effect on the sensitivity of an experiment and is thus an obvious key element for a successful array experiment.

2.2.2 Quality criteria

In analogy with conventional karyotype, a molecular karyotype should have a well-defined quality. However, the few reports at the early stage of the introduction of array technology that use array CGH for the detection of constitutional

chromosomal imbalances lacked uniformity in addressing the quality of an array and identifying appropriate threshold values [230, 260, 206, 208, 224]. First, the number of clones on the array with successful hybridizations should be reported. Second, the SD of the log_2 transformed intensity ratios should not exceed 0.096. Third, a minimum threshold can be defined as plus or minus four times the SD of all the log_2 transformed intensity ratios of the normal clones. To define the SD of the clones with normal intensity ratios, a reiterative process to define the threshold can be used. First, the SD is calculated using the values of all clones and, subsequently, the values of the clones surpassing the threshold are eliminated and the SD is recalculated. Finally, a higher threshold value was defined as $log_2\left(\frac{3}{2}\right) - 2 \cdot SD$. If the intensity ratio of only a single locus surpasses the lower threshold, it should also surpass the higher threshold to trigger confirmatory experiments. However, if two flanking loci surpass the lower threshold, this equally triggers confirmatory experiments.

2.2.3 Pitfalls for pre- and postnatal diagnosis: polymorphic loci

Variations in chromosome lengths and banding are well known to the cytogenetic community. Common chromosomal variants have been documented over the last 35 years and practitioners have good knowledge of the most common benign variants. Novel benign variants are still being uncovered. Polymorphisms may extend several megabases [92, 235, 152], and thus multiple array loci. Not surprisingly, at the higher resolution level obtained by molecular karyotyping, similarly polymorphic loci are detected and, due to the higher resolution, the number of variants that are observed has equally increased. The first reports using whole genome array CGH in constitutional cytogenetic analyses have identified, respectively, 2/20 and 5/50 single clone anomalies in patients that were also present in their phenotypically normal parents [260, 224]. During the analysis of 30 patients at the Center of Human Genetics described in the manuscript that corresponds to this section of the dissertation [258], it was estimated that a total of 18 loci (or 0.5% of all loci on the array) within the clone set were polymorphic as defined by intensity ratios in different independent experiments that are concordant with a deletion or duplication of a locus. However, proof that these loci are truly polymorphic sites still awaited confirmation at that time. While currently a polymorphism is described as a benign genetic variant, it seems likely that the distinction between benign polymorphisms and disease-causing genomic alterations will become blurred. Segmental duplications without phenotypic alterations have been detected. Using a high-density molecular karyotype, Sebat *et al.* [213] detected many copy number alterations in the population, some of which have previously been linked or associated with various phenotypes such as predispositions for cancers or neurological diseases. In addition, it is likely that smaller intra-clonal variations and polymorphisms also exist. Albertson and Pinkel [7] reported on the variation of the intensity ratios of a single spot derived from DNA on chromosome 6 containing the apolipoprotein. They suggest that this variation is due to the known polymorphisms within this gene.

Indirect evidence that more intraclonal polymorphisms exist comes from our experiments with very low SD. In such optimal experiments the low cut-off levels result in many more DNA fragments that are (false) positive for a deletion or duplication than in a less optimal experiment with a higher SD. However, such non-Gaussian behavior of certain spots could also have other causes such as suboptimal printing or hybridization efficiencies.

2.2.4 Discussion

Over the five years that have passed since this early publication, polymorphic clone information has become integrated in the genome annotations, as was predicted in the manuscript. Indeed, various large-scale concerted efforts have been set up to address the question of how such polymorphic sites in the genome can be ascertained. As predicted, continued feedback from a series of dedicated laboratories is leading to validated databases of candidate or proven genomic polymorphisms, such as the database of Genomic Variants (See Section 2.4.2 for an elaborate discussion).

To provide a frame of reference for this early work on polymorphisms, the list of putative polymorphic clones published in the manuscript underlying this section of the dissertation [258] was cross-referenced with the current public body of available knowledge on clone polymorphisms. Of the 18 clones listed as probably polymorphic, RP6-22D12 could no longer be mapped to chromosome 3 and was excluded from this analysis. Three clones that couldn't be mapped at the time were excluded as well. Well names such as RPNN-XXYZZ are not well suited as unique clone identifiers since it is not always clear what clone set is referenced. Additionally, with each *genome build* and database release, small and even large alterations in mapped position can occur. Sometimes, clones are even mapped to an entirely different chromosome after a new genome assembly effort. For the 14 remaining clones, it was verified whether they overlap with currently known polymorphic loci (cross referencing with the list of copy number variations listed in the seminal Redon *et al.* paper [190]). It was also verified whether segmental duplications lie in the vicinity of the clone mapping position. For this, the database of Segmental Duplications from the University of Washington was used. This was done because segmental duplications are believed to be a major catalyst of large-scale variation in the human genome [220]. Segmental duplications themselves are also found to be significantly enriched within regions of copy number polymorphism. Finally, it was also verified whether known patients from the DECIPHER database overlapped the clone under investigation. The results are shown in Table 2.1. From this table, we learn that a fourth of all investigated clones neither show the presence of segmental duplications nor any known polymorphic loci in the Redon set. Of the four clones that have not been found to overlap with known variable loci, still, THREE correspond to existing DECIPHER patients. This implies the conclusions in this early work have proven valid in the light of the current state of the art of genome variance databases.

Table 2.1: Cross-referencing early results on polymorphism by Vermeesch *et al.* [258] with the current knowledge on copy number variant loci. Each reporter is cross referenced with the list of copy number variations from the Redon *et al.* paper [190] (number of gains as well as losses), and checked for the presence of segmental duplications inside or in the vicinity of its mapping position, through the database of Segmental Duplications from the University of Washington. Finally, it was verified whether known patients from the DECIPHER database overlapped the clone under investigation.

Clone	Chromosome	Redon Gain	Redon Loss	Redon	WashU	Decipher
RP5-1108M17	1	142	32	YES	yes	1
RP11-294I20	2	0	0	-	Yes	1
RP11-24H13	4	27	0	YES	No	0
RP11-6C14	4	0	0	-	No	0
RP11-17M8	8	0	6	YES	No	0
RP11-356M23	8	0	6	YES	Yes	0
RP11-292F22	10	2	253	YES	Yes	0
RP11-114K7	11	0	0	-	No	1
RP11-161M6	16	3	61	YES	No	0
RP11-274A17	16	112	14	YES	Yes	1
RP11-342F21	17	0	0	-	No	1
RP11-374N3	17	56	0	YES	Yes	7
RP11-537N4	19	0	3	YES	Yes	0
CTA-299D3	22	0	0	-	Yes	6

2.2.5 Back to the future

We have evaluated an early list of polymorphic loci and investigated them in the light of current knowledge on copy number variation in the human genome. At the time the paper was published, molecular karyotyping was defined as a genome-wide array CGH experiment, but no resolution has been defined. The minimum resolution of a molecular karyotype was said to be at least equal but preferentially to surpass the resolution obtained by conventional karyotyping. It was stated that the maximum resolution that could be obtained using BAC- or PAC-based clone arrays would be a genomic tiling path array of 32.000 targets, hence covering the entire genome one-fold, which had already been achieved at that time [111]. Even higher resolutions were obtained by using arrays with overlapping BAC/PAC clones [111], smaller insert fragments [36], PCR products [150], or oligonucleotides [141]. Although molecular karyotyping at higher resolution was predicted to lag behind in clinical adoption, currently, commercial platforms are available and are making their entry into the diagnostic setting. Statistical analysis to extract valid clinical features from raw log ratios or individual reporter copy number calls are still an issue under investigation, and while both research and commercial software platforms are available to perform this initial data analysis, the results remain subject of intense discussion. Even if standards would be defined on how to perform the raw data analysis in an uniform way so that assays of this type become robust enough to be amenable to use in clinical practice reliably, there still remains the issue of clinical interpretation, or the hunt for meaning and relevance of copy number variations in a patient sample. Indeed, with the increasing resolution of array CGH assays, whether it be clone, oligo or loss of heterozygosity based, data mining, data integration, intelligent LIMS and expert systems will be needed for diagnosis support. While the remainder of this chapter will focus on statistical data analysis of raw array CGH data (in the context of a loop design) and on initial attempts at intelligent database and analysis systems for array CGH experiment results, the next chapter of this dissertation will focus on bringing these genotype results into a diagnostic context.

Considering a significant degree of polymorphic clones in the genome, the identification and validation of these polymorphic loci has been, and will remain to be a daunting task. Possibly, novel improved approaches to identify polymorphic loci may resolve this issue. Finally, low copy repeats, often containing small copy number variations, make up 5 – 10% of the genome. The complex nature of genome architecture will be further elucidated as the resolution of genetic screens increase and prices per assay decrease.

Finally, one caveat on array CGH analysis still stands today. Molecular karyotyping is likely to replace, in part, current karyotyping technologies based on staining chromosomes both in pre- and postnatal diagnosis. In addition to the many advantages, molecular karyotyping also has some caveats as compared with conventional karyotyping, *e.g.*, it fails to identify balanced translocations and ploidy variations. Because arrays rely on the occurrence of genomic copy number differences between patient and control samples, balanced transloca-

tions, which by definition do not have genomic losses or gains, cannot be detected. Also, ploidy variations are likely to escape detection by array CGH as the technique relies on normalization of the intensity ratios. Double-dye intensity derived from a triploid DNA sample would subsequently be normalized and thus not be detected. Because of these limitations, it seems likely that we will continue to enjoy the view of banded chromosomes in the foreseeable future, and that banded chromosomes will remain an invaluable tool in the genetic diagnostic laboratory.

2.3 A loop design analysis for Array CGH data

2.3.1 Introduction

Comparative genomic hybridization microarrays for the detection of constitutional chromosomal aberrations constitutes the application of microarray technology most rapidly being adopted into routine clinical application. Through genotype-phenotype association, which is more thoroughly discussed in Chapter 3, it is also an important technique towards the discovery of disease-causing genes and genome-wide functional annotation in man.

When using a two-channel microarray of genomic DNA probes for array CGH, the basic setup consists in hybridizing a patient against a normal reference sample. Two major disadvantages of this setup are (1) the use of half of the resources to measure a reference sample, which is in itself not or of only limited informative value, and (2) the fact that deviant signals possibly can possibly be caused by benign copy number variation in the "normal" reference sample instead of a real aberration in the patient genome.

To account for these shortcomings of the classical two way approach, we propose an experimental loop design that compares three patients in three hybridizations. We develop and compare two statistical methods (linear models of log ratios and mixed models of absolute measurements). In an analysis of 27 patients seen at the Center for Human Genetics, Leuven, the loop design setup in combination with the linear model analysis overcomes the limitations of the classical setup. Furthermore, we observed that the linear models of the log ratios are advantageous over the mixed models of the absolute intensities. These advances contribute to the quick adoption of array CGH as a routine diagnostic tool. They lower the detection limit of mosaicisms and improve the assignment of copy number variation for genetic association studies.

For the statistical analysis, we consider two novel approaches for array CGH: linear modeling of the log ratios and mixed modeling of the absolute signal intensities. We compare both methods on a test data set consisting of 27 patients (9 loop designs) and we implement the method with the best signal-to-noise ratio as a user-friendly web application.

Experimental designs that make the best use of available resources are essential for the widespread adoption of array CGH as a routine clinical tool for the diagnosis of constitutional chromosomal aberrations. Reduction of false posi-

tives and negatives guarantees the best service to the patient and the best use of economic resources, which are key factors in a clinical environment. Furthermore, correct assignment of "benign" copy-number variations to the right sample will be important in upcoming studies of association between copy-number variations and disease. Finally, this design will also lower the detection limit for mosaicism (i.e., chromosomal aberrations present in only a percentage of the cells).

Array CGH for the detection of congenital chromosomal aberrations is also a key method for the genome-wide discovery of gene function. Patients with chromosomal aberrations provide a natural form of forward genetics screening. Through genotype-phenotype correlations, phenotypes can be associated to chromosomal regions containing only a few to a few tens of candidate genes [248]. Further prioritization of candidate genes using bio-informatics approaches and validation in small animal models (for example, zebrafish or fruitfly) allows rapid identification of disease causing genes [5].

2.3.2 Acknowledgements and credits

The methods, validation and implementation into clinical practice are the result of a close collaboration between ESAT and CME. Joke Allemeersch [1] is credited for evaluation, comparison and implementation of the statistical models. Femke Hannes [2] is credited for the validation on real patient data. Bert Coessens [3] later added a module for normalisation of regular, two-way designs. A description of their efforts is included into this section in order to offer a complete view on the approach; in scope of this dissertation, we will focus explicitly on the web architecture that was created to leverage the results of both the theoretical work and statistical script prototypes of Joke Allemeersch, and the validation efforts of Femke Hannes, into a web based platform as is central to the scope of Array CGH Data Management and Analysis of this dissertation. The final method was implemented as a web application and is available at www.esat.kuleuven.be/loop. It is currently being used by different genetic testing centers across Europe. All microarray data are publicly available in GEO (GSE6538).

2.3.3 A change in experimental setup

Array Comparative Genomic Hybridization or *molecular karyotyping*, as introduced in chapter 1.2, detects copy-number aberrations and variations at high resolution on a genome-wide scale and has been applied to detect chromosomal imbalances in patients with congenital anomalies and mental retardation. Array CGH is a highly effective technique that is entering routine clinical use much faster than other microarray technologies. The most frequent experimental setup for array CGH consists in comparing genomic DNA of a patient (test)

[1] ESAT, K.U.Leuven; currently working at the V.I.B. Microarray Facility
[2] CME (Center for Human Genetics), Gasthuisberg, Katholieke Universiteit Leuven
[3] ESAT, K.U.Leuven

with that of a normal individual (reference) using a two-channel microarray consisting of DNA segments spread across the whole genome. In the case of the clinical platform used as a basis for the work described in this section of the dissertation, the DNA segments consist in PCR-amplified BAC [4] clones. However, the discussion applies equally to spotted long oligo [5] platforms. So, we will refer to our probes as *reporters*. DNA from the test and reference samples is extracted, labeled with different fluorescent dyes (usually Cy3 and Cy5), hybridized to the microarray, and then scanned by two-channel laser scanner. Aneuploid chromosomal regions are detected as probes with a deviant log ratio of the intensities of the test against reference signal (approximately $\log_2(1/2)$ for a deletion and $\log_2(3/2)$ for a duplication). Usually the experiment is repeated in a dye-swap with the fluorescent labeling of test and reference exchanged. The signals are then averaged over the dye-swap replicates to reduce the signal-to-noise ratio.

The new design proposed here is a loop design in which three hybridizations are carried out with three test patients that are compared with each other: Patient 1 versus Patient 2, Patient 2 versus Patient 3, and Patient 3 versus Patient 1, as shown in Figure 2.4.

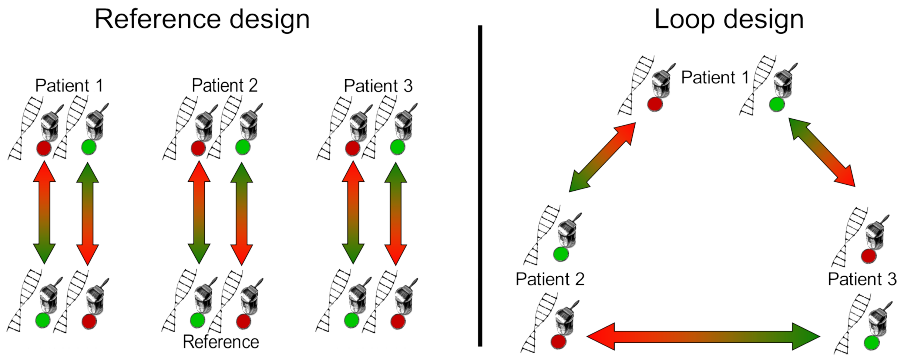


Figure 2.4: The loop design. Schematic overview of a reference and a loop design, in which three patients are compared. In the reference design, the three patients are compared to the DNA of a normal individual (reference). In the loop design the three patients are compared two-by-two.

This design measures the intensities of three test samples in a statistically balanced way and requires no normal reference sample. Hence, only three arrays are used to analyze three patients and to obtain two measurements for each of them. For the classical dye-swap design, half of the resources are consumed to measure the reference sample of a normal individual and, therefore, six arrays would be necessary to obtain as many measurements from the test samples.

[4] DNA vector into which relatively large DNA fragments can be inserted and cloned in a bacterial host, used here as reporters on microarrays

[5] A short DNA probe whose hybridization is sensitive to a single base mismatch

Variation

Normal individuals show an extensive genomic variability called copy number variation (CNV) [190, 213, 109]. The extent of this variation is surprisingly high, covering at least 10% of the genome, and likely to have major implications for human variation and disease. In the classical dye-swap design, a deviant log ratio for one reporter in the test sample could just as well be associated with a variation in the reference sample. The difficulty in disambiguating deviations between the test and reference sample prevents us also from replacing the reference sample with a second test sample in the dye-swap design. The loop design, on the contrary, unambiguously associates a deviation to the correct sample by looking for a unique pattern of log ratios. For example, a duplication in Patient 1 will be associated approximately to a positive log ratio in the Patient 1 vs. Patient 2 hybridization, a negative log ratio in the Patient 3 vs. Patient 1 hybridization, and a null log ratio in the Patient 2 vs. Patient 3 hybridization. No deletion or duplication in another patient will display the same pattern, so the association is unambiguous.

2.3.4 New statistical models for loop design analysis

Two main philosophies vie for dominance for the statistical modeling of microarray signal intensities: (1) linear models of log ratios of intensities and (2) mixed models of absolute intensities. Two-channel microarrays were originally developed so that taking the ratio between the Cy3 and Cy5 intensity of a spot would eliminate multiple sources of variations (in particular, the amount of DNA material per spot). Linear models of log ratios keep in track with this philosophy and extend it by observing that log ratio measurements from different hybridizations containing equivalent samples are interdependent. They formulate these dependencies as a set of linear relations that are then inverted to obtain tighter estimates for a smaller set of independent statistical effects (which remain in essence similar to log ratios). By contrast, mixed models of absolute intensities aim at disentangling the signals from the Cy3 and Cy5 channels by expressing the intensities as the sum of an extensive set of fixed and random effects dependent on many factors that systematically affect the microarray measurements (dye effect, array effect, spot effect, etc.). Although the mixed models we consider are linear mixed models and thus also linear models, we will for simplicity refer in this section to the two classes of methods as being "linear models" vs. "mixed models".

Mixed model of absolute signal intensities We will apply a mixed model along the lines proposed by [267]. These models were originally applied to cDNA microarrays, but they can be tailored to the analysis of array CGH. Before applying a mixed model, we want to correct hybridization signals for possible spatial effects (which cannot be easily corrected with the mixed model). We therefore apply a 2D spatial LOESS correction to the hybridization intensities and obtain LOESS corrected log ratios. The mixed models proposed by Wolfinger [267]

consist of two successive models: the normalization and the reporter-specific model. The normalization model corrects for array, dye, and patient effects. The fitted model can be written as follows:

$$y_{cij} = \mu + P_i + A_j + \mathrm{PA}_{ij} + r_{cij},$$

where y_{cij} are the Cy3 and Cy5 intensities, μ is the overall average, P_i is the fixed patient effect with three levels (i = Patient $1, \ldots, 3$), A_j estimates the random array effect (also with three levels j = Array $1, 2$, or 3), and PA_{ij} fits the interaction effect between the patient and array effect, and in this way it also corrects for the dye effect. For each reporter, we extract the residuals r_{cij} from the normalization model and fit a reporter-specific model

$$r_{cij} = C_c + \mathrm{CP}_{ci} + \mathrm{CA}_{cj} + \varepsilon_{cij},$$

where r_{cij} are the residuals obtained from the normalization model, C_c is the overall average for Reporter or Clone c, CP_{ci} is the fixed patient effect for Reporter c with three levels (i = Patient $1, \ldots, 3$), CA_{cj} estimates the random spot effect for Reporter c (also with three levels j = Array $1, 2$, or 3), and ε_{cij} fits the random error effect. Our main interest is in the estimates of the CP_{ci} effects, which reflect the difference between the patients for Reporter c. Specifically, we assess whether the contrasts Patient 2 vs. Patient 1 ($= \mathrm{CP}_{c2} - \mathrm{CP}_{c1}$) and Patient 1 vs. Patient 3 ($= \mathrm{CP}_{c1} - \mathrm{CP}_{c3}$) are equal to zero with a Wald's F-test. In the case where the contrast is significantly larger than zero for a chosen significance level α, we call this contrast *positive*. In the case where it is smaller than zero, it is called *negative*. Else we assign 0. Based on this hypothesis testing, the reporters are classified as duplicated or deleted according to the classification shown in Table 2.2. For example, if the contrast Patient 1 vs. Patient 3 ($= \mathrm{CP}_{c1} - \mathrm{CP}_{c3}$) is positive and the contrast Patient 2 vs. Patient 1 ($= \mathrm{CP}_{c2} - \mathrm{CP}_{c1}$) is negative for a reporter, then this reporter is likely to be duplicated for Patient 1. In some rare cases, we obtain as result a reporter that has, for example, a negative value for both contrasts $\mathrm{CP}_{c1} - \mathrm{CP}_{c3}$ and $\mathrm{CP}_{c2} - \mathrm{CP}_{c1}$, which is none of the combinations in Table 2.2. In this case, we call the reporter *strange*.

Linear model of log ratios An alternative statistical tool is a linear model of the log ratios, as proposed by Smyth *et al.* [229]. In contrast to the linear mixed model, this technique fits the 2D spatial LOESS corrected $\log_2$ ratios directly. In this particular case, we can choose the following contrasts $\mathrm{C}_{c1} = \log_2(P_{c1}/P_{c3})$ and $\mathrm{C}_{c2} = \log_2(P_{c2}/P_{c1})$, where P_{ci} corresponds to the true underlying signal for Reporter c for Patient i. These contrasts correspond to the samples that were directly compared on the first two slides in Figure 2.4 and the observed log ratios on these slides should on average be equal to the contrast. The data of the third slide should then correspond on average to $\mathrm{C}_{c3} = \log_2(P_{c3}/P_{c2}) = -C_{c1} - C_{c2}$. The linear model that fits the data can be written as

$$E\begin{pmatrix} y_{c1} \\ y_{c2} \\ y_{c3} \end{pmatrix} = X\mathrm{C}_c = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} C_{c1} \\ C_{c2} \end{pmatrix},$$

Table 2.2: Classification of the reporters. Each reporter is classified as subject to a gain (*positive*), loss (*negative*), or normal copy number (0) for two chosen contrasts, i.e. Patient 1 versus 3 and Patient 2 versus 1. Based on the results a reporter can be recognized as duplicated or deleted for Patient 1, 2, or 3, according to this scheme.

Classification	Log Ratio patient_1 / patient_3	Log Ratio patient_2 / patient_1
Duplication for patient_1	positive	negative
Duplication for patient_2	0	positive
Duplication for patient_3	negative	0
Deletion for patient_1	negative	positive
Deletion for patient_2	0	negative
Deletion for patient_3	positive	0

where E denotes the expectation of a random variable, X is the matrix of linear dependencies, C_c is the vector of contrasts for Reporter c, and y_{ci} denotes the $\log_2$ ratio for Reporter c measured on the ith slide. For each reporter, the least squares estimates of the three contrasts are obtained. To classify the contrasts as significantly gained or lost, or not copy number altered, we apply the moderated t-statistic as implemented in the BioConductor package `LIMMA`, which implements linear models for microarray data analysis [229]. The p-values from the moderated t-test were corrected to control the false discovery rate with Benjamini-Hochberg [20]. Similarly to the mixed model, we can detect reporters that are duplicated or deleted for a patient, based on the p-values of the contrasts. For a chosen cut-off value α, we decide whether a reporter is not copy number altered (0), gained (*positive*), or lost (*negative*) for a contrast. Based on the two contrasts, we can again classify a reporter as duplicated or deleted for a patient according to Table 2.2. Again, we can on rare occasions obtain *strange* reporters.

2.3.5 Methods

Array CGH

A 1 Mb resolution BAC array was performed as described by Menten *et al.* [158]. In short, 3500 BAC en PAC clones from the Welcome Trust Sanger Institute were amplified by two rounds of DOP-PCR [63]. The purified amino-linked PCR products were spotted in duplicate on 3-D CodeLink Bioarray System slides (Amersham Biosciences, Piscataway, NJ). 150 ng of patient DNA was labeled by random prime labeling system (BioPrime Array CGH Genomic Labeling System, Invitrogen) using Cy3- and Cy5-dCTPs (Amersham Biosciences). Probe concentration and labeling efficiencies were measured using the nanodrop ND-1000 spectrophotometer (Nanodrop Technologies, Rockland, DE). The probe was placed on the slide under a glass cover slip (24x24 mm). The slides were

incubated for 42-72 hours at 37°C under humified conditions. Post hybridization washing was performed by soaking in 1xPBS for 10 minutes and followed by a stringency wash on 42°C for 30 minutes in 50% formamide 2x SSC solution and 10 minutes in 1x PBS at room temperature. To finalize, the slides were dried by centrifugation for at least 1 minute.

Image and data analysis

After the washing, a two-channel scan was performed with an axon laser scanner GenePix 4000B (Molecular Devices, Union City, CA) at 532 nm and 635 nm using GenePixPro 6.0 program. The results were converted into GPR format and were directly uploaded to the web-based application. The median spot intensities were corrected with the local median background, and only those spots with a signal above background (*i.e.*, foreground intensity larger than local background intensity plus twice the local standard deviation of the background) were retained for the analysis. In this way, only a few spots are lost, as almost all spots are above background (on average 96.6%). The ratios of the Cy5 to the Cy3 intensities were computed for each reporter and base 2 log transformed. The log ratios are normalized using a 2D spatial LOESS normalization, in which one applies a LOESS regression to fit the $\log_2$-ratios (M-values) on the coordinates on the slide as predictor variables. All data have been uploaded in GEO, with Accession Number GSE6538.

Benchmark data set

For the comparison of the analysis approaches, we consider a data set consisting out of nine loop designs or 27 patients with mental retardation (MR) and multiple congenital anomalies (MCA). The patients were seen at our genetics center (Center for Human Genetics, U.Z.Leuven). Conventional karyotyping showed chromosomal imbalances in 11 patients. Analysis of the patients was carried out with a 1Mb BAC array. In the first instance, the data analysis strategy that we had previously developed was applied to this data. In this procedure, a region is called aberrant, if one clone passes the threshold of $4 \times SD$ and if two or more flanking clones were passing the threshold of $\log_2\left(\frac{3}{2}\right) - 2 \times SD$ as described by Vermeesch *et al.* [258], and discussed earlier in this chapter. If a deletion or a duplication larger then 3Mb was detected, FISH was performed to confirm the results of the array. In case of a duplication smaller than 3Mb, we performed quantitative PCR (qPCR) [158]. Out of 27 patients, 16 show one or multiple clone anomalies, whereas 10 patients are apparently normal, at least according to the results of the array. One of the patient was a carrier of an inversion, which cannot be detected by array CGH, but can be with conventional karyotyping; as such this patient did not contribute data to the benchmark. A short summary of the data set is shown in Table 2.3.5. In total, this data set comprises 635 aberrant clones: 274 deleted and 361 duplicated clones.

Experiment	Patient	Deletions/duplications	length	Confirmed with
1	1	deletion on 13	25	karyotyping
	2	duplication of X (sex mismatch)	149	karyotyping
2	1	duplication of 18	102	karyotyping and FISH
3	1	deletion on 10	6	
	2	duplication on 7	20	
	3	duplication on 15	43	
4	2	deletion on 4	15	FISH
	3	Inversion (p11-q13)		karyotyping
5	1	deletion on 12	14	FISH
6	1	deletion on 9	6	
	3	deletion on 12	7	
7	2	duplication on 5	13	
	2	deletion on 18	16	
8	1	duplication on 13	15	FISH
	1	deletion on 13	12	FISH
	2	balanced translocation (2;6)(q33,1;p23)		karyotyping
9	1	duplication on 7	11	
	1	deletion on 7	15	
	2	deletion of X (sex mismatch)	158	karyotyping
	3	duplication on 21	8	

Table 2.3: Loop design test data set. The nine loop design experiments, listed in the table, will be used as a test data set to compare the two methods, linear model and mixed model. For each aberration present in the experiment the number of deleted or duplicated reporters (length) is indicated.

2.3.6 Statistical models: a discussion

The mixed model and linear model provide two distinct ways to analyze the loop design experiments. To decide which method is preferable, we will first check which estimation method best separates the aberrant from the non-aberrant reporters. This will already give an indication to which method is preferred. Secondly, we will compare the false positive and false negative rates for a chosen set of cut-off values α. Based on this information, we will decide which method to use and choose a cut-off value to call a reporter significantly duplicated or deleted.

For the comparison of the analysis approaches, we consider a data set consisting out of nine loop designs from patients seen at our genetics center (Center for Human Genetics, U.Z.Leuven; see Materials, Data, and Methods). Fifteen of the 27 patients involved in the nine loop designs showed one or more confirmed deletions or duplications of a chromosomal segment. Two experiments (Experiments 1 and 9) in our test data set include a sex mismatch. As for these experiments, the Y chromosome is absent for at least one of the patients, the measurements on the Y chromosome were excluded for both experiments from all computations. Because the X chromosome has regions with chromosome Y homology, the intensity ratios of chromosome X reporters are also more variable, and hence the X chromosome was also excluded from the computations for Experiments 1 and 9. In total, this data set comprises 328 aberrant reporters: 116 deleted and 212 duplicated reporters. Over all 9 experiments, we have a set of 30.668 measurements for non-aberrant reporters.

Signal-to-noise ratios

Assessing which method is best capable of distinguishing between the intensities of aberrant and non-aberrant reporters can be done by computing *signal-to-noise-ratios* (SN) (for both deletions and duplications separately) as

$$\mathrm{SN}_{\mathrm{dupl/del}} = \frac{|\mathrm{mean}_{\mathrm{dupl/del}} - \mathrm{mean}_{\text{ non-aberrant}}|}{\sqrt{\frac{1}{2}\left(\mathrm{var}_{\mathrm{dupl/del}} + \mathrm{var}_{\mathrm{non\text{-}aberrant}}\right)}}.$$

As we have collected a data set with 212 duplicated reporters, 116 deleted, and 30.668 non-aberrant reporters, we can compute the SN values based on the absolute values of the contrasts Patient 1 vs. Patient 3 and Patient 2 vs. Patient 1, for both the linear model and the mixed model. The results are shown in Table 2.4. As for deleted clones the difference between their log ratios ($\pm\frac{1}{2}$) and log ratios of non-aberrant clones is larger than for the log ratios derived from duplicated clones ($\pm\frac{2}{3}$), the signal-to-noise-ratio is of course larger for the deleted reporters than for the duplicated reporters. The linear model leads to a significant reduction in the noise, especially for the non-aberrant reporters, and this results in a larger signal-to-noise-ratio. Therefore, these statistics are favorable to the linear model.

Table 2.4: Signal-to-noise-ratios. For both methods (i.e., the mixed model and the linear model), the number of reporters, average and standard deviation of the $\log_2$-ratios are given for the non-aberrant, duplicated, and deleted reporters. Based on these numbers the signal-to-noise-ratios are computed.

	Mixed model	Linear model
$N_{\text{non-aberrant}}$	30668	30668
$\text{mean}_{\text{non-aberrant}}$	0.00701	-0.00064
$\text{s.d.}_{\text{non-aberrant}}$	0.11306	0.06914
N_{dupl}	212	212
$\text{mean}_{\text{dupl}}$	0.48515	0.48777
$\text{s.d.}_{\text{dupl}}$	0.12920	0.11967
$\mathbf{SN_{dupl}}$	**3.93861**	**4.99782**
N_{del}	116	116
mean_{del}	0.76779	0.78468
s.d._{del}	0.22275	0.18787
$\mathbf{SN_{del}}$	**4.30702**	**5.54778**

True positive and false positive rate

First, we compute for a chosen set of significance levels α, the percentage of the duplicated and deleted reporters that are correctly classified as duplicated and deleted, respectively, in our test data set, according to both methods. The true positive (TP) rates are shown for the mixed model and the linear model in Figure 2.5 in function of the significance level α. For the linear model, the TP rate reaches a maximum of 0.954 for significance level $\alpha = 0.009$. For higher α, surprisingly, the TP rate drops a little. This effect is caused by the fact that some reporters become *strange* reporters for larger significance levels α. For the classification with the mixed model, the TP rates grow slowly as the significance level increases. Within this range of significance levels α, it never reaches the maximum TP rate value obtained with the linear model. Perhaps it comes closer to the result obtained with the linear model if we allow for even larger significance levels α, but this will increase the FP rate and the number of strange reporters to an unacceptable level.

Outside the duplicated and deleted regions, other reporters were also classified as duplicated or deleted. These positives can be false positives due to technical artifacts, or they can indicate true biological variations. At this point, we will not make the distinction between both kinds of aberrant reporters, as it does not affect the method comparison, and we will refer to this set of positives as non-confirmed positives. At a later stage, this set of non-confirmed positives will be examined in-depth, for one method and one significance level α.

The number of these non-confirmed positives is shown in Figure 2.3.6. The figure shows that there is no clear difference in the non-confirmed positive rates between both methods. For low significance levels α the linear model has a slightly smaller number of non-confirmed positives.

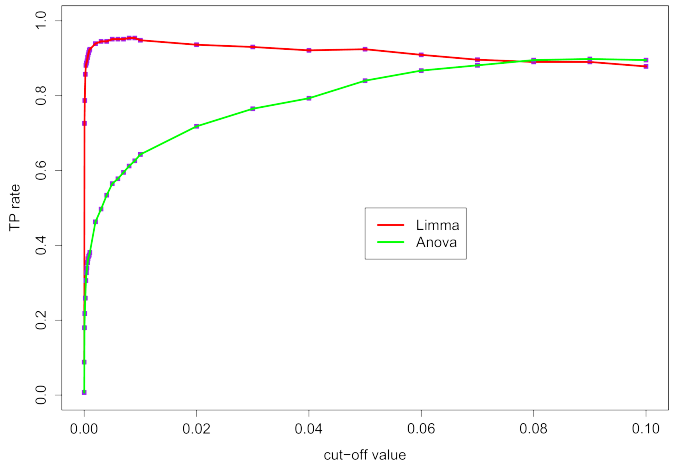


Figure 2.5: True positive rate. For the mixed model and the linear model, the true positive (TP) rates are plotted for a chosen set of significance levels α with a green and red line, respectively.

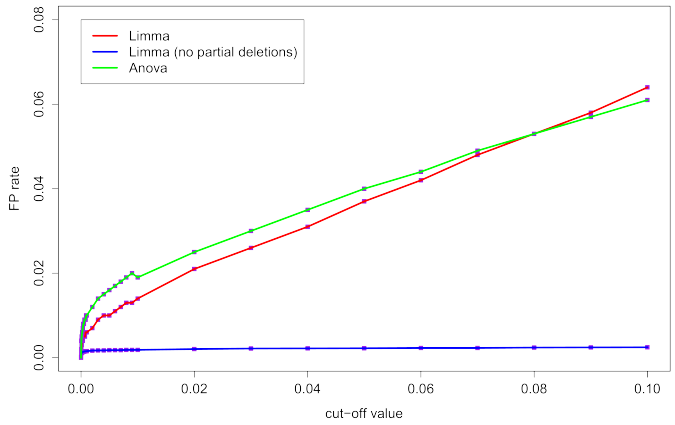


Figure 2.6: False positive rate. For the mixed model and the linear model, the false positive (FP) rates are plotted for the range of significance levels α with a green and red line, respectively. The FP rate for the complete deletions or duplications, obtained via the linear model is indicated in blue.

The combined results on the signal-to-noise ratio, the TP, and non-confirmed positive rate lead to the conclusion that the linear model is the preferable method.

Optimization of the linear model

In the previous section, we focused on how well the different methods fit the measurements by assessing its capabilities to divide the non-aberrant reporters from the deviant reporters and by comparing the FP and TP rates. This indicated that the linear model was best suited to distinguish these groups of reporters, although also the linear model has a fairly high FP rate. However, we did not yet benefit from all available information.

Completely and partially deleted reporters To further optimize the reporter classification, we will use the fact that if, for example, a reporter of Patient 1 is deleted or duplicated, its contrast Patient 2 vs. Patient 1 should theoretically be equal to $\log_2\left(\frac{2}{1}\right) = -1$ or $\log_2\left(\frac{2}{3}\right) = 0.58$, respectively. However, nonlinear saturation effects in the signals cause a deviation from these values. Instead of taking the theoretically expected values (i.e., ± 1 and ± 0.58), we will estimate the expected values based on the linear model estimates of the contrasts for the group of confirmed deletions and duplications, after exclusion of the deletions and duplications on the X and Y chromosome. This results in an average for the absolute log ratios of 0.86 for the deleted reporters and 0.54 for the duplicated reporters.

Therefore, if we detect with the linear model a reporter that is likely to be duplicated or deleted, we extract its contrasts C_{c1} and C_{c2}. If their absolute value is not larger than 0.54 or 0.86, respectively, we use an adapted version of the moderated t-test, as implemented in the `LIMMA` (Linear Models for Microarray Data) package. We use the same standard deviation of the contrasts, as computed within the previous `LIMMA` procedure, and test in a one-sided way the hypothesis $H_0 : C = 0.54$ versus $H_a : |C| < 0.54$ for the duplicated reporters and $H_0 : |C| = 0.86$ versus $H_a : |C| < 0.86$ for the deleted reporters. If we cannot reject the hypothesis at a significance level α_{partial} for both tests, we call the reporter *completely* deleted or duplicated, else we call the reporter *partially* deviant. As a significance level, we choose $\alpha_{\text{partial}} = 0.01$.

For this significance level, the non-confirmed positives restricted to the reporters that are completely deleted is plotted as a blue line in Figure 2.3.6. This non-confirmed positives rate ranges between 0.001 and 0.002.

The non-confirmed positives The non-confirmed positives rate obtained in the previous sections is not a direct indication of the false positives rate, as they can comprise not only false positives, but also both true positives or polymorphic reporters.

In our data set of nine loops, we saw that after extraction of clones that were proven to be copy number variable regions, 19 single clones complete deletions or duplications are picked up by our method, at a significance level of

$\alpha = 0.001$. Seven of these clones have been previously described as polymorphic [258]. One clone (RP1-93N13) overlaps with a CNV present in the normal population (Genomic Database of Variants; `http://projects.tcag.ca/variation/`http://projects.tcag.ca/variation/). To determine whether the remaining reporters are false positives or true CNVs, qPCR or FISH [158] was performed. For six clones, the deletion or duplication was confirmed. For one clone we could not determine the copy number variation due to lack of DNA. The remaining four clones were all measured in Experiment 1 and caused by low spot quality, as they had a large discrepancy between the mean and the median pixel intensities and a large standard deviation of the pixel intensity. Hence, our method reliably detects aberrant clones. An overview of these results is shown in Table 2.5.

Table 2.5: The non-confirmed positives. Outside the duplicated and deleted regions, 15 reporters were also classified as completely duplicated or deleted. Eight of them could be confirmed as being polymorphic clones according to Vermeesch [258] and the Database of Genomic Variants. Six of them confirmed with FISH or qPCR. One clone could not be confirmed, due to lack of DNA.

Clone	Status	Validation method
RP11-114K7	See publication [258]	
RP11-161M6	[258]	
RP1-225D2	[258]	
RP6-14C6	[258]	
RP11-342F21	[258]	
RP6-22D12	[258]	
RP1-93N13	Database of Genomic Variants	
RP11-2P5	Single clone deletion	qPCR
RP11-48E16	Single clone deletion	qPCR
RP11-127A9	Single clone deletion	qPCR
RP11-469N6	Single clone deletion	qPCR
RP11-576I16	Single clone duplication	qPCR
RP11-152O18	Single clone deletion	FISH
RP11-361M10	Not confirmed	Lack of DNA
RP11-404P12	Low spot quality	
RP5-982E9	Low spot quality	
RP11-301H15	Low spot quality	
RP4-742J24	Low spot quality	

At $\alpha = 0.001$, also 13 clones were classified as being *strange*. Ten of these clones overlapped with CNVs that are present in the normal population, according to the Database of Genomic Variants. (Note also that not all normal CNVs have already been identified.) Therefore, we suspect that a *strange* clone can often be explained as a polymorphic clone that is shared by two patients.

Taken together these results indicate that at $\alpha = 0.001$ our procedure has both low false negatives (a TP rate around 95%) and essentially no false positives

for completely deleted or duplicated clones (most clones detected as positive or strange outside the regions known in the benchmark have been confirmed by qPCR or FISH; several of them being known normal CNVs).

2.3.7 Web application

The method is implemented in a web based framework and is available at `www.esat.kuleuven.be/loop`www.esat.kuleuven.be/loop. A demo account with test data is available. The application and the statistical analysis have been tested and refined on an in-house series of over 400 patients. Between March 2005 and August 2008, the system has stored and analyzed over 3000 assays by about 40 lab technicians and/or researchers across different genetic testing centers throughout Europe.

The application offers four components: an upload wizard for GPR [6] files, a slide view that provides an overview of all uploaded hybridizations (shown in Figure 2.8), a loop design view offering reports of all loop design experiments (shown in Figure 2.9), and a two-way design view offering reports on regular experiments, which can also be handled by the application.

Upload Module The upload wizard expects the GenePix GPR file format. Other scanner manufacturers have included the possibility to export in this format in their scanner software. The wizard is shown in Figure 2.7. At this point, the user is allowed to add a single *.gpr* file to the slide repository. Files are added one at a time. The user can combine multiple slides into a loop design later on. In order to add a slide to the repository, 5 bits of information need to be provided: the *.gpr* file itself, who conducted the experiment, when it was performed, and what the unique names ('references') are for the Cy5 sample and the Cy3 sample. It is important for these references to be correct. The system will use it to infer how loop designs are set up. The upload wizard expects that unique identifiers be provided for Cy5 and Cy3 samples: these are used to verify loop designs as valid, and are used as sample references throughout the tool. Entering a slide into the repository can take some time, as .gpr-files are rather large. Finally, it is possible to add notes or comments to the slide, if necessary.

Slide View Module The slide view provides an overview of basic information on uploaded hybridizations. It is shown in Figure 2.8. Each hybridization record can be folded open to view technical information, and timing and lab information can be edited. In this view, three hybridizations that make up a single loop design are checked, and selected experiments are combined into a loop design. The list view only shows recent uploads. A full list can be obtained as well.

Loop Design View Module Overview of finished loop design analysis runs. The overview shows an identifier, the date the analysis was performed, experi-

[6] GenePix Results (GPR) File



Figure 2.7: Adding scanner result files to the repository This page allows the user to add a single *.gpr* file to the slide repository. Files are added one at a time. The user can combine multiple slides into a loop design later on. In order to add a slide to the repository, 5 bits of information need to be provided: the *.gpr* file itself, who conducted the experiment, when it was performed, and what the unique names ('references') are for the Cy5 sample and the Cy3 sample. It is important for these references to be correct. The system will use it to infer how loop designs are set up. Entering a slide into the repository can take some time, as .gpr-files are rather large. Finally, it is possible to add notes or comments to the slide, if necessary.

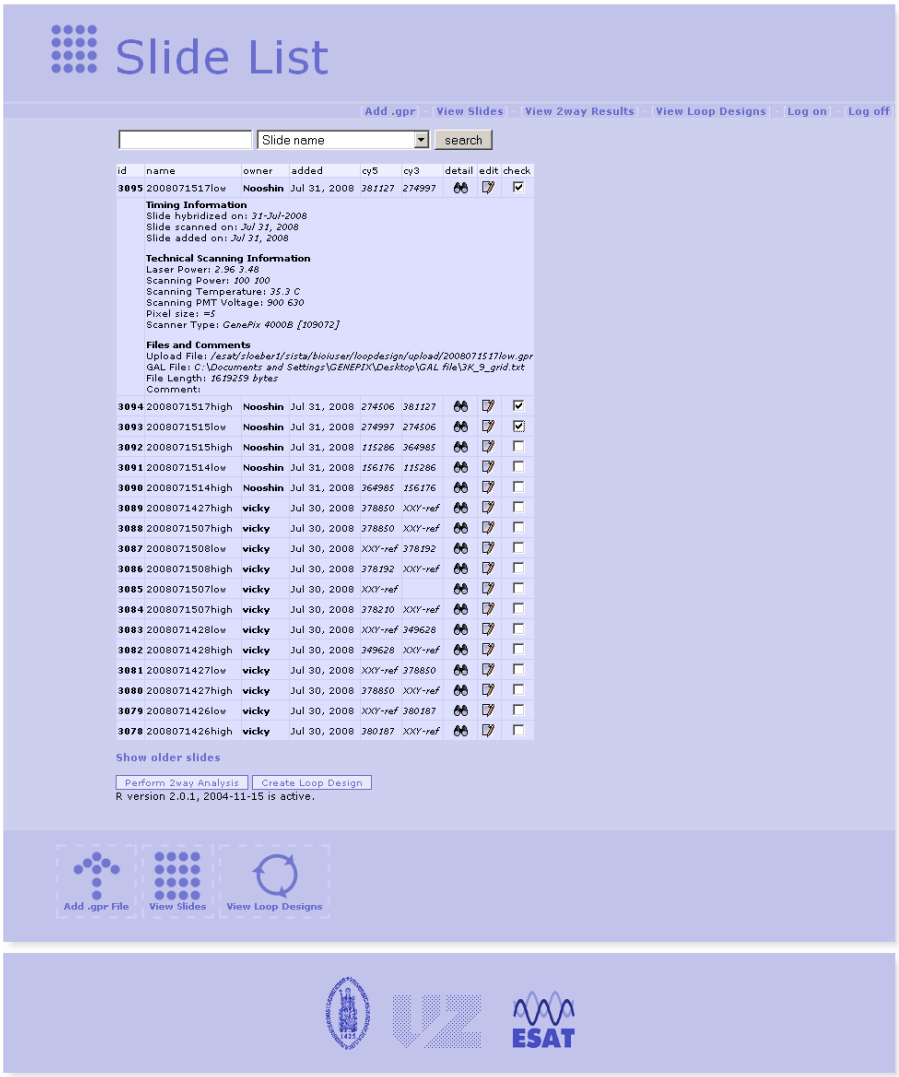


Figure 2.8: List of slides present in the slide repository. The slide view provides an overview of basic information on uploaded hybridizations. Each hybridization record can be folded open to view technical information, and timing and lab information can be edited. In this view, three hybridizations that make up a single loop design are checked, and selected experiments are combined into a loop design. The list view only shows recent uploads. A full list can be obtained as well.

ments and design details such as patient identifiers and respective dyes, and a status. Status can be 'loading files', 'running LIMMA', or 'Done', depending of the phase of the current run. When finished, the analysis run will enable four links in the results column: one overview (described in Figures 2.10, 2.11 and 2.12), and a detailed patient view for each individual patient, as shown in Figure 2.13.

After submitting three hybridizations as a new loop design, a new entry is added in the loop design view. A status comment indicates what phase of the analysis is active. When done, four reports can be viewed: an overview, and three individual patient reports. The overview report will display experiment design and quality assessment information, Cy5 and Cy3 slide background images, and MA-plots before and after 2D LOESS normalization. Significantly aberrant reporters are shown in an overview, highlighting reporters previously marked as polymorphic. Potentially polymorphic reporters are shown in Table 2.6. If needed, a list of significantly aberrant reporters and a list of all reporters can be downloaded from this report as a tab delimited text file for further processing. A graphical overview shows normal and aberrant reporters ordered by chromosomal position for all hybridizations. The individual patient reports show quality statistics and a graphical display of significantly aberrant reporters. On this overview, the user can zoom in to individual chromosomes. A table provides an overview of aberrant reporters, which can again be downloaded as a tab delimited file.

Loop Design experiment report The overview of a single finished loop design experiment is shown in Figures 2.10, 2.11 and 2.12. The overview shows which slides were included in the current loop design experiment. It then goes on providing information which allows the array quality to be assessed. For this, it reports on the spread of the log_2 background intensities on the slides and the number of spots that report above background. A spot is called to be above background if $F > B + 2 * stdev(B)$, with F = foreground, B = background. In the table, the number and percentage of the spots that are above background, is shown for both channels combined, and for the green and red channel separately.

For all hybridizations, the MA-plots [7] are shown both before (left) and after (right) the spatial LOESS normalization is performed. The measurements in the MA-plots are colored blue if they are above background and green if they are below. The empty spots are colored red. The report goes on showing details on the LIMMA analysis, which fits a linear model through the ratios, incorporating the design of the experiment. An empirical Bayes method is used to estimate the standard deviations of the log ratios. With these standard deviations, the moderated t-test is performed. This test has an augmented number of degrees of freedom compared to the classical t-test and is therefore more powerful. We will call a clone differentially expressed if it has a p-value smaller than 0.001, corrected to control the false discovery rate with Benjamini-

[7] M (intensity ratio) versus A (average intensity) plot to visualize intensity-dependent ratio of raw microarray data

Reporter ID	# detected	Chromosome
_1Mb_PACE9	9	1
NONSC1B4	9	15
SC9BACMbset_1F12	7	17
NONSC24B9	7	5
SC10BAC-1Mbset-1E12	6	10
NONSC16A2	5	5
NONSC43G1	5	17
NONSC24A3	5	-
NONSC34A4	3	16
NONSC33A5	3	19
NONSC32E3	5	16
SC10BAC-1Mbset-1A3	4	10
SC6PAC1Mbset1F9	4	6
NONSC31C3	4	17
Cancer_1G8	3	5
SC13_1Mb_BAC1E4	3	13
SC13_1Mb_BAC1A5	4	Y
NONSC10B6	2	5
SC1_1Mb_BAC1G5	3	1
NONSC32B9	2	15
SC22_0.75_BACB8	2	22
NONSC8F4	3	17
NONSC10C7	3	7
NONSC2C8	2	8
NONSC3C7	2	2
NONSC8D3	2	2
telomereB8	2	14
telomereG1	2	2
NONSC40A4	2	8
NONSC40D1	2	4
NONSC41E2	2	8
NONSC11G2	2	8
NONSC10G9	3	2
NONSC10C3	2	16
Cancer_1A10	3	11
NONSC40E7	3	16

Table 2.6: The list of potential polymorphic reporters. Next to their identifier, an indication of the number of times it was detected as deviant, along with its chromosome (if determined).

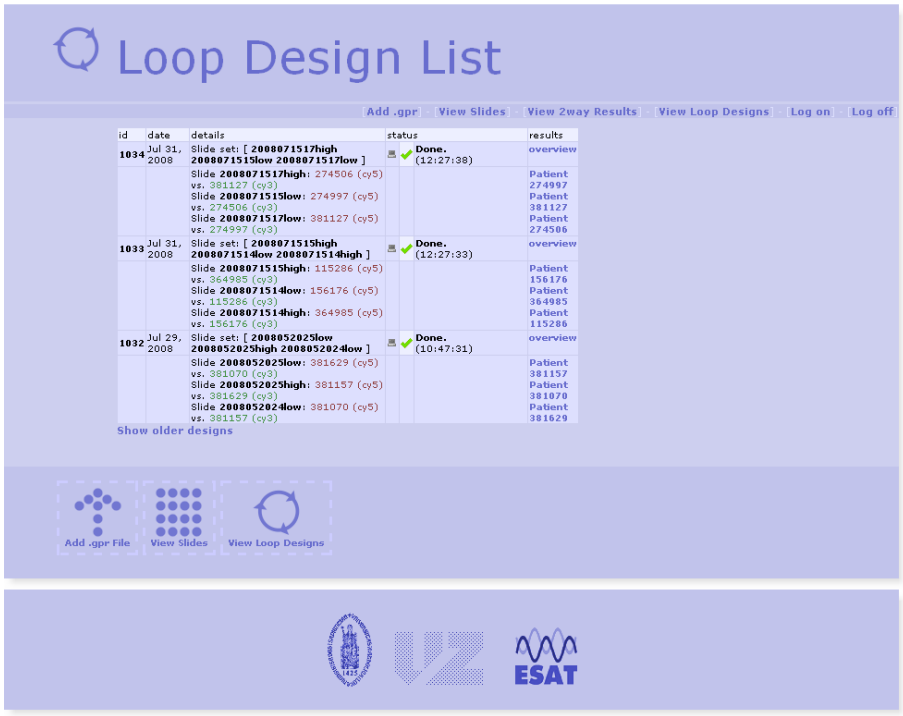


Figure 2.9: The loop design list view. Overview of finished loop design analysis runs. The overview shows an identifier, the date the analysis was performed, experiments and design details such as patient identifiers and respective dyes, and a status. Status can be 'loading files', 'running LIMMA', or 'Done', depending of the phase of the current run. When finished, the analysis run will enable four links in the results column: one overview (described in Figures 2.10, 2.11 and 2.12), and a detailed patient view for each individual patient, as shown in Figure 2.13.

Hochberg. The classification scheme for clone calling is shown. Through the use of a LIMMA model, a table is produced that shows the three log ratios for each clone. The table contains some redundancy as each of these three columns is a combination of the two other columns, due to the design of the experiment. The next three columns show the classification. A '1' indicates that the clone is classified as in state of gain, a '-1' indicates loss and 0 shows no difference between the two patients. Clones that can't be classified according to these rules are called 'STRANGE'. This classification is shown in the column 'Class'. The columns 'Chrom', 'kbbegin', and 'kbend' display the location of the clone on the chromosome. The next three columns show for each patient the number of measurements that are above background. Clones with less than 2 measurements for a patient are omitted. The last column gives an indication if they are likely to be completely or partially deleted or duplicated. This was done by testing whether the absolute value of two of the estimates was significantly lower than 0.86 or 0.54, respectively, with p-value 0.01. Clones that are supposed to be polymorphic are colored in light-blue. The table of significant clones can also be obtained in a *.txt* file. A more extensive table of all clones is also stored and available for immediate download. Finally, detailed chromosome ordered graphs for all 3 hybridizations are shown.

Single patient report The overview of a single patient report within a finished loop design experiment is shown in Figure 2.13. The report shows deletions and duplications for the patient under review. It starts with quality statistics for the related hybridizations through a table that shows the standard deviation of the log ratios for each hybridization, after removing the observations for the clones on chromosome X, Y, and the aberrant clones. Like in the global overview described in Figures 2.10, 2.11 and 2.12, here also an indication for the array quality is given through the number of spots above background. Here also, we call a spot above background if $F > B + 2 * stdev(B)$. In the table, the number and percentage of the spots that are above background is shown: once for both channels combined, and once for the green and red channels separately. Finally, both a graphical and tabular overview are shown to display the significant clones obtained with the LIMMA techniques. First, the plot itself will show the average of the LIMMA-estimates for the different clones, ordered according to their chromosome location. Grey spots are not classified as deleted or duplicated for this specific patient. The colored clones are duplicated or deleted for the patient under review. In case a clone is likely to be only partially deleted (*i.e.* significantly below 0.86) or duplicated ((*i.e.* significantly above -0.54), it is drawn as a triangle. Finally, the results in the table can be downloaded as a tab-delimited file.

Architecture The web application consists of a set of Java Server Pages (JSP) that are run in Apache Tomcat Java application servers on two different machines. A single MySQL database instance with daily back up is used as a back-end data store and holds user and hybridization loop design data. GPR

The design of the experiment.

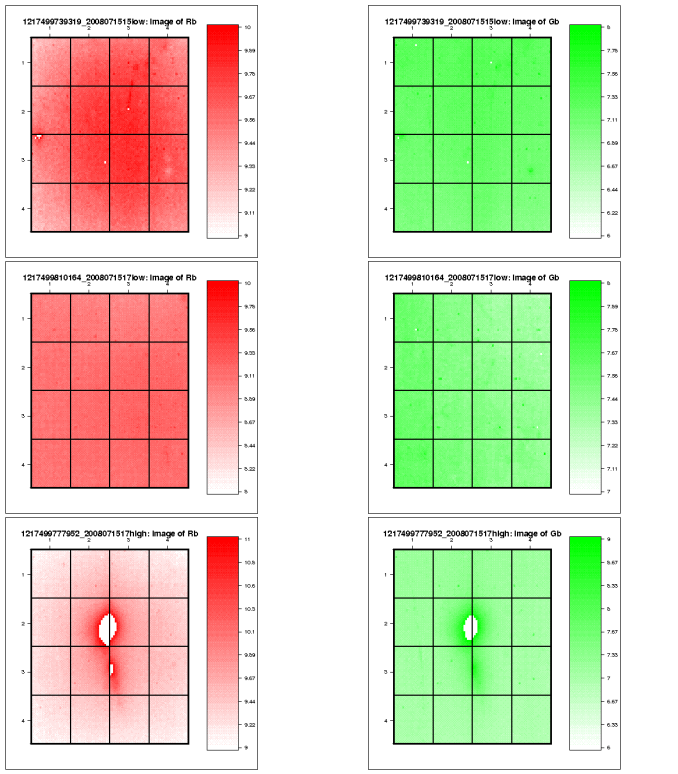
We have the following hybridisations:

	Hybridisation	Patient in Cy5	Patient in Cy3
1	1217499739319_20080715151low	274997	274506
2	1217499810164_20080715171low	381127	274997
3	1217499777952_20080715171high	274506	381127

Assessing the array quality.

In order to check the quality of the arrays, we look at the spread of the log base 2 background intensities on the slides and the number of spots 'above background'.

In the following plots, the log base 2 background intensities are shown for the Cy3 and Cy5 intensities. This function can be used to explore whether there are serious spatial effects in the data, for example, print-tip or cover-slip effects.



A second indication for the array quality is the number of spots above background. Here, we call a spot above background if Fg>Bg+2*sd(Bg). In the table, the number and percentage of the spots that are above background, is shown; for both channels combined, and for the green and red channel separately.

Hybridisation	# in both channels	% in both channels	# in Red channel	% in Red channel	# in Green channel	% in Green channel
1217499739319_20080715151low	7159	92.446	7161	92.472	7501	96.862
1217499810164_20080715171low	7416	95.764	7419	95.803	7564	97.676
1217499777952_20080715171high	7285	94.073	7286	94.086	7568	97.727
For 274997 combined	7145	92.265	7148	92.304	7489	96.707
For 381127 combined	7270	93.879	7272	93.905	7540	97.366
For 274506 combined	7119	91.929	7120	91.942	7486	96.668

Figure 2.10: The loop experiment report. Overview of a single finished loop design experiment. The overview shows which slides were included in the current loop design experiment. It then goes on providing information which allows the array quality to be assessed. For this, it reports on the spread of the log_2 background intensities on the slides and the number of spots that report above background. In the table, the number and percentage of the spots that are above background, is shown for both channels combined, and for the green and red channel separately

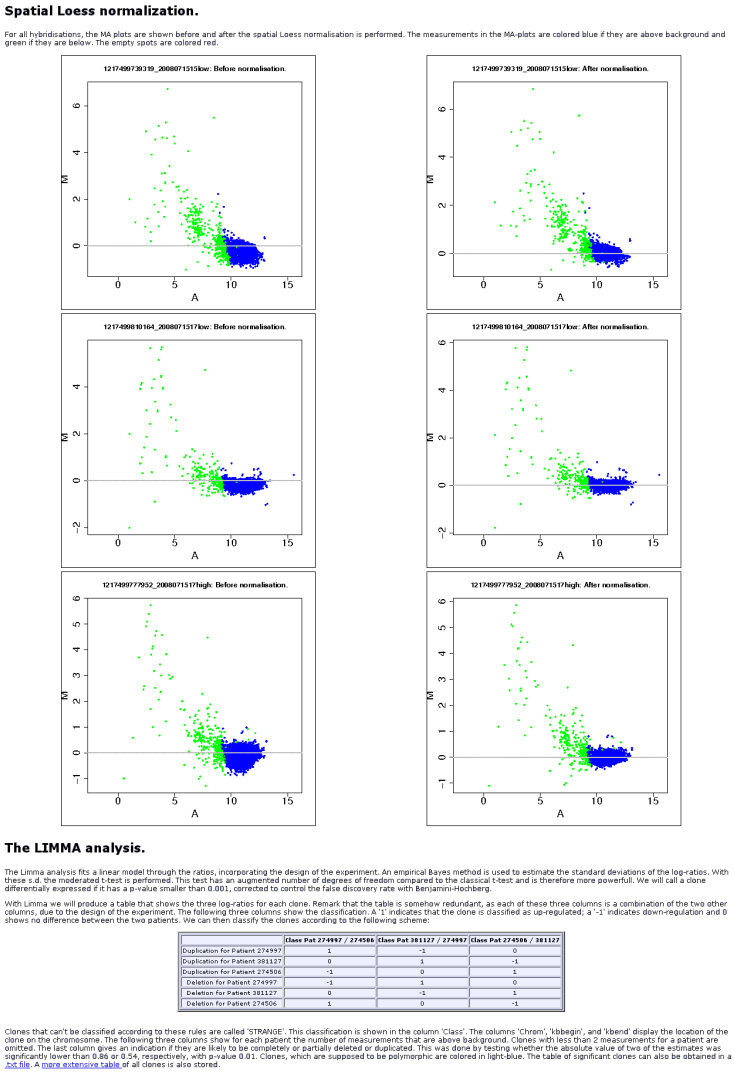
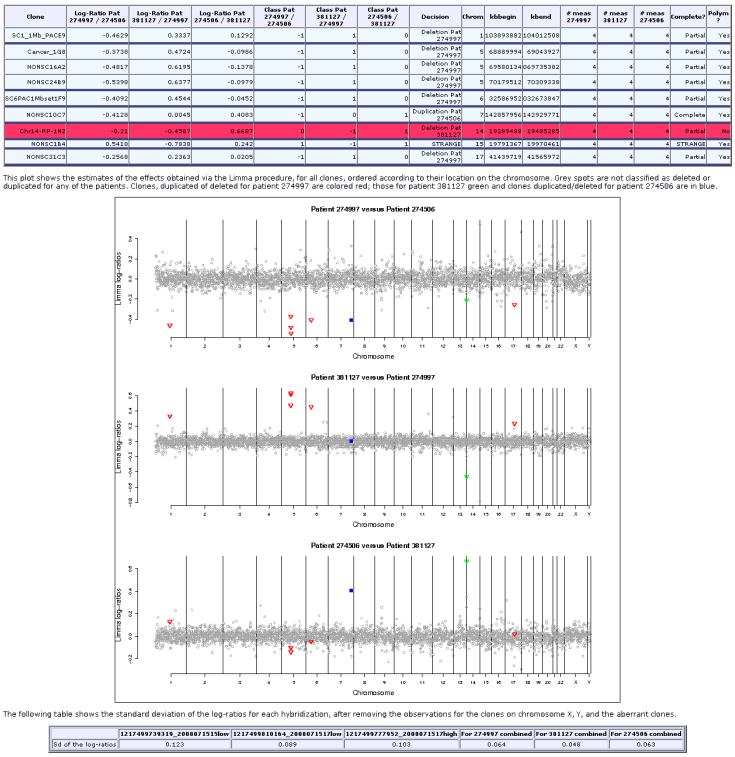


Figure 2.11: The loop experiment report. Overview of a single finished loop design experiment. For all hybridizations, the MA-plots are shown both before (left) and after (right) the spatial LOESS normalization is performed. The measurements in the MA-plots are colored blue if they are above background and green if they are below. The empty spots are colored red. The report goes on showing details on the LIMMA analysis, which fits a linear model through the ratios, incorporating the design of the experiment. The classification scheme for clone calling is shown



Clone	Log-Ratio Pat 274997 / 274506	Log-Ratio Pat 381127 / 274997	Log-Ratio Pat 274506 / 381127	Class Pat 274997 / 274506	Class Pat 381127 / 274997	Class Pat 274506 / 381127	Decision	Chrom	kbbegin	kbend	# meas 274997	# meas 381127	# meas 274506	Complete?	Polym ?
SC1_1Mb_PACE9	-0.4629	0.3337	0.1292	-1	1	0	Deletion Pat 274997	1	103893882	104012508	4	4	4	Partial	Yes
Cancer_168	-0.3738	0.4724	-0.0986	-1	1	0	Deletion Pat 274997	5	68889994	69043927	4	4	4	Partial	Yes
NONSC16A2	-0.4817	0.6195	-0.1378	-1	1	0	Deletion Pat 274997	5	69580134	69735382	4	4	4	Partial	Yes
NONSC2489	-0.5398	0.6377	-0.0979	-1	1	0	Deletion Pat 274997	5	70179512	70309338	4	4	4	Partial	Yes
SC6PAC1Mbsetf9	-0.4092	0.4544	-0.0452	-1	1	0	Deletion Pat 274997	6	32586952	32673847	4	4	4	Partial	Yes
NONSC10C7	-0.4128	0.0045	0.4083	-1	0	1	Duplication Pat 274506	7	142857956	142929771	4	4	4	Complete	Yes
Chr14-RP-1N2	-0.21	-0.4587	0.6687	0	-1	1	Deletion Pat 381127	14	19289488	19485285	4	4	4	Partial	No
NONSC184	0.5418	-0.7838	0.242	1	-1	1	STRANGE	15	19791267	19970461	4	4	4	STRANGE	Yes
NONSC31C3	-0.2568	0.2363	0.0205	-1	1	0	Deletion Pat 274997	17	41439719	41565972	4	4	4	Partial	Yes

This plot shows the estimates of the effects obtained via the Limma procedure, for all clones, ordered according to their location on the chromosome. Grey spots are not classified as deleted or duplicated for any of the patients. Clones, duplicated of deleted for patient 274997 are colored red; those for patient 381127 green and clones duplicated/deleted for patient 274506 are in blue.

The following table shows the standard deviation of the log-ratios for each hybridization, after removing the observations for the clones on chromosome X, Y, and the aberrant clones.

	1217499739319_200807151Slow	1217499810164_200807151Flow	1217499777952_200807151High	For 274997 combined	For 381127 combined	For 274506 combined
Sd of the log-ratios	0.123	0.089	0.103	0.064	0.049	0.063

Figure 2.12: The loop experiment report. Overview of a single finished loop design experiment. Through the use of a LIMMA model, a table is produced that shows the three log ratios for each clone. The table contains some redundancy as each of these three columns is a combination of the two other columns, due to the design of the experiment. The next three columns show the classification. A '1' indicates that the clone is classified as in state of gain, a '-1' indicates loss and 0 shows no difference between the two patients. Clones that can't be classified according to these rules are called 'STRANGE'. This classification is shown in the column 'Class'. The columns 'Chrom', 'kbbegin', and 'kbend' display the location of the clone on the chromosome. The next three columns show for each patient the number of measurements that are above background. Clones with less than 2 measurements for a patient are omitted. The last column gives an indication if they are likely to be completely or partially deleted or duplicated. Clones that are supposed to be polymorphic are colored in light-blue. The table of significant clones can also be obtained in a *.txt* file. A more extensive table of all clones is also stored and available for immediate download. Finally, detailed chromosome ordered graphs for all 3 hybridizations are shown

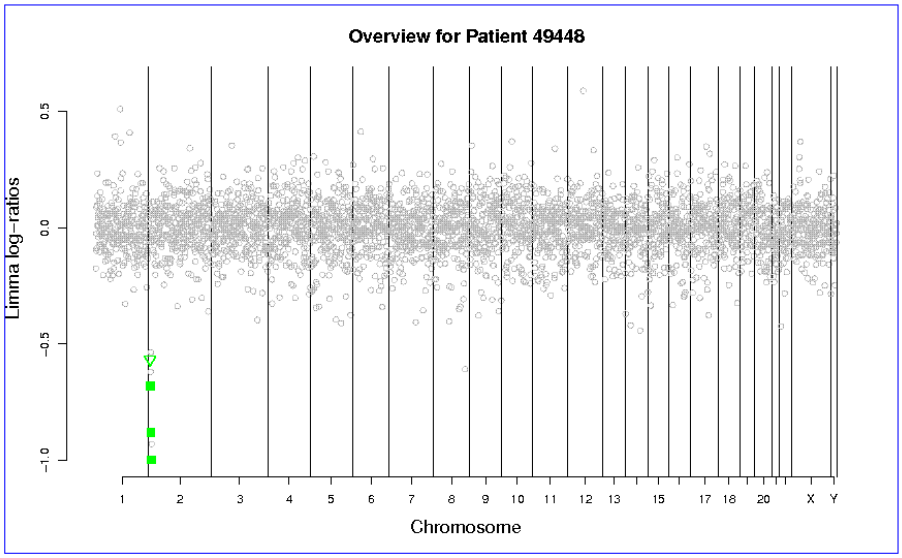
Deletions and duplications for Patient 49448.

Quality statitistics for the related hybridizations:

1217230586074_2008071504low	1217230677159_2008071505LOW	For 49448 combined
0.28	0.28	0.15

Hybridisation	# in both Channels	% in both channels	# in Red channel	% in Red channel	# in Green channel	% in Green channel
1217230586074_2008071504low	7519	97.095	7519	97.095	7612	98.295
1217230677159_2008071505LOW	7568	97.727	7570	97.753	7611	98.283
For 49448 combined	7503	96.888	7503	96.888	7602	98.166

The significant clones obtained with the Limma techniques:



Clone Id	EMBL accession	Class	Chrom	kbbegin	kbend	# meas	Complete?	Polym ?
NONSC10B4	RP11-292A20	Deletion	2	4779724	4937997	4	Complete	No
NONSC14E5	RP11-485O17	Deletion	2	6648268	6738494	4	Complete	No
nonsc41C6	RP11-327H5	Deletion	2	3921580	4126930	4	Partial	No
nonsc41H7	RP11-496C3	Deletion	2	7448599	7626975	4	Complete	No

Figure 2.13: A single patient report within a loop experiment. Overview of a single patient report within a finished loop design experiment. The report shows deletions and duplications for the patient under review. It starts with quality statistics for the related hybridizations through a table that shows the standard deviation of the log ratios for each hybridization, after removing the observations for the clones on chromosome X, Y, and the aberrant clones. Like in the global overview described in Figures 2.10, 2.11 and 2.12, here also an indication for the array quality is given through the number of spots above background. In the table, the number and percentage of the spots that are above background is shown: once for both channels combined, and once for the green and red channels separately. Finally, both a graphical and tabular overview are shown to display the significant clones obtained with the LIMMA techniques. First, the plot itself will show the average of the LIMMA-estimates for the different clones, ordered according to their chromosome location. Grey spots are not classified as deleted or duplicated for this specific patient. The colored clones are duplicated or deleted for the patient under review. In case a clone is likely to be only partially deleted or duplicated, it is drawn as a triangle. Finally, the results in the table can be downloaded as a tab-delimited file.

files are stored on RAID [8] disks with daily backup, as are images and statistics results. Statistics and visualization scripts were written in R [9] and make use of BioConductor [10] modules. Instances of RServe [11], a daemon interface to R that accepts and handles remote calls from Java [12], are installed on two Unix machines for increased availability.

2.3.8 Discussion on the LOOP approach

A first point of discussion is to which platforms the proposed statistical models are applicable. The loop design is directly dependent on using a two-channel microarray. Some array CGH platforms are available using single-channel microarrays. In this situation, the loop design is simply not applicable.

A second point of discussion is a main assumption that underlies our analysis For two-channel microarrays, which is that two patients never share the same aberration. If two patients have an aberration for the same reporter, the statistical model cannot correctly interpret the result—resulting mostly into a *strange* reporter. For the detection of congenital anomalies, this effect is prevented by putting into a loop design patients who have clearly distinct phenotypic patterns. The rare cases where a common aberration would still be present can be rescued at the validation stage. Incorrect assignment to the third patient would not be validated by FISH or qPCR and, in this case, the possibility of a common aberration for the two other patients should be kept in mind. Reporters flagged as strange can be validated in all three patients to clear out the situation. However, this assumption is unrealistic when studying tumors or when doing pre-implantation diagnosis. In the case of tumors, many chromosomal regions can be affected and overlap between patients is essentially unavoidable. In the case of pre-implantation diagnosis, procedure was developed at ESAT and CME where one of the eight blastomere cells of an 8-cell embryo from in vitro fertilization is assessed by single-cell array CGH [132]. This procedure does not perturb the implantation and development of the embryo and allows the detection of aberrations of the embryo. However, in this procedure it is essentially impossible to guarantee that the aberrations from different blastomeres do not overlap (for example, some aberrant embryos have chaotic genomes where many chromosomes are affected).

In some rare cases, deviant reporters cannot be assigned as duplicated or deleted, but are instead labeled as *strange*. We do not expect those reporters to correspond to aberrations common to two patients. Rather we observed that most of these *strange* reporters corresponded to polymorphic reporters. This can correspond to the situation where two patients share a complex polymorphic reporter [190]. Further investigation even suggests that those polymorphic reporters may have multiple variants. This will be investigated more deeply in

[8] Redundant Array of Independent Disks
[9] `www.r-project.org`
[10] `www.bioconductor.org`
[11] `www.rforge.net/Rserve/`
[12] `www.java.com`

later studies.

Our analysis was characterized by the fact that for our optimized threshold we have a low number of false negatives and close to no false positives. Given that a secondary validation is available through FISH, which can catch false positives but not false negatives, we would want to increase our threshold further to decrease the number of false negatives at the cost of a few false positives that can then be caught by the FISH validation. However, increasing the threshold further results rather in more reporters being labeled as *strange*, which we prefer to avoid.

At a significance level of 0.001, for 0.251 % of the clones the intensity ratios were suggestive of partial deletions or duplications. Most likely, the intermediate intensity ratios reflect the presence of sub-clonal CNVs. Recently, several articles show high levels of small CNVs in the human genome [247, 190, 109, 213]. Moreover, qPCR was performed at the CME on some of these clones using several PCR primers and could confirm such partial deletions (data not shown). However, since no systematic analysis of all these clones has been performed, we do not know the actual number of false positives.

In our specific setting, linear models clearly outperformed mixed models. We do not draw any conclusion with regard to the suitability of one class of method versus the other in general. In our setting, we hypothesize that the mixed model was less robust to deviation from the underlying normality assumptions or that the more compact (fewer parameters) estimation procedure of the linear model increased its robustness. The improvement of the mixed model in this setting appears to be an interesting research direction.

Microarray CGH for the diagnosis of congenital chromosomal aberrations is progressing rapidly from the research lab to the clinic. In our laboratory, over 500 patients have been assessed in an almost routine clinical setup. What value do such improved statistical procedures add to the diagnosis? First of all is the issue of cost per experiment. While for gene expression microarrays the pressure on cost comes from the limitations of research budgets, the pressure on cost for array CGH for congenital aberrations comes from health care management considerations. To replace conventional karyotyping technology, the cost of array CGH must be close to that of conventional karyotyping. At present, commercial array CGH diagnostics cost two to three times more than a karyotype. Home-made platforms come closer to the cost of conventional karyotyping. To keep cost down, some laboratories do not perform a dye swap when using the basic array CGH setup (patient vs. normal reference). In this case, a single array is used per patient. However, only one measurement per patient will be available in this case instead of two with our design, which means a higher level of false positives and negatives. False positives will be caught by the secondary FISH or qPCR validation, but false negatives will not. This results in lower quality diagnostics for the patient.

Secondly, CNVs are hard to assign with certainty to the patient instead of the "normal" reference. While these CNVs are currently largely handled as noise, some of them are likely to act as genetic modifiers that are risk factors for disease or modulate its penetrance and phenotypic spectrum [190]. Association

studies are currently being started to evaluate the importance of such CNVs in both congenital and acquired disorders. Unambiguous assignment of the "normal" CNVs to the correct sample is therefore an important challenge.

As a last point, increasing the signal-to-noise ratio between aberrant and normal clones through our statistical procedure is likely to improve the detection of low-grade mosaicism (*i.e.*, chromosomal aberrations present in only a percentage of all cells), which is currently difficult to detect by array CGH. In a previous study [132],a power analysis model was presented for the detection of low-grade mosaicism where the signal-to-noise ratio was a critical factor in determining the limit of detection for mosaicism.

In conclusion, our results indicate that the experimental loop design, together with a statistical analysis by a linear model, provides an efficient procedure for the detection of chromosomal aberrations in congenital anomalies by array CGH. It is significantly superior to the classical setup by doubling the use of resources and unambiguously assigning variation to the correct patient. These improvements have a direct impact on the diagnosis offered to the patient using the microarray technology that is closest to routine clinical use.

2.3.9 Future work

With the advent of higher density clone based assays and especially the commercial oligo-based assays containing an increasing number of probes with every release, combined with the persistent interest of genetic diagnostic and research laboratories in 3-way design Array CGH assays, there is a clear need for adapting the methods described to experiments that boast data volumes an order of magnitude higher than the classical assays of only thousands of measure points per chip. Although the implementations of the algorithms underlying the LOOP application are amenable to optimization for speed, it remains clear that this has strong repercussions on computation power needs. Therefore, a server-based and preferably web-oriented approach is advantageous: already, multiple centers benefit from the server based architecture of the LOOP application described in this chapter, and as the future will bring a strong growth of computational and storage needs, this will become increasingly important. For this reason, the LOOP framework will be absorbed into the collaborative cytogenetics database framework Store+Bench, which will be described in Chapter 4 of this dissertation, where intelligent databases for clinical interpretation are discussed. In what follows, we will illustrate how the LOOP application itself can go a long way in research on copy number variant regions of the genome and that there are clear benefits when the focus of this tool shifts from a stand-alone multi-user web application towards a collaborative framework such as Store+Bench.

2.4 Human copy number variation

2.4.1 Importance of copy number variation mapping

Copy Number Variations or CNV's refer to genomic segments that are at least 1.000 base pairs in size and for which copy number differences have been observed in comparison to reference genome assemblies [62, 70]. Although DNA copy number variation has been associated with genomic disorders for a long time, the realization that this type of genomic variation is widespread in healthy individuals and in fact ubiquitous in mammalian genomes is recent [190]. Copy number variants are now believed to account for a substantial amount of genetic variation.

Techniques and efforts

It is believed that CNVs may account for a significant proportion of normal phenotypic variation since many CNVs include genes which in turn results in differential levels of gene expression [70]. Currently there are a number of efforts ongoing that are directed towards a more comprehensive cataloging and characterization of CNVs. The results of these studies will undoubtedly provide a strong new source of information for determining the impact of genomic diversity on biological function, human genetic disease, and even evolution [70].

Until recently, the analysis of the human genome lead researchers to believe that SNPs [13] were the main source of genetic and phenotypic human variation [62]. Although classical band stain karyotyping and FISH assays have enabled researchers to pick up microscopic variability for a long time, recently, through high-resolution and high-throughput techniques such as molecular karyotyping, submicroscopic variants have proven to be common. The advent of whole genome scanning technologies such as molecular karyotyping has uncovered that the extent of structural variation in the human genome is larger than expected. These structural variations include deletions, duplications, insertions, inversions and translocations of genomic fragments.

In this section, we will focus specifically on copy-number variability, also called copy-number polymorphism (CNP). The precise and comprehensive identification and characterization of CNVs would greatly benefit the functional analysis of human genome variation, and complement current genome-wide association studies based on SNPs, making an important contribution to human diversity and disease susceptibility.

CNV's can be identified and charted through a multitude of techniques, including SNP and CGH arrays. New technologies such as PEM (paired-end mapping) [126], clone-end re-sequencing [55], and whole-genome re-sequencing [214] are being used to chart copy number variations. In this chapter, we focus on Array-CGH [109, 213, 138, 190]. Both array assays based on genome-wide large insert clone as oligonucleotide assays can be used. In this chapter, we focus on Array CGH.

[13] Single Nucleotide Polymorphisms

2.4.2 Charting CNVs through accumulated diagnostic assay data

With the advent of large data volumes that comes with the entry of Array CGH into clinical practice, opportunities arise for CNV studies on this data. Interesting possibilities include population specific CNV studies, or studies charting CNV's per phenotype: while healthy individuals are bearers of many copy number variations, so called benign CNV's could actually still play a role in genetic disease as a risk or susceptibility factor. It is not straight forward to draw a clear line between a state of health and a state of disease: when does obesity become morbid, what about milder cases of heritable conditions such as autism, *etc.*

In first reports on copy-number polymorphisms, 76 (Sebat *et al.* [213], using tiled oligos) and 255 (Iafrate *et al.* [109], using BAC-arrays) loci were identified to display copy-number variation in the human genome. Buckley *et al.* point out that we are only at the point of mining the tip of an iceberg, and that a pool of individuals significantly larger in size needs to be analyzed to assess the frequency of CNVs using similar standards as applied for SNPs. There, a change detected in $< 1\%$ of all individuals is considered a mutation, rather than a polymorphism. Buckley *et al.* further point out that this suggested number might even need to be revised due to the fact that overall frequencies of rare versus common CNVs are currently unknown. Additionally, there is far less overlap in both data sets than one would expect.

Already, public efforts arise to chart Copy Number Variations. The Database of Genomic Variants [14] is a curated catalogue of structural variation in the human genome identified in healthy control samples. As of August, 2008, the catalogue contained a total of 27799 entries, including 17641 actual CNVs, 498 inversions, 660 InDels (insertions and deletions between 100 and 1000 base pairs, these variations are by definition smaller than CNVs), on a total of 5672 CNV loci. The objective of the Database of Genomic Variants is to provide a comprehensive summary of structural variation in the human genome. The Database of Genomic Variants provides a useful catalog of control data for studies aiming to correlate genomic variation with phenotypic data.

Another relevant resource in this context is the Washington University database of Segmental Duplications. The human genome contains numerous blocks of highly homologous duplicated sequence. This higher-order architecture provides a substrate for recombination and recurrent chromosomal rearrangement associated with genomic disease. [220]. These segmental duplications play a role in normal variation as well, and this database features a set of potential rearrangement hot-spots. Sharp *et al.* used a custom segmental duplication BAC microarray to screen a panel of normal individuals [220] representing populations from four continents, identifying over 100 regions of copy number polymorphism (duplications and deletions). Interestingly, their study shows that CNPs were usually not confined to a single population, suggesting that these either are

[14] urlprojects.tcag.ca/variation/

recurrent events, having occurred independently in multiple founders, or were present in early human populations.

2.4.3 LOOP hybridizations as valuable data source

The Database of Genomic Variants collects information from published studies and only takes genome variation information into account obtained from healthy individuals and control samples. If we leave this constraint out of the equation, data from patient studies can be considered as an information source as well. As the body of experimental results of the LOOP tool reached 1000 experiments, it became worthwhile to investigate clone behavior over such a large body of samples.

The result of this effort is a prototype application that allows to investigate individual BAC clones over all past hybridizations performed. It can only be considered a first step towards charting copy number variations based on patient data, since there are a number of important caveats that need to be addressed before it can be considered a sound and robust research tool. These caveats include:

- Sample data originates from patients that were screened for genomic aberrations because clinical indications pointed to the possibility of genetic defect, and not from healthy individuals.

- There is no control on the construction of the population cohort except from the geographical location of the genetic testing center. Patients can be referred from other centers, there is no restriction or verification on ethnicity or population group.

- Although patients are often referred for MR/MCA, there is no grouping performed based on phenotype information.

- The data are extracted for use before a normalization process has taken place to correct for dye bias, 2D slide effects, or other experimental effects that obfuscate the biological measurement.

- The resulting data were not manually or automatically analyzed to remove deleted or duplicated regions spanning a high number of reporters. Removing these larger aberrations could make sense because these stand a higher chance of not corresponding to the CNVs we are looking for. These large aberrations would probably link to the patient phenotype, and hence not represent the normal and apparently benign variation we are after.

- The samples are not separated into male and female groups.

- The result files that originate from the LOOP design tool data set were used directly, without separating the pairs of three hybridizations into virtual result files for three individual patients.

- replicates within a single hybridization were not aggregated. No spots are excluded.
- The results were not separated according to slide design. While all slide layouts were designed from the SANGER clone sets (Tile path, 1Mb, 32K and 30K), different slide layouts use different sub-sets of these clones. Hence, some clones are measured (far) more often than others.

Data extraction

The prototype application is made available under the working title 'Clone Wars'. The underlying data set was constructed by grouping all *.gpr* files present in the LOOP tool data repository. From these tabular files containing all information extracted from the original image scans of the microarrays, only the columns with (a) the clone name and (b) the background corrected log ratio for the corresponding measurement were withheld. This step is non-trivial due to differences in file formatting across designs and was done through PERL [15] regular expressions. Clone names were formatted in a uniform way to facilitate querying. The formatted clone names and log ratios were loaded into a MySQL [16] database. A web interface supporting simple queries was created as a layer on top of this database.

Querying interface

The web interface consists of one single page, asking the user for input and consequently generating a graph as output. The required input consists of a clone name (typically the clone well name or clone identifier), and information necessary to construct a histogram: an upper and lower boundary defining log ratios to be taken into account, and the number of bins.

Keeping in mind the caveats listed in Section 2.4.3 and the disadvantages of the fact that this prototype currently uses a simple histogram approach (*cfr. infra*, 2.4.3), we can attempt at a visual interpretation of clone profiles and correlate suspected clone polymorphism with existing databases available online, such as the Database of Genomic Variants, the copy number variations identified by Redon [190], and the University of Washington Segmental Duplication data set, all discussed above. The results of a selection of clones are shown in Figures 2.14, 2.15 and 2.16 for supposedly normal clones and Figures 2.17, 2.18 and 2.19 for putatively polymorphic clones. All profiles were generated between log ratio boundaries of -1 and 1, using 100 bins for histogram generation. The first three profiles show clones that are not reported as polymorphic, or overlapping with segmental duplications, in any of the mentioned data sets. The last three profiles were reported in one or more of these publicly available resources. These last three profiles explicitly show a multi-modal distribution, suggesting these clones are present in variable copies in the genomes of the individuals tested.

[15] Practical Extraction and Report Language, `www.perl.org`
[16] popular open source database, `www.mysql.com`

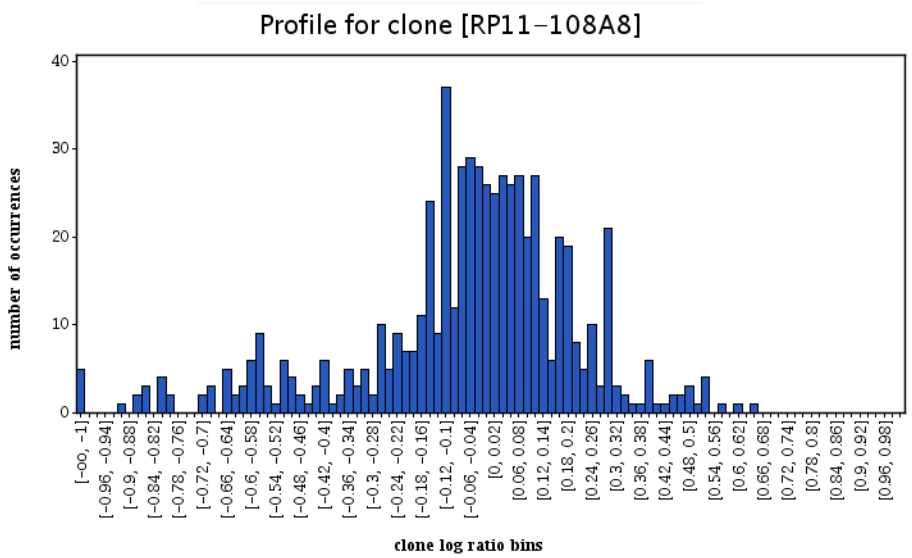


Figure 2.14: Log ratio profile over all loop design experiments for clone RP11-108A8. No DGV (Database of Genomic Variants), SegDup (Washington University Segmental Duplications), or Redon CNV loci are present in this region.

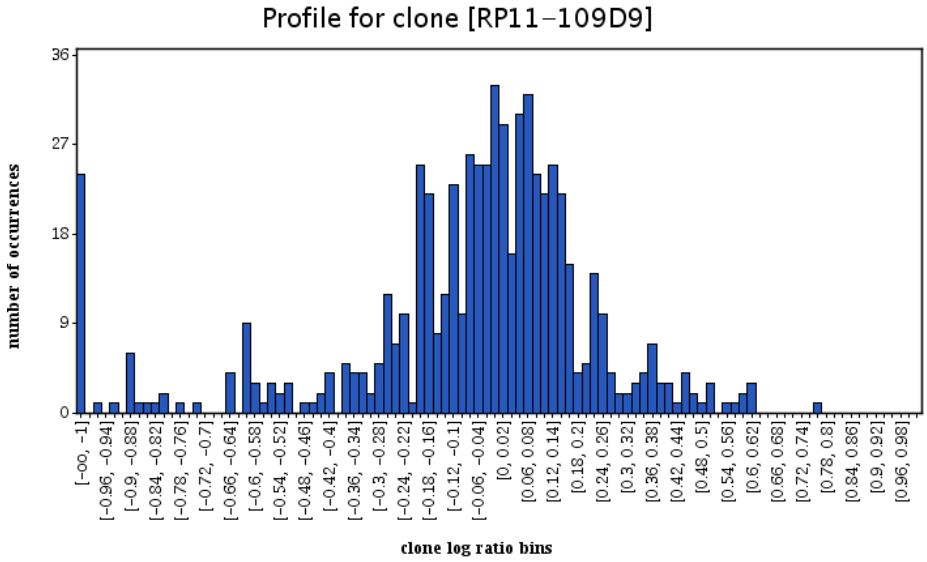


Figure 2.15: Log ratio profile over all loop design experiments for clone RP11-109D9. No DGV, SegDup, or Redon CNV loci are present in this region.

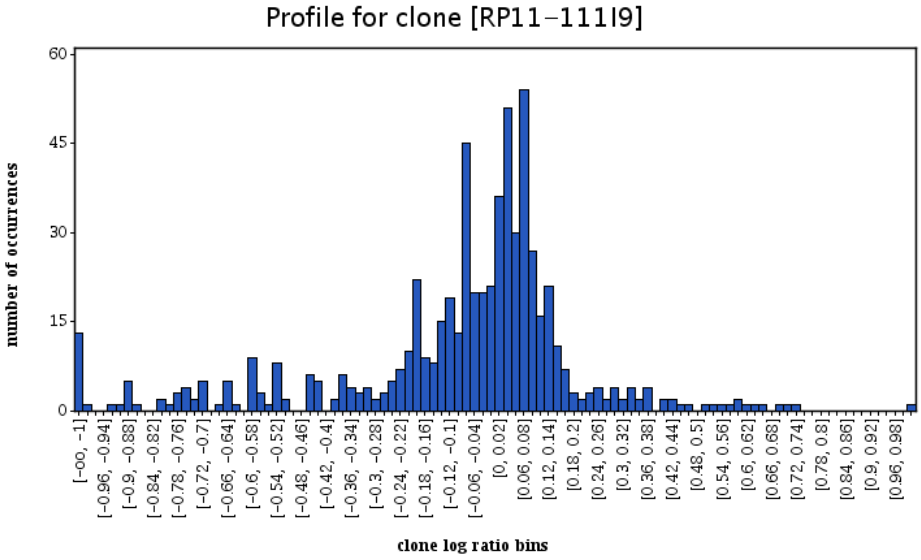


Figure 2.16: Log ratio profile over all loop design experiments for clone RP11-111I9. No DGV, SegDup, or Redon CNV loci are present in this region.

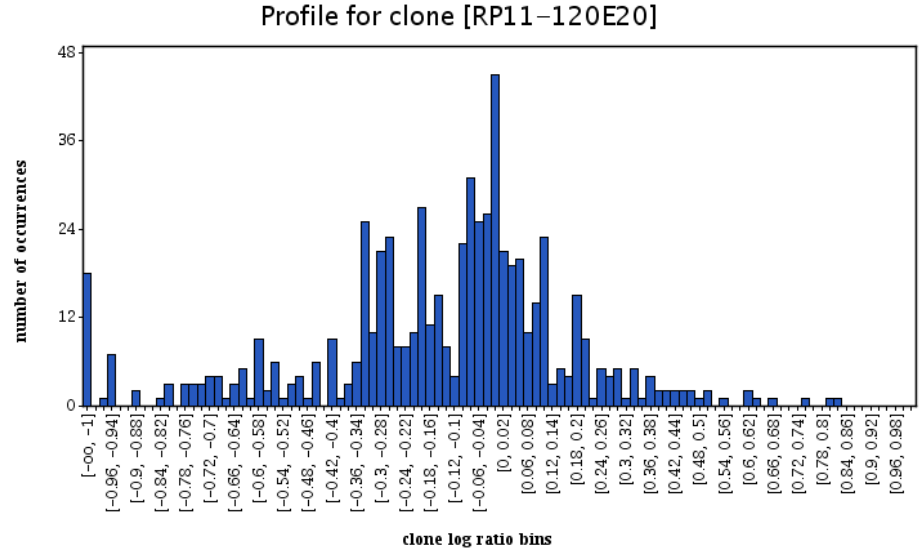


Figure 2.17: Log ratio profile over all loop design experiments for clone RP11-120E20. Multiple loci were reported for this clone in DGV, in SegDup, and in the set of Redon Copy Number Variations.

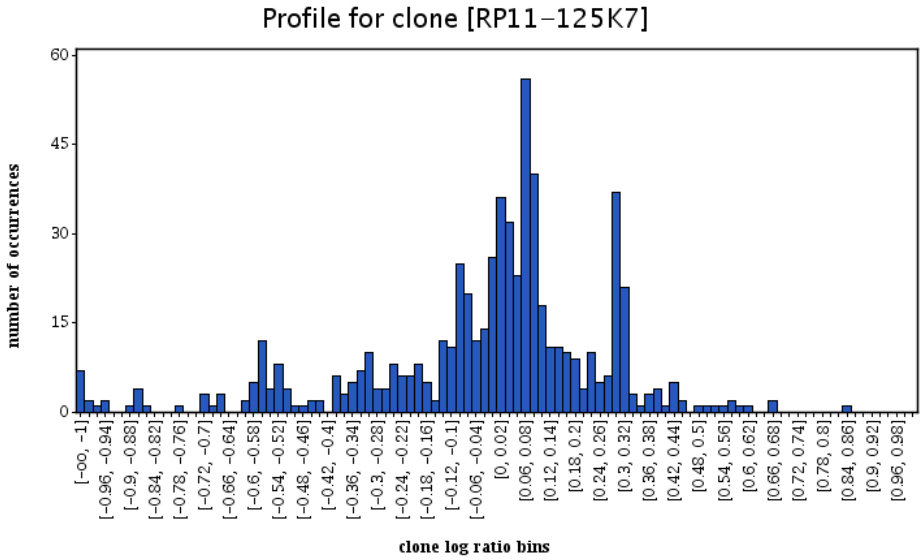


Figure 2.18: Log ratio profile over all loop design experiments for clone NONSC6D10. No DGV loci are reported in this region, but multiple SegDup and Redon CNV loci are present in this region.

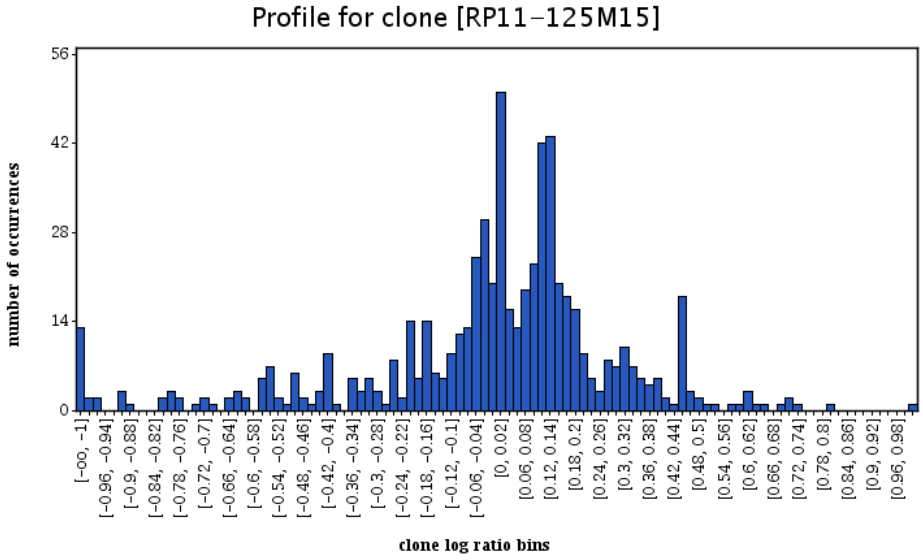


Figure 2.19: Log ratio profile over all loop design experiments for clone RP11-125M15. A DGV Copy Number variation is reported in this region, but no SegDup or Redon CNV loci are present in this region.

Although these preliminary results point to the use of data sets acquired through a diagnostic routine setting instead of a panel of normals explicitly constructed in a research context, a number of open ends remain. Apart from the technical data preparation caveats mentioned in Section 2.4.3, specific issues arise with the use of manual, visual interpretation of histograms. In this section, we will describe the issue, and discuss a possible approach to address it. By doing this, future work on this application is defined.

One of the disadvantages of the histogram approach in this application, is the fact that while histogram can be used to estimate the density of probability for a certain range of, in this case, log ratio values, histogram graphs depend heavily on the staring points of the bins, in addition to the number of bins. The effect can be surprisingly large [255]. While averaging histograms with different bin starting points can be a solution to obtain a smoother image of the clone behavior, more work is needed to obtain probability density functions in the strict sense of the definition. Specifically, if we were to obtain explicit estimations for the probability density, we could trivially infer an estimate for the fraction of the population showing the copy number variation represented by a specific mode of the histogram graph. This can be done under the assumption that if the surface under a specific peak below the main mode of the density function amounts to, *e.g.*, 10%, then in about 1 out of 10 people in the normal population, this copy number variation could expected to be present.

The non-parametric estimation of probability density functions is a large topic. Peter Konings [17] is credited for suggesting a kernel-based approach to density estimation based on histogram data. One of the advantages of this approach is that packages providing kernel-based methods are readily available for major statistical programming languages, such as R [18]. The commonest form [255] of density estimation is a *kernel density estimate* of the form:

$$\hat{f}(x) = \frac{1}{nb} \sum_{j=1}^{n} K\left(\frac{x - x_j}{b}\right) \tag{2.3}$$

for a sample $x_1, \ldots, x_n$, a fixed kernel $K()$ and a bandwidth b; the kernel is normally chosen to be a probability density function. The choice of bandwidth is a compromise between smoothing enough to remove insignificant bumps and not smoothing too much to smear out real peaks.

For the three putative polymorphic clones discussed above, the kernel estimated density functions are shown in Figures 2.20, 2.21 and 2.22.

2.4.4 Charting CNVs: conclusions

It is clear that we are only at the beginning of charting the complex variability of the human genome in a clinical and medical context. The fact however stands that the advent of high resolution full genome assays clears the path for new

[17] ESAT, K.U.Leuven

[18] `www.freestatistics.org/cran/`

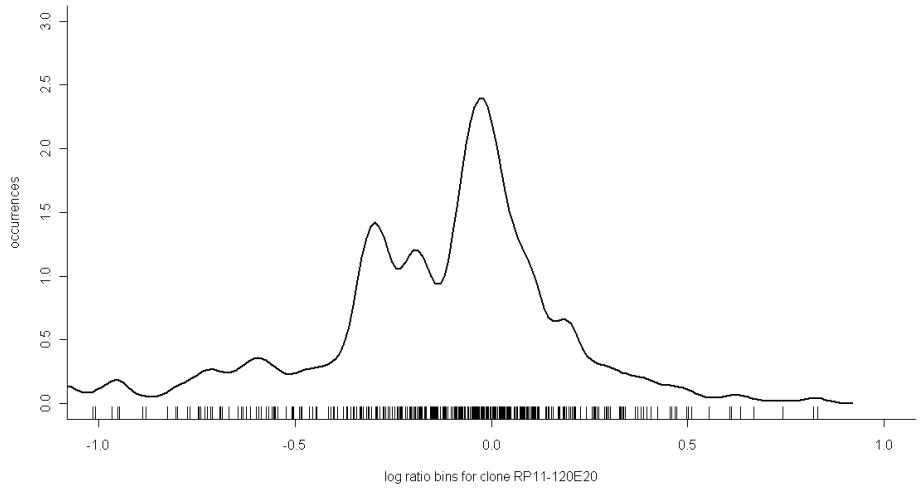


Figure 2.20: Kernel estimated density function rendered on histogram data for clone RP11-120E20. All images have been generated with Gaussian kernels on a sample size of 1000 hybridizations with a bandwith of $b = 1000^{-\frac{1}{5}}$ or approximately 0,25.

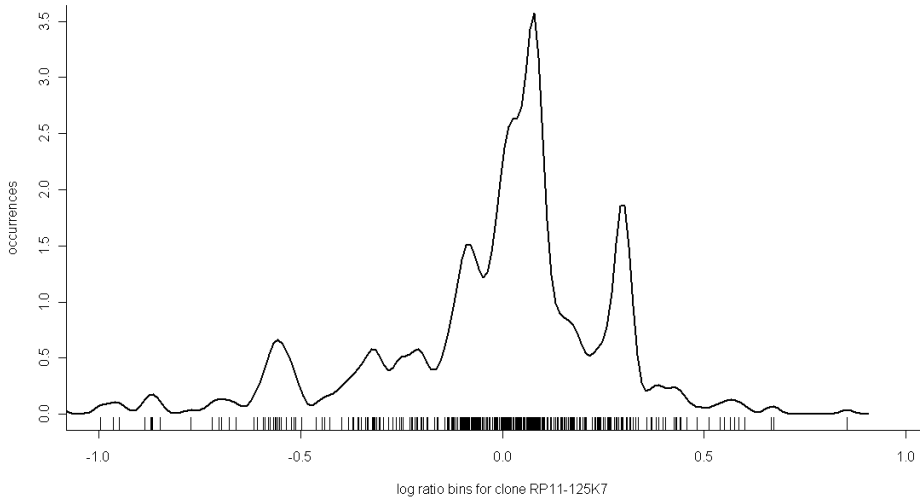


Figure 2.21: Kernel estimated density function rendered on histogram data for clone RP11-125K7.

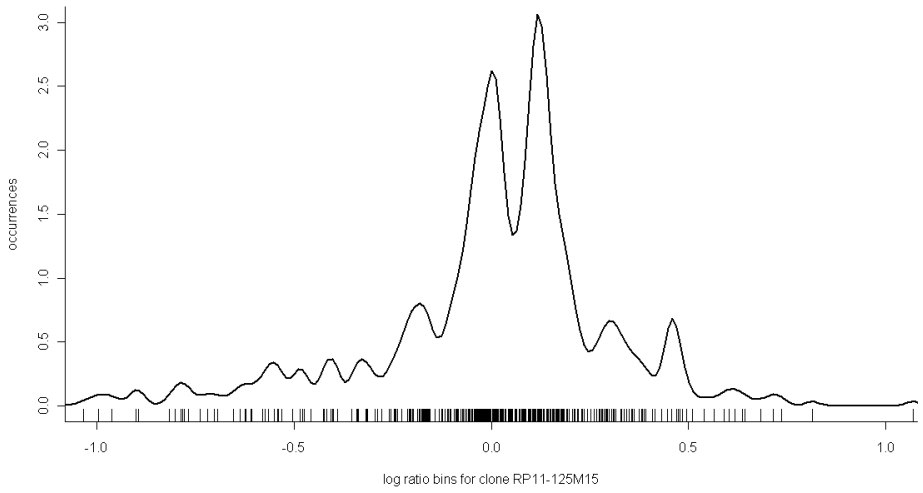


Figure 2.22: Kernel estimated density function rendered on histogram data for clone RP11-125M15.

research directions that will cause to flow clinically applicable functional genetics knowledge from research into everyday medical practice.

We have identified a number of data processing requirements that need to be fulfilled in order for the clone profiling to be more robust and for results to be more reliable. We have also identified an approach to make results more sound and interpretable. The advent of higher resolution oligo assays will add to the challenge of data interpretation, opening up the opportunity to chart copy number variations in the human genome. Interpreting both normal individuals and patient data for which phenotype information is available will in turn allow researchers to perform data mining studies that identify regions in some way implied in the development of specific phenotypes, and dive deeper into the realm of complex genetics that is defined by susceptibility loci, incomplete penetrance, multifactorial traits, and even environmental effects.

2.5 Array repositories

As already illustrated in Section 2.3, centralized data aggregation over multiple experiments is useful in both a diagnostics and research setting. This does not only serve the purpose of providing a one-stop-shop where clinicians and researchers alike can analyze and interpret the raw results of individual experiments, but, as shown in Section 2.4.2, opens up possibilities for other types of analysis of unsuspected added value, especially when these central data repositories grow large. This section deals specifically with such data repositories.

One of the challenges inherent to the technology of Array CGH is the management and analysis of large numbers of data points generated in each indi-

vidual experiment. We have discussed *LOOP*, the loop design repository, in Section 2.3. This application is web based, and doubles as an online storage application for array results that includes an analysis pipeline.

Other database systems are also conceived as analysis pipeline tools that double as storage applications, making stored data amenable to research related data mining as well as lab work flow support. The availability of the human genome sequence as well as the large number of physically accessible oligonucleotides, cDNA, and BAC clones across the entire genome has made the need for such platforms clear and has triggered and accelerated their development. ArrayCGHBase is such a database tool, and will be discussed in the next section.

2.5.1 ArrayCGHBase

ArrayCGHBase is a comprehensive analysis platform for microarray based comparative genomic hybridization (array CGH) experiments consisting of a MIAME (Minimal Information About a Microarray Experiment) supportive database underlying a data mining web tool, to store, analyze, interpret, compare, and visualize array CGH results in a uniform and user-friendly format. The database is compatible with all existing array CGH platforms (BAC, SNP, oligo, and cDNA). Data can be exported in a multitude of formats, amongst others to map copy number information on the genome using the Ensembl or UCSC genome browser, or to import data into Progenetix [18]. ArrayCGHBase is web based, platform independent and allows users to access the analysis suite through the internet or a local intranet after installation on a private server. The complete package can be freely downloaded under a Creative Commons License for local installation on a privately hosted web server. For local use, additional free software is required such as the MySQL database, Apache web server, and PHP hypertext preprocessor.

While the analysis platform has been developed at the University of Ghent, it has been part of an ongoing research collaboration with CME and was subject of a joint publication. This database was developed as a reply to BASE, a similar analysis platform that is largely focused on expression arrays. Features specific to the context of array CGH were implemented in ArrayCGHBase, with a focus on use for analysis of DNA copy number changes. ArrayCGHBase is described in a manuscript by Menten *et al.* [159].

Rationale of development

ArrayCGHBase is not only discussed in this dissertation because of the fact that it was part of a joint research project and a co-publication. More importantly, ArrayCGHBase was the first application of its kind to combine several key aspects in Array CGH Data storage and analysis, where up until then only separate software systems were available that individually addressed some of the needs of the analysis pipeline. These aspects include the need for databases for data storage, such as the BioArray Software Environment (BASE), a plat-

form for comprehensive management and analysis of microarray data [201]; applications for clustering and visualization of microarray data (seeGH [43], M-CGH [264], CGHAnalyzer [151, 88], aCGH-smooth [117] and ChARM [166]). Other aspects are the need for an interface with public genome databases that contain reporter information, functionalities that adhere to those of Laboratory Information Management Systems (LIMS), and various storage methods for recording biomaterial annotations. Since none of the software packages or databases available at that time combined all necessary features, a web based open source array CGH analysis platform, ArrayCGHBase, was developed.

Key features

ArrayCGHBase is an open source package developed in PHP, a hypertext preprocessor [19], that stores all data in a relational database in a MIAME [32] (Minimal Information About a Microarray Experiment) compliant format. The tool integrates DNA sample information, lab protocols, extracted data, and contains a plug-in architecture for data transformation, analysis, and graphical display, allowing users to develop their own modules. It features user and project administration, allows management of DNA samples, annotation of lab protocols, and it runs analyses. Data is exportable in several formats. Key features are its data processing and visualization routines, input data and local background correction and normalization, QC (Quality Control), and scoring of chromosomal regions with aberrant copy number. The tool boasts various useful chromosome visualizations, that are rich and interactive.

2.5.2 Towards functional interpretation and intelligent databases

So far, we have described quality criteria and threshold approaches for the introduction of Array CGH into clinical practice, and we have discussed how data analysis and storage is done for a specific type of array designs, coupled with a more detailed analysis on copy number polymorphism based on thresholds, and on data gathered in the loop design assay repository. We have then described storage and analysis platforms that provide stronger visualization and analysis functions. We have discussed ArrayCGHBase, which links to R and BioConductor to offer region calling algorithms, that allow – when tuned – the unsupervised step from individual clone log ratio calls to regions of mostly adjacent clones that point to a continuous copy number aberration in the patient genome at that specific locus.

The next and necessary step is to further investigate the information that is obtained, whether it be a set of log ratio calls, or a list of called aberrations; and to interpret the results in a diagnostic context. This task surpasses the mere need of robust search, storage, and strong visualizations. This next step is an absolute requirement to guarantee a high throughput diagnostic laboratory work flow: typically, tens and often hundreds of aberrations will result from a

[19] `www.php.net`

standard region calling algorithm executed on the raw results from a high density assay, and clinical interpretation in the context of the patient's phenotype and clinical information becomes a tedious, long, manual and error prone task. Hence, a next generation of genetic assay platform is needed, lying closer to the realm of biomedical research, but strongly focusing on the optimization of the lab work flow.

ArrayCGHBase proves very useful as an assay database storage, and raw data analysis and visualization tool. Further on in this dissertation, we will describe how the analysis pipeline can be completed by offering tools to diagnostically interpret the results in the context of phenotypes, the historical aspect of the database itself, both research and diagnostic collaboration between geographically separated labs, external databases, web sites and information sources, and especially, the body of existing biomedical knowledge that is present in publicly available literature, and is made amenable to interpretation through text-mining. This lead us directly to the next two chapters of this dissertation, where we discuss how test results (*i.e.*, regions of aberrant copy number and the genes that lie in those regions) can be interpreted. For this, we will discuss tools that were developed in the context of this PhD as well as external tools that use knowledge from literature and a variety of other data sources to place genomic loci in their functional context, and where, in a subsequent chapter, intelligent tools and databases with a direct impact on diagnosis in clinical routine are discussed, focusing on all above aspects, on supervised or automatic linking between the genotype and the phenotype of a patient, and even including additional features like automated reporting.

Chapter 3

Functional interpretation of genetic information

"Nothing in biology makes sense except in the light of evolution."

Theodosius Dobzhansky (1973)

In the previous chapter, we have mainly discussed raw data analysis and quality criteria for the introduction of Array CGH into clinical practice. We have touched on the complex issue of benign polymorphisms, and saw a way to analyze loop design assays. Now, a next and necessary step is to make sense of this genomic information, and to put lists of called aberrations containing genes in a biomedically relevant context. This chapter will later lead us to the part of this dissertation where we elaborate on actual clinical interpretation: in these sections, we discuss how test results (i.e., regions of aberrant copy number and the genes that lie in those regions) can be interpreted. Tools, both developed at ESAT and elsewhere, will be described which use knowledge from literature and a variety of other data sources to place genomic loci in their functional context.

Array CGH is used increasingly often as a primary genetic screening method in diagnosis and research [215, 62, 203, 177]. The technique is uncovering pathogenic sub-microscopic chromosomal imbalances in patients with developmental disorders. Most patients carry different chromosomal anomalies, and anomalies occur across the whole genome [158, 111, 73, 51, 212]. These imbalances pinpoint the location of genes that are involved in human development [262]. Since most imbalances encompass regions harboring multiple genes, the challenge is (1) to identify those genes responsible for the specific phenotype and (2) to disentangle the role of the different genes located in an imbalanced region.

We start with an overview of tools and databases that are available to identify candidate genes for novel submicroscopic chromosomal imbalance syndromes. This work was published as a review paper in Genetics in Medicine. It provides both a wide overview of tools available to manage candidate genes,

and it will serve as an introduction to sketch the context for the three tools that will be discussed further on, and that were developed or co-developed in this PhD work: aBandApart, which uses literature to find links to genomic banding patterns and human phenotypes and diseases; aGeneApart, which does the same for genes, and Endeavour, a tool that uses not only literature but many more data sources for candidate gene prioritization.

3.1 Candidate genes for novel submicroscopic chromosomal imbalance syndromes

The high resolution at which array CGH has been used to define candidate regions for putative genes responsible for human genetic diseases is instrumental in defining and refining the critical region for a disease or phenotype and reducing the number of candidate genes for (an aspect of) the phenotype. [215]. This higher resolution has led to a dramatic increase in gene identification through molecular karyotyping, and it is likely that the function of many more genes will be identified in this way [50].

However, some specific challenges apply to correlating genotype and phenotype in the context of human disease. First, it is clear that etiology of rare chromosomal imbalances greatly benefits from large-scale efforts in collection and organization of case reports from different genetic testing and research centers around the world. Especially for rare diseases, the need for large and well-annotated case report resources is obvious.

Secondly, the identification of critical genes and pathways involved in a disease or biological process is helped by interpreting aberrations within the context of broader knowledge [177]. In understanding the functional basis of genetic conditions, it is therefore instrumental to incorporate information from different sources, other than mere genotype and phenotype information present in case reports. Integration of publicly available data sources pertaining to genome and gene function permits the development of bio-informatics methods for candidate gene selection. Such information sources range from the large body of biomedical literature to protein-protein interaction, pathway, and genome annotation databases in general.

In this section, we discuss relevant databases that have recently been developed to elucidate the role of genes in different aspects of the phenotype. We go on by giving an overview of published methods and tools that can help in the gene discovery process. Finally, we identify relevant issues in management and use of genotype-phenotype databases, and elaborate on issues encountered when annotating phenotypic characteristics to patient case reports ('phenotyping').

3.2 Computational approaches in correlating genotype and phenotype

A primary goal in the context of constitutional cytogenetics is elucidating the role of genes in different aspects of a phenotype without falsely associating normal variations to disease. Although the databases mentioned in the following sections enable associations between phenotype and genotype to be stored, queried, annotated, and exchanged, they fall short in identifying the genes underlying the phenotypic anomalies. Many tools and methods have been set up to filter high-probability candidates. In what follows, we provide a list of the most important resources and published computational methodologies to generate genotype-phenotype leads and select gene candidates for further investigation.

3.2.1 Querying genotype-phenotype correlations in literature databases

With a focus on the human as the organism of interest, a primary resource for connecting genes to disease-related phenotypes in a general rather than in a case-based manner is the Online Mendelian Inheritance in Man database (OMIM). It contains curated records of genetically inherited human disorders with references to causative genes or genetic loci. Despite the highly reliable information it contains, its usefulness in computational analyses remains limited due to the unstructured way in which the phenotypes are described. Therefore, several approaches have arisen to transform the information in OMIM to be amenable for computational analysis. van Driel *et al.* [249], for example, created a human phenotype similarity map by text-mining OMIM. Starting from an OMIM record or disease name, their MimMiner application (`www.cmbi.ru.nl/MimMiner`) retrieves the most phenotypically similar disorders. Based on the observation that similar phenotypes are often caused by functionally related or interacting genes, a researcher can easily go through the list of genes associated with similar phenotypes to select leads for further investigation. Most of the methods for computational prioritization of disease genes described in this section (see Table 3.1) make use of the information in OMIM, either as a starting point to further investigate genotype-to-phenotype associations [72, 77, 2, 250], or as a reference to benchmark their proposed method [5, 176, 69, 243, 129, 170].

Akin to OMIM, the MEDLINE database of biomedical literature contains a large amount of useful information on genotype-phenotype relations in free-text format. Here, too, several published approaches exist to quickly guide researchers to information relevant to their research interest. The iHOP resource created by Hoffmann *et al.* [101] provides an intuitive access to the published literature by hyper-linking abstracts and sentences via the gene or protein symbols and names they contain. The approach discussed later in this chapter (Section 3.4.1) uses over-representation statistics to correlate specific biomedical terms from various targeted biomedical vocabularies with cytogenetic bands cited in MEDLINE abstracts. There, through a web application named aBan-

dApart (Section 3.4.1 and published manuscript [253]), researchers can easily fetch the most relevant concepts for a genetic region of interest, look up chromosomal bands associated with a query term, and retrieve related literature. Another rich source of phenotypical information about genes is provided by Entrez Gene's GeneRIFs (Gene References into Function) [160]. These direct associations between genes and published literature allow construction of accurate textual representations of a gene using standard text-mining techniques [85, 86]. Yet only Aerts *et al.* [5] and Lage *et al.* [129] make use of this valuable information source for candidate gene prioritization.

Other noteworthy resources include PhenomicDB [90] and PhenoGO [145]. PhenomicDB's value lies in how it integrates phenotype information from multiple species-oriented databases into one repository. Phenotype-genotype associations are grouped based on gene orthology to allow exploration across different species. PhenoGO is mentioned here because it is a good example of how more advanced text-mining techniques can help to bridge the gap between functional annotations of genes and phenotype descriptions [145]. The PhenoGO system uses Natural Language Processing of MEDLINE abstracts to connect phenotypic contextual information with GO annotations (and hence genes) and other biomedical ontologies.

3.2.2 Finding phenotype-rich genotypical features

Apart from retrieval of phenotype information from databases, certain aspects (or features) of the gene or protein sequence can also be used to infer or predict associated phenotypes. For instance, it is known that disease genes tend to code for longer proteins and are in general more highly conserved through evolution. Both López-Bigas and Ouzounis [139] and Adie *et al.* [1] take this approach to calculate the correlation between genes and disease. Both started from the list of genes known to be involved in hereditary disease in the morbid map table of OMIM to define discriminating features and subsequently classified all known human genes using a decision tree based model. López-Bigas and Ouzounis used information about length, phylogenetic extent, degree of conservation, and paralogy of proteins in their Disease Gene Prediction (DGP) method. For their Prospectr method, Adie *et al.* used a more elaborate feature set that reflects the structure, content, and evolutionary conservation of both the DNA and protein sequence. The outcome of both methods is a score that indicates the probability of a gene to be causal to disease.

While these kinds of methods provide valuable information, their applicability in connecting genes to specific phenotypes or diseases remains limited. This is mainly because they do not rely on the existing knowledge of a particular disease, contrary to the methods described in the following paragraph.

3.2.3 Pinpointing phenotype-related genes: guilt by association

The published methods described here can be broadly divided in three categories. An overview of methods is presented in Table 3.1. The first category covers *ab initio* methods. These try to identify genes by defining whether their characteristics or features are related to a specific disease. Examples of such features are genomic location (*e.g.*, within a linkage region), sequence features, sequence phylogeny, functional annotation, gene expression, etc. Most methods take into account a combination of features to prioritize the candidates. The method by Turner *et al.* [246], POCUS, calculates statistical over-representation of Gene Ontology annotations and InterPro protein domains for genes in a given set of genomic loci to identify successful leads. GeneSeeker is a web application implemented by van Driel *et al.* [250] that filters the genes in a certain region based on user-specified characteristics of interest (tissues, phenotypic features of a syndrome, etc.). The Genes2Diseases (G2D) application presented by Perez-Iratxeta *et al.* [176] calculates the association between a gene and a disease based on the co-occurrence in a set of MEDLINE abstracts of MeSH terms in the 'Diseases' and 'Chemical and Drugs' categories with the gene's Gene Ontology annotations. Other methods in this category include TEAM [68], the method by Tiffin *et al.* [243], and the genomic convergence approach described by Hauser *et al.* [97].

A second category of methods to link genes and phenotypes are network methods. Here, the emphasis is on the creation of an interaction network of genes or proteins. The rationale behind these methods is that similar phenotypes are often caused by functionally related genes, (i.e., genes that belong to the same functional process, take part in related pathways, or code for proteins that are part of the same protein complex). They differ mainly in the way the protein network is constructed and how interaction partners of known disease proteins are associated with known disease phenotypes. Franke *et al.* [69] created a Bayesian classifier to first predict protein-protein interactions not present in a gold-standard data set, using GO annotation, gene co-expression, and protein-protein interaction data. Then, their Prioritizer application establishes whether candidate genes in known disease loci are closer together in the network than expected. Oti *et al.* [170] use a hybrid protein-protein interaction network to find interaction partners of known disease proteins. They go on by checking whether the genes coding for these partners were in a disease associated locus for which no genes were previously identified. Lage *et al.* [129] follow a similar approach by constructing a quality-controlled human protein interaction network and deriving candidate protein complexes that contain the product of each of the candidate genes involved. The input phenotype is then compared to phenotypes of disease causing proteins present in these complexes and the protein coding candidate genes are scored accordingly using a Bayesian predictor.

We identify similarity methods as a third category, because here, prioritization of candidate genes is based on similarity between candidate and known disease genes, rather than on putatively involved features or on their presumed

disease causing interaction partners. The Endeavour application by Aerts *et al.* [5] uses a sound statistical framework based on order statistics to reconcile a large number of different data sources. Used data includes both existing knowledge (literature, functional annotations, pathway information, etc.) and experimentally derived data (gene expression, protein interaction, etc.) to balance out a bias towards known genes. Candidate genes are compared to a user-selected or automatically retrieved list of training genes that represent the disease or phenotype under study, and prioritized according to their similarity with the training set thus obtained. It is worth noting that Endeavour is one of the only computational methods with which new disease gene was revealed that was later validated *in vivo*, was revealed [5]. Adie *et al.* [2] devised a similar method named SUSPECTS. Here, a more generic candidate gene scoring approach is used. Contrary to Endeavour, the set of training genes can not be customized, only four data sources can be included in the analysis, and the method shows no flexibility towards filtering the candidate genes at a user-defined locus. A new method, CAESAR, is described by Gaulton *et al.* [77] and is the most recent addition to similarity based methods. It also takes a more generic approach in scoring candidates. CAESAR uses a myriad of different data sources and integrates similarity measures from these different information spaces through the use of four different arithmetic operations. Although already an older method, the approach of Freudenberg and Propping [72] is also worth a mention in this category.

It must be noted that some approaches can be classified under more than one category. This is, for instance, the case with the CGI method of Ma *et al.* [147] in which *ab initio* derived gene-condition co-expression bi-clusters are combined with data from a protein interaction network. Endeavour also takes into account data from known protein interaction networks (BIND and Kegg) and, like SUSPECTS, from disease probabilities described earlier (DGP and Prospectr).

An upcoming trend in computational gene identification is the use of a concert of prioritization methods based on a combination of prioritization results. This approach was presented by Tiffin *et al.* [242] and Elbers *et al.* [57] who both conducted a study to find genes commonly associated with obesity and type II diabetes.

3.3 Challenges for automated genotype-phenotype correlations

As more and more biological data are stored on computers, the problem of efficient retrieval and analysis of these data becomes the most important scientific bottleneck. This problem is particularly acute in biology because biological data are notorious for their complex form and semantics [231]. Case report databases can only become valuable public or private resources when the data is of sufficient quality and is rigorously evaluated, annotated, and interpreted within the

Computational method	Available online	Resources used	Application to complex traits	Supported species	In-vivo validation	Data integration method	Web site and reference
POCUS (2003)	No	InterPro, GO, UniGene	-	Human	No	Over-representation statistics	Turner *et al.* [246]
TEAM (2004)	Yes (download)	GO, gene expression data	Yes	Human	No	Filtering functionality	humgen.med.uu.nl/~lude/team, Francke *et al.* [68]
(Tiffin *et al.* 2005)	No	MEDLINE, eVOC	-	Human	No	Term co-occurrence statistics	www.sanbi.ac.za/tiffin_et_al [243]
GeneSeeker (2005)	Yes	MGD, GDB, MEDLINE, OMIM, UniProt, GXD	-	Human, mouse	No	Boolean logic	www.cmbi.ru.nl/GeneSeeker/ [250]
G2D (2005)	Yes	MeSH, MEDLINE, GO	Yes	Human	No	Fuzzy set theory	www.ogic.ca/projects/g2d_2/[176]
(Freudenberg & Propping 2002)	No	OMIM, GO	-	Human	No	Generic scores	[72]
SUSPECTS (2006)	Yes	OMIM, HGMD, GAD, Prospectr, InterPro, GO, gene expression data,	-	Human	No	Generic scores	www.genetics.med.ed.ac.uk/suspects/ [2]
Endeavour (2006)	Yes (download)	MEDLINE, EST, KEGG, GO, TRANSFAC, Jaspar, InterPro, BIND, DGP, Prospectr, gene expression data	Yes	Human, mouse, fly	Yes	Order statistics	www.esat.kuleuven.be/endeavour/ [5]
Caesar (2007)	No	MP, eVOC, GO, OMIM, Entrez Gene, EnsEMBL, UniProt, InterPro, BIND, HPRD, KEGG, MGD, GAD	Yes	Human	No	Generic scores	visionlab.bio.unc.edu/caesar/ [77]
(Oti *et al.* 2006)	No	HPRD, DIP, interactions from high-throughput experiments	-	Human	No	Generic approach	[170]
Prioritizer (2006)	Yes (download)	BIND, HPRD, Reactome, KEGG, GO, SMD, GEO, GeneNetwork	Yes	Human	No	Bayesian classifier	www.prioritizer.nl [69]
CGI (2007)	No	Yeast gene expression compendia (knock-out, stress response, and cell cycle), MPPI, GO	Yes	Yeast, human	No	Markov Random Field theory	[147]
(Lage *et al.* 2007)	No	MINT, BIND, IntAct, KEGG PPrel, KEGG ECrel, Reactome	Yes	Human	No	Bayesian classifier	[129]

Table 3.1: An overview of published methodologies for gene prioritization. Remarks: the 'resources used' column only contains the data sources used in the prioritization methodology, not the data sources used to validate or benchmark the approach; the column Application to complex traits contains Yes if the paper describing the method explicitly mentioned its applicability to or application on complex traits, a dash otherwise; the Supported species column contains the species the method was applied to previously, not necessarily all species the method could be applied to

richest possible context [177]. This is not a straightforward task. In a review paper on novel computational tools that allow researchers to amass, access, integrate, organize and manage phenotypic databases, Lussier and Liu [146] state that the development of phenotypic databases lags behind the advance in genomic databases, and creates the need for novel computational methods to unlock gene-disease relationships. In the next section, we will discuss database quality issues with regard to phenotyping in the context of chromosomal aberration and phenotype databases. We also discuss some specific challenges in gene prioritization for constitutional cytogenetics.

To avoid an inconsistent evaluation of a phenotype by multiple clinical geneticists, a single observer can be designated so that classification criteria can be uniformly applied (for example, Zhang *et al.* [276]). In their genotype-phenotype mapping efforts for Cri du Chat Syndrome, Zhang *et al.* attributed inconsistent results in previous mapping efforts partly to issues regarding inconsistent evaluation of the phenotype by multiple observers and lack of consideration for age dependence of prominence of phenotypic characteristics.

Appointing a single observer is clearly impossible for studies on rare disorders where case reports are gathered from a multitude of genetic research and testing centers. However, for case reports to be amenable to large scale data integration, exchange, and mining, phenotypic annotations need to be uniform and unbiased. Formalizations are crucial for organizing and executing experiments, as well as storing and sharing the experiment results [231].

In a context that spans multiple diagnostic or research entities (cross-departmental, cross-clinic, or in international collaborations), the proper use of dedicated ontologies can partly address this issue, allowing clinical geneticists to use a uniform vocabulary of clearly defined phenotype features to annotate case reports. Sound ontologies are instrumental to mapping function to gene products in the genome [241, 60]. However, even if a detailed and highly descriptive standardized vocabulary of phenotype characteristics is available, some important issues remain. Firstly, phenotypic traits can be age dependent or linked to a certain developmental stage, and can evolve over time. Secondly, phenotypes can vary in penetrance or severity, leading to the need for qualifiers and not just concepts. Thirdly, across databases, phenotype annotations often happen at different levels of granularity, in different formats, and with different aims [146].

Representation of phenotypic information is more complicated than biological data, and consequently there are few data standards and models for managing phenotypes within human repositories. With OMIM as an example, Lussier *et al.* state that while OMIM has the largest collection of human diseases [146], the unstructured narrative content of its phenotypes makes it unsuitable for computational analysis, data mining and fusion, and integration between databases, as was mentioned before. It is clear that, in addition to proper interpretation of clinical features, unambiguous and complete identification and annotation of developmental anomalies, dysmorphic features, and any phenotype aspect in general is crucial for databases to be useful and interoperable. Several common terminologies to describe phenotypic aspects of a patient are presently available. Some of them can be licensed or obtained under

certain conditions, others are freely available. Some well known ontologies and vocabularies are listed in Table 3.2.

Ontology	Description
GO	Gene Ontology, a systematic terminology for functional features of genes and proteins
ICD-9	International Classification of Diseases Clinical Modification
LDDB, LNDB	Oxford Medical Dictionary London Dysmorphology and Neurology Databases
MPO	Mammalian Phenotype Ontology
SNOMED	Systematized Nomenclature of Medicine
UMLS	Unified Medical Language System, groups and links a host of ontologies

Table 3.2: Well known and widely used ontologies and vocabularies relating to phenotypic traits and human disease

DECIPHER and BENCH (formerly CGHGate, a database tool for storage, reporting and mining Array CGH Case reports, `www.cartagenia.com`) make use of a structured vocabulary present in LNDB (London Neurology Database), a hierarchical list of human dysmorphology concepts.

As both DECIPHER and BENCH surpass the level of mere functional interpretation of genetic information and firmly belong to the realm of intelligent databases for clinical and diagnostic interpretation of both genotype and phenotype information, they will be discussed in great detail in the next chapter.

Both databases use human dysmorphy vocabularies, taxonomies or hierarchies. Although LNDB is adequate for describing human dysmorphology in the context of constitutional developmental disorders, this vocabulary has some issues with regard to disambiguation between concepts and with uniqueness and consistency of identifiers. It was not designed to be used as a standard for phenotype annotation amenable to mining, database integration and automated annotation.

OMD LNDB is not the only vocabulary that suffers from such issues. Soldatova [231] states that ontologies are often primarily designed to provide biologists with a common vocabulary for standard annotation purposes, and are not always structured with standard practice in mind. This approach is not compatible with the increasing use of computational reasoning in biology and its dependence on ontological data. Soldatova further states that although expert biologists may be able to deal with poorly designed and inconsistent ontologies, this is not currently possible for computer programs that do machine learning or text-mining. As such programs are set to dominate the analysis and retrieval of biological data, Soldatova argues that biological ontologies should be designed with these needs in mind as well.

Some challenges are specific to gene prioritization for human development and constitutional cytogenetics. For one, it is important to note that phenotype

characteristics are often complex traits that are not a function of state, but rather an end or even intermediate point that can be reached through different and very unrelated developmental processes. In short, variations or mutations in different genes may yield identical or related phenotypes. This contributes to the complexity of gene prioritization for phenotype traits. Secondly, environment interactions during human development are likely to be an important cause of heterogeneity in phenotype. Attribution of phenotype traits not only to the genotype but also to the environment (nature vs. nurture) increases the order of complexity of the task at hand.

While parts of a phenotype can be explained by the action of a single gene, other characteristics are caused by multiple genes. For this reason, it is important that tools for phenotype based candidate gene prioritization are conceived with complex disorders in mind.

Positional or epigenetic effects may play a role in developmental disorders, so that genes responsible for the phenotype may actually lie outside the aberrant region.

As discussed earlier in this dissertation, Redon *et al.* [190] recently drew attention to the large extent to which nonpathogenic copy number variations are present throughout the human genome through analysis of Array-CGH and single nucleotide polymorphism genotyping data. The authors show that this affects 12% of the human genome, around the same level as single nucleotide polymorphism variation. Understanding benign copy number polymorphism is further complicated by the fact that some so-called normal variation may underlie a phenotypic characteristic such as disease susceptibility [177] or involvement in a late-onset phenotype.

In the following sections, we will elaborate on specific examples of tools that aid functional interpretation of genetic information. Specifically, aBandApart, aGeneApart, and Endeavour will be discussed. In the next chapter, we will discuss ways to address the issues mentioned in the sections above, and focus on clinical interpretation, phenotype annotation, collaborative aspects of patient repositories, and on data mining, reporting and visualization through intelligent database tools.

3.4 Statistically linking genotypes and phenotypes

3.4.1 Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations

Biomedical literature provides a rich but unstructured source of associations between chromosomal regions and biomedical concepts. By mining MEDLINE abstracts, we annotate the human genome at the level of cytogenetic bands. The method described in the following sections creates a set of chromosomal aberration maps that associate cytogenetic bands to biomedical concepts from a variety of controlled vocabularies, including disease, dysmorphy, anatomy, development,

and Gene Ontology branches. The association between a band (*e.g.*, 4p16.3) and a concept (*e.g.*, *microcephaly*) is assessed by the statistical over-representation of this concept in the abstracts relating to this band. This method is validated using existing genome annotation resources and known chromosomal aberration maps, and is further illustrated through a case study on heart disease. Our chromosomal aberration maps provide diagnostics support to clinical geneticists, aid cytogeneticists to interpret and report cytogenetic findings, and support researchers interested in human gene function. The method is available as a web application, aBandApart, at `www.esat.kuleuven.be/abandapart/`, and is discussed in further detail in a published manuscript [253].

3.4.2 Identification of loci involved in development, dysmorphy and pathology

Forward genetics, i.e. identification of gene mutations that underlie a phenotype of interest in a particular individual, is a key strategy to characterize gene function. In humans, where mutagenesis screens are impossible, genomic information from patients with developmental disorders can serve as the basis for disease gene discovery. Different positional cloning strategies, such as cytogenetic studies and linkage and association studies, can subsequently identify the chromosomal region where the disease gene is located.

To speed up the process of gene discovery, some attempts have been made to associate genomic rearrangements (such as sub-chromosomal deletions and duplications) to congenital malformations based on clinical and cytogenetic information from patients. Brewer *et al.* analyzed detailed clinical and cytogenetic information associated to a large number of autosomal deletions [34] and duplications [35] to construct a chromosome map showing associations of congenital malformations and chromosomal regions. Notably, these maps have not been updated since their publication.

Research groups with an interest in the etiology of, for example, congenital malformations often lack an extensive pool of patients to conduct large and informative association studies. Several public and private databases are being constructed to support such efforts by aggregating case reports and encouraging the exchange of patient information to complement private patient pools. Examples are the Catalogue of Unbalanced Chromosome Aberration in Man [207], the Human Cytogenetics Database and ECARUCA [1], DECIPHER [2], the Chromosome Anomaly Collection [3], the Mitelman Database of Chromosome Aberrations in Cancer [4], the Mendelian Cytogenetics Network DataBase [5], Orphanet [6], *etc.* These efforts differ in setup but aim at aggregating chromosomal aberration information and charting phenotypes and case reports. Some catalogs are avail-

[1] `www.ecaruca.net`
[2] `decipher.sanger.ac.uk`
[3] `www.som.soton.ac.uk/research/geneticsdiv/`
[4] `cgap.nci.nih.gov/Chromosomes/Mitelman`
[5] `www.mcndb.org`
[6] `www.orpha.net`

able only in print or at a licence fee, and some databases require registration. Other are open and searchable by the public, but include no specific means for data mining.

The information available in the public corpus of biomedical literature is a powerful alternative resource for patient reports and cytogenetic findings to conduct association studies. This corpus can be seen as a *de facto* genotype-phenotype association database. Moreover, it is not limited to case reports listing congenital malformations. Apart from disease related concepts, it is a rich source of information with regard to anatomy and development, systems and tissues, and molecular functions and biological processes as well. An example of such a relevant abstract is shown in Image 3.1, where both phenotypes as cytogenetic banding patterns are highlighted.



Figure 3.1: Example of a PUBMED abstract containing relevant concepts: phenotypes are highlighted in blue, banding patterns in green for duplications and red for deletions.

In what follows, we describe a method we have developed to automatically create chromosomal aberration maps from MEDLINE abstracts that mention (ranges of) cytogenetic bands. Through the use of multiple structured vocabularies, association with a band is not limited to a disease or syndrome, but also covers dysmorphology, human development, and cell biology. The online application built on this method forms a bridge to the relevant and most current literature for further analysis by the researcher, rather than merely providing a catalogue of genotype-phenotype associations. It thereby facilitates studies in the etiology of disease and the identification of disease genes. This resource is

freely accessible and will stay up-to-date through regular automatic updates.

Related literature-mining methods

A number of tools and methods are currently available and offer capabilities for mining associations from literature between disease and genomic locations, although none have a scope identical to our method.

G2D [176] is a method for the prioritization of genes according to their relation to inherited disease. It allows a user to enter an OMIM disease identifier and a genomic region of interest. Through sequence and biomedical database analysis, G2D then identifies genes potentially associated with the disease.

HCAD [100] (Human Chromosome Aberration Database) is a web-based text-mining tool supporting analysis of human breakpoint data by mining the scientific literature to generate information on all human breakpoints.

Korbel *et al.* mine MEDLINE to identify clusters of gene-phenotype similarities based on information on prokaryotic genomes [125]. The results are not available through a web interface.

Tiffin *et al.* use an anatomical ontology to integrate text-mining of biomedical literature and data-mining of available human gene expression data [243]. Their method prioritizes candidate genes according to their expression in disease-affected tissues.

iHOP [102] (Information Hyperlinked over Proteins) uses genes and proteins from multiple organisms as hyperlinks between sentences and abstracts to access and navigate PubMed.

MimMiner [249] is restricted to mining the OMIM database and ranks related phenotypes for a given phenotype or OMIM identifier. GeneSeeker [250] is a related tool that aims at the identification of genes underlying human genetic disorders by combining data on cytogenetic locus, phenotypes and expression patterns, to generate a list of candidate genes.

GFINDer [154] mines text data present in OMIM to annotate genes with gene ontology concepts and statistically selects relevant annotation categories. Phenotype descriptions are normalized to handle synonymy and are hierarchically structured.

Our method relates to these approaches but differs in several aspects. Firstly, instead of extracting MEDLINE references linked to OMIM entries, or mining only text present in OMIM, MEDLINE abstracts are directly mined for cytogenetic bands and biomedical concepts. While curated databases offer high quality annotations and hence reduce noise, the use of abstracts allows mining to be more complete and up to date. Secondly, gene prioritization tools like G2D build an internal representation for the disease or phenotype under study through the intermediate association of MeSH and GO [9] terms. This allows relating genes to phenotypes by means of chemicals, molecules, etc. In our method, this internal association process is rendered explicit through the choice of controlled vocabularies that allow the user to elucidate over-represented associations between loci and concepts. Thirdly, most of these tools offer only a disease specific approach (in some cases using other annotations internally)

whereas aBandApart explicitly allows for additional perspectives or user interests, such as dysmorphology, anatomy, development, molecular function, etc.

The aBandApart approach is a novel analysis method based on abstracts present in MEDLINE for cataloguing biomedical concepts according to their association with chromosomal bands, which can be considered as a cytogenetic approach to genotype-phenotype correlation. Rather than prioritizing candidate genes, it focuses on cytogenetic bands and offers a portal into relevant literature. Through its different approach and goal, it can be considered complementary to tools that already exist.

3.4.3 Automatically building chromosomal aberration maps

Three elements are necessary to automatically build a chromosomal aberration map from MEDLINE abstracts: (1) identification of cytogenetic bands, (2) identification of concepts from multiple vocabularies, and (3) assessment of the statistical over-representation of a concept among the abstracts relating to a band.

To discover over-represented association between concepts and cytobands, we must first locate cytogenetic band identifiers and concepts from the vocabularies (and their synonyms) in the MEDLINE corpus. We have extended Lucene [96], a high-performance text indexing engine written in Java, to parse all MEDLINE abstracts and extract cytogenetic bands, ranges of bands, and biomedical concepts that are present in our different structured vocabularies.

Identification of cytogenetic bands

The International System for Human Cytogenetic Nomenclature (ISCN) gives a universal terminology of the description of chromosomal anomalies based on cytogenetic staining techniques [216]. This nomenclature guarantees that all chromosomal anomalies are reported in a standardized way. Hence, reports in literature typically mention bands to delineate a genomic region at various levels of cytogenetic resolution. Because of this specific nomenclature, bands can be unambiguously extracted from text in the majority of cases. A similar approach is adopted in HCAD, where the nomenclature for translocations is used.

Although band patterns delineate chromosomal regions at a less detailed resolution than markers, base-pair positions, BAC clone identifiers, or genes, this approach is advantageous because of its effectiveness. Indeed, in most cases, chromosomal deletions and duplications have so far been resolved and reported only if their size was of the order of a cytogenetic band. Also, more accurate identifiers of genomic location are not used frequently or consistently enough in abstracts to construct a large and reliable mapping between genomic location and literature.

A range is a delineation of consecutive cytobands, possibly even spanning a centromere. Whenever such a range is encountered in an abstract, all the intermediate cytobands are associated to the abstract as well. A custom ontology resolves all bands in a range: a document mentioning 1p21.2-q23.1 will

be annotated to all bands in between. In addition, an association to a certain abstract is transferred from a certain cytoband upwards through different levels of cytogenetic resolutions. This implies documents mentioning 3q26.32 will be annotated to 3q26 as well.

Based on this premise, we constructed a map that links MEDLINE abstracts to cytogenetic bands. This highly specific map was then used to characterize individual cytogenetic bands based on the content of the abstracts they are linked to. As the contents of the literature indices underlying aBandApart are updated regularly, the validation is based on a version of the tool that was frozen at the state of MEDLINE on September 6th, 2005. Within that MEDLINE corpus, we identified 36.092 abstracts mentioning at least one cytogenetic band or range of bands. From this set, 293.808 associations between bands and concepts were extracted. As nearly 60.000 publications are added to this corpus every month, the number of abstracts and associations is expected to grow steadily as the system is continuously brought up to date.

A potential source of concern for the text-mining algorithm is that man is not the only organism for which banding patterns can be discerned through cytogenetic staining. Band nomenclatures also exist for other organisms. Genome architecture differs among species, which implies that assertions on human genotype-phenotype correlations are contaminated by literature dealing with nonhuman organisms for which a similar band pattern nomenclature is used. To assess the importance of this problem, we need to know the prevalence of documents dealing with nonhuman species in our corpus. We considered the complete set of documents that mention one or more cytogenetic bands and indexed this set using a vocabulary of both common and scientific organism names based on English animal related lists (nouns and adjectives), as well as the NCBI taxonomy (`www.ncbi.nlm.nih.gov/Taxonomy/`). From this vocabulary, 489 distinct terms and phrases were detected at least once in the document set. The most frequently occurring species are shown in Table 3.3.

rank	phrase	rank	phrase
14865	human	126	pig
3664	mouse	107	primates
1252	rat	98	papillomavirus
590	rodent	70	cat
474	hamster	70	bacteria
240	bovine	68	zebrafish
214	melanogaster	67	sheep
183	chicken	63	canine
178	porcine	63	troglodytes
135	rabbit	61	monkey

Table 3.3: The most frequently occurring species in a set of 36.082 cytogenetic MEDLINE abstracts mentioning cytogenetic bands

Note that the results from Table 3.3 do not imply that 14.865 documents discuss human cases and 3.664 documents discuss mouse: on the one hand, the term *human* does not necessarily occur in all abstracts on human. On the other hand, the terms *human* and *mouse* can co-occur, since some abstracts discuss patients as well as model

organisms. Although the mere occurrence of terms and phrases relating to organisms does not clearly elucidate the topic of a document, this brief analysis allows us to estimate how species are distributed as subjects of documents.

A clear majority of all references to organisms in our test corpus is human. The second most frequent organism is mouse and is referenced four times less often in the test documents. However, it does not add noise to the cytogenetic band detection because its band staining patterns are indicated with capital letters followed by a number. The third most frequent organism is rat, as *rat* occurs in 3.47% of the test document set. As the rat chromosome nomenclature closely follows the human cytogenetic nomenclature [136], abstracts dealing with rat band patterns are a potential source of contamination—however, they represent only a small fraction of the abstracts.

The problem is further reduced because of at least two reasons. Firstly, only a fraction of these rat related documents actually contaminate the genome-to-literature map. We manually verified a random sample of 30 documents containing the term *rat*. Only a third contained cytogenetic bands that indeed referred to the rat genome, the other documents all contained bands that referred only to the human genome. This suggests contamination of the genome-to-literature map by nonhuman band patterns is smaller still. Secondly, not all bands stand the risk of contamination. Human bands at high resolution (*e.g.*, 4q15.32) do not occur in rat. In addition, for chromosome 1 (for example) and at the same cytogenetic resolution for rat and human, only 12 of 21 rat bands and only 12 of 24 human bands occur in both nomenclatures.

This brief analysis shows that the contamination effect must be kept in mind, but does not weigh significantly on the results of our method.

Vocabularies

Geneticists, pediatricians or physicians in general, dysmorphologists, molecular cell biologists, and etiologists are all interested in making genotype-phenotype correlations. They have however each a different focus—for example, a different level of emphasis on clinical practice versus molecular biology research. To retrieve knowledge that is interesting to a specific researcher at a given time, we increase the specificity of the text-mining results by limiting its scope through controlled lists of concepts derived from biomedical vocabularies and ontologies.

These lists or sets of linked concepts confine the results of our information extraction method to the current interest of the researcher: different domain-specific vocabularies define from which perspective to annotate the genome. The available options include dysmorphology, anatomy-specific, gene- or protein-centered, gene ontology, and disease-related perspectives on the literature. An overview is shown in Table 3.4.

Words as well as phrases are detected as concepts. In the case of ontologies, no relational information is kept, except for synonymy, which is taken into account when applicable (*e.g.*, with LDDB as a vocabulary, the occurrence of the phrase *small head* will trigger an association to *microcephaly*).

The choice of controlled vocabularies is crucial to the scope and applicability of this method. The vocabulary sources were selected with both a research and diagnostics perspective in mind. For example, options range from a rather general Heritable Disease vocabulary (OMIM) to a specific Dysmorphology concept hierarchy (LDDB). Also, each vocabulary focuses on a different level of biological detail, from small (molecular, biochemical) over intermediate (cellular and tissue level) to large

name	function	example	size
MeSH	Medical Subject Headings	chemicals, medical concepts	16.998
GO.B	Biological Processes	'cell growth', 'signal transduction'	1.120
GO.C	Cellular Components	'proteasome', 'nucleus'	402
GO.M	Molecular Functions	'ATPase activity'	701
GO.E	Gene Ontology	all of the above	2.170
LDDB	London Dysmorphology Database	'microcephaly' or 'small head'	808
OMIM	Genetic Disorders	'attention deficit hyperactivity disorder'	1.716
CBIL	Human Anatomy	'heart muscle'	303
OHDA	Embryo Development	'early stage, fetus'	380
TDMS.s	Systems, Tissues and Sites	'cardiovascular system'	392
TDMS.l	Microscopic Lesions	'disseminated intravascular coagulation'	204

Table 3.4: Different Controlled Vocabularies in aBandApart. A total of 11 vocabularies are present, shown above with an example concept and the number of concepts in each vocabulary

(organs and anatomy).

The sources for these vocabularies were chosen based on how authoritative they are in their respective field. Several of these vocabularies have already proven their value in previous work on gene profiling and prioritization. For example, the GO [9] derived vocabularies boost prioritization performance in Endeavour [5], our web based method for candidate gene prioritization by genomic data fusion. Additionally, GO, MeSH and OMIM vocabularies have proven their merit in TXTGate, a web tool in support of previous work on text based gene and gene group profiling [86]. The Dysmorphology vocabulary is also widely used in its field: firstly, the Oxford Medical Dictionary Dysmorphology and Neurology databases that build on the LDDB taxonomy are a widely used clinical reference. Secondly, LDDB is the elementary dysmorphology taxonomy within DECIPHER, the DatabasE of Chromosomal Imbalance and Phenotype in Humans using EnsEMBL Resources, developed and hosted at the Wellcome Trust Sanger Institute (`decipher.sanger.ac.uk`).

Below, individual aBandApart vocabularies, their size, construction and origin are discussed in detail. The vocabularies themselves can be obtained from the authors upon request.

Firstly, the OMIM and LDDB derived vocabularies provide hereditary disease and dysmorphology phenotype specific views.

Secondly, a number of vocabularies have been constructed to mine literature at different levels of detail. From GO [9], vocabularies at the molecular and cellular level are constructed. TDMS offers tissues, organs and systems. Finally, two anatomical vocabularies are provided, one of which is development specific.

Finally, medical and chemical terms and synonyms based on MeSH make up a general purpose vocabulary.

OMIM This database is a catalog of human genes and genetic disorders that focuses primarily on heritable genetic diseases. Although OMIM does not contain direct information on chromosomal aberrations, it is relevant and useful as a resource of hereditary disease phenotypes. From its downloadable textual information we have extracted a vocabulary of disease related concepts out of which 1.642 entries occur in our cytoband-related subset of the PUBMED corpus.

London Dysmorphology Database Most clinical geneticists are familiar with the Oxford Medical Databases. LDDB contains information on 3.428 dysmorphic syndromes and has a hierarchically structured feature vocabulary which we have manually annotated with synonymous phrases to increase recall in our method. In our band annotated corpus, 796 dysmorphies are annotated through 1.286 synonyms. This dictionary is an authoritative source [162] of information about dysmorphic and neurogenetic syndromes.

Gene Ontology GO [9] provides consistent descriptions of gene and gene product attributes in the form of three structured controlled vocabularies that each provide a specific angle of view (biological processes, cellular components and molecular functions). The GO effort is deliberately term centered to allow for uniform queries across different databases. Our method does incorporate synonymy information from GO. GO is built and maintained with the explicit goal of applications in text-mining and semantic matching in mind. Hence, the Gene Ontology is an ideal source for domain specific views in our method, and makes up four controlled vocabularies: (1) the whole set of GO concepts, for general associations to gene and gene product attributes; (2) cellular components, which may comprise anatomical structures (*e.g.*, rough endoplasmic reticulum or nucleus) or a gene product group (*e.g.*, ribosome, proteasome or a protein dimer) [9], (3) biological processes, defined as series of events accomplished by one or more ordered assemblies of molecular functions, and (4) molecular functions, which describe activities at the molecular level.

TDMS Tissue and Lesions Vocabularies At another level up from the molecular and cellular scale, specific vocabularies are provided that are geared at organs, tissues and systems. Two vocabularies have been extracted from phrase lists used in a Laboratory Data Acquisition System set up at the USA National Institutes of Health. The word lists of their Toxicology Data Management System are subset in a vocabulary with microscopic lesions on the one hand, and a vocabulary with microscopic sites, systems, tissues, and organs on the other hand. This allowed us to complete the set of vocabularies ranging from the very small (molecular functions) over spatially larger concepts (cellular locations) to tissues and organs, which are part of the TDMS vocabularies. At the macroscopic end of this spectrum, CBIL offers anatomical structures.

CBIL Anatomy To focus on structures of larger scale than cellular and tissue levels, an anatomy specific vocabulary was extracted from the Hierarchical Controlled Vocabulary of Anatomy Terms from the Computational Biology and Informatics Laboratory at the University of Pennsylvania. The controlled vocabulary is based on anatomy terms taken from the Mouse Gene Expression Database at the Jackson Laboratory, and was extended to incorporate human anatomy. It was then further revised in a number of areas, such as the haematolymphoid system and the brain.

Ontology of Human Developmental Anatomy The Edinburgh Human Developmental Anatomy [108] lists the tissues present during the first 50 days after conception. This vocabulary is based on detailed anatomy information and standard named tissues for analysis of normal and abnormal human embryos. Space-associated data is included. Bard *et al.* based this anatomical ontology on literature and on a

detailed examination of histological material. It includes all the basic tissues recognizable to an experienced histologist and was designed for describing tissue at a fairly fine resolution (*e.g.*, in gene expression experiments).

Medical Subject Headings MeSH is the National Library of Medicine's controlled vocabulary thesaurus. From it, we constructed a vocabulary that takes into account all phrases up to 6 terms, and maps all narrow and equal synonyms, leaving out broad synonyms. Apart from the 22.997 descriptors and their synonyms, over 150.000 entries and synonyms from the separate Supplementary Concept thesaurus are included as well, adding a general focus of chemical records to the vocabulary. Terms and phrases range from general to specific and constitute a general purpose vocabulary with broad coverage of the biomedical field. We recommend this vocabulary for first exploration of a genomic region, and for when none of the specific vocabularies described above apply.

Statistical over-representation

Cytogenetic bands and concepts can occur together in a single document just by chance. Firstly, consider an abstract where one band is mentioned together with one disease and that this disease is then compared to a second disease. Merely relying on co-citation within single documents would have such an abstract cause a spurious association between the band and the second disease. Secondly, a similar situation occurs when a document discusses several bands and contains multiple, loosely related case reports. This situation implies that we cannot accept a genotype-phenotype association based on the mere co-occurrence of the genomic location identifier (a cytogenetic band) and a concept from one of the vocabularies. Our method reports all co-occurrences together with a p-value indicating how much confidence an association deserves. To quantify this level of over-representation, we assume a hypergeometric distribution as a model.

The hypergeometric distribution is a discrete probability distribution that describes the number of successes in a sequence of C draws from a finite population without replacement. The population has labeled (success) and unlabeled items. The hypergeometric distribution describes $p(O)$, the probability that in a sample of C distinctive objects drawn from the global population, exactly O objects are labeled.

In the context of this work, the question is whether more papers link a cytoband b (*e.g.* '4p16.3') to a concept c (*e.g.* 'microcephaly') than one might expect by chance. If this is the case, the link between the concept and the cytoband can be thought of as being over-represented in the text corpus.

Let A be the total number of abstracts annotated to all known cytobands and concepts. This is the size of the PUBMED sub-corpus in our text indices. Let C be the number of papers containing concept c or its synonyms and B the number of papers associated to b as described in section 3.4.3. We want to qualify the strength of the link between band b and concept c.

By only inspecting abstracts from the corpus that are linked to concept c, in fact, a draw is performed with C abstracts in it where some are and some are not linked to band b. Let O_{bc} be the observed number of papers that are associated to band b *and* mention concept c or one of its synonyms. To know whether the number of b-linked papers in that drawn sample is unusually large, we need to know the probability of drawing O_{bc} papers or more extreme outcomes. This corresponds to calculating the

cumulative probability $p(X \geq o)$ and can be calculated by the cumulative distribution function of a hypergeometric random variable X with parameters as described.

Since the hypergeometric distribution is a discrete probability distribution, the cumulative probability can be calculated easily by adding all corresponding single probability values. This probability constitutes a p-value since it is the probability of seeing something as extreme or more extreme than what was observed.

The p-value is then given by the hypergeometric cumulative distribution function

$$p_{bc} = 1 - H_{\text{cdf}}(O_{bc}|A, B, C) \tag{3.1}$$

$$= 1 - \sum_{i=0}^{O_{bc}-1} \frac{\binom{B}{i}\binom{A-B}{C-i}}{\binom{A}{C}} \tag{3.2}$$

$$= \sum_{i=O_{bc}}^{\min(B,C)} \frac{\binom{B}{i}\binom{A-B}{C-i}}{\binom{A}{C}} \tag{3.3}$$

The p-value p_{bc} is the probability that we observe by chance O_{bc} documents or more that associate band b to concept c. It is the probability of observing O_{bc} or more documents linked to band b when drawing C concept-related documents without replacement from a corpus of A abstracts. Symmetrically, it equals the probability of observing O_{bc} or more documents linked to concept c when drawing B band-related documents.

It is important to note that for small numbers of concepts and documents, the p-values possibly provide a distorted view on the actual relevance of the band-concept association. Even though p-values for small counts still correctly represent the probability of observing this or a higher number of co-citations, it is clear that the p-value should be regarded with caution. The web application will show actual counts with each p-value, to allow the user to assert confidence in the association at hand. For a detailed discussion of this issue, see Section 3.4.5.

Relation to other distributions When the population size is large compared to the sample size, the hypergeometric distribution is approximated reasonably well by a binomial distribution. This approach is computationally less intensive. Both distributions were compared. Although a binomial approach proved justifiable, the hypergeometric distribution was chosen because it did not prove detrimental to performance. All p-values are precalculated at indexing time.

The hypergeometric test based on the hypergeometric distribution is identical to a one-tailed Fisher's exact test. This can be verified by writing down the 2×2 contingency table.

Statistically related tools Our method uses different domain vocabularies as concept sources. Gene Ontology as a whole, together with its three sub-branches, constitute 4 of our vocabularies. A range of existing tools operate on Gene Ontology alone to identify over-represented concepts for (groups of) genes that result from expression array experiments. In these tools, the hypergeometric distribution and binomial approximation are prominent statistical methods. The hypergeometric approach and the equivalent Fisher's exact test constitute a standard approach in the majority of the tools, as discussed by Khatri *et al.* [121]. In this review paper, they further state that

although different distributions are used in different tools, it seems that in most cases the differences between the models are not dramatic.

Gentleman *et al.* describe the use of a hypergeometric distribution in GOStats and GOHyperG [200] to find concepts from Gene Ontology that are over-represented for genes. Our work follows a similar philosophy, though applied to a series of unstructured vocabularies and literature co-citations of bands and concepts instead of genes and GO-terms.

FunSpec (`funspec.med.utoronto.ca`) inputs a list of yeast gene names, and outputs a summary of over-represented concepts. The tool calculates p-values using the hypergeometric distribution.

Other tools use the hyper-geometrical distribution or equivalent Fisher's Exact test for finding over-represented concepts from Gene Ontology, including Fatigo [6], GOBar [133], GoMiner [274], GOToolBox [153], GeneMerge [41], GOTree Machine [275], OntologyTraverser [271], GOCluster [269], and GOHyperGAll in R BioConductor [52].

In BlastSets, Barriot *et al.* use the hypergeometric distribution to calculate the probability of having at least the observed number of elements in common between two sets of sequences for which biological relationships are inferred from different data sources.

Web application

We constructed a web application to illustrate and publicize our method and to make validation efforts reproducible. The tool functions in two directions.

On the one hand, users indicate a cytogenetic band on a genome view. These identifiers can also be entered manually. The tool will characterize this band with statistically over-represented vocabulary concepts found in the literature. Users indicate which controlled vocabulary is to be used, according to their current research interest. For example, when aBandApart is queried with 4p16.3 and a disease vocabulary, the most significant concepts are *achondroplasia*, *Wolf-Hirschhorn syndrome*, *Huntington disease*, *multiple myeloma*, *cherubism*, *dwarfism*, and *hypochondroplasia*, all of which are disorders confirmed to be associated to that region.

On the other hand, users can start from a concept and query the database for statistically over-represented chromosomal regions. If the concept is not found, the application will suggest alternatives with similar spelling. Over-represented bands are listed together with their p-values, and the raw counts that were used to calculate each p-value. The highly over-represented bands are highlighted in red on the same genome chart that is used for input of cytogenetic bands. Links to relevant literature are provided with the cytoband profile.

3.4.4 Results

To illustrate our approach, we discuss results for searches related to heart disease. A detailed validation of our method follows as we discuss the performance on a set of 90 known gene-disease associations. We conclude by evaluating the correspondence of our results to chromosomal aberration maps composed by Brewer *et al.* [34, 35].

Heart disease

We first illustrate the approach by querying the system for *heart* while selecting CBIL, the human anatomy vocabulary. The concept *heart* has a total of 1.324 documents

associated to it. The five most relevant hits are shown in Table 3.5.

band name	BC	B	p-value
22q11	164	1092	0
22q11.2	83	755	1.28e-26
20p12	19	113	3.03e-10
21q22.2	16	171	5.88e-06
7q11.23	20	301	1.12e-04

Table 3.5: Five most relevant hits for query *heart* on vocabulary CBIL. The concept *heart* has a total of 1324 documents associated to it. The four columns show the hit, the number of documents that are linked to both band and concept, the number of documents linked to the band (hit), and the p-value.

A very strong correlation is found for 22q11, and specifically, 22q11.2. Closer examination of these first two hits reveals that this association relates to the well-known DG/VCFS syndrome (DiGeorge / velocardiofacial syndrome). The zero p-value occurs because DG/VCFS, known as the 22q11.2 deletion syndrome, is the most common chromosomal deletion syndrome found in humans [270]. Cardiac defects are strongly penetrant in those patients. The third best association, linking *heart* to 20p12, is corroborated by literature on the Alagille syndrome [128], a pleiotropic disorder with involvement of the liver, heart, skeleton, eyes, and facial structures. The fourth, 21q22.2, is identified through literature analysis as a chromosomal region critical for heart defects related to Down syndrome [127]. The fifth most relevant result is 7q11.23. When 7q11.23 is submitted as a query with the CBIL anatomy vocabulary, a link with the cardiovascular system is apparent. Results with highly significant p-values ($p < 0.01$) are shown in Table 3.6.

As an illustration of how working with different domain vocabularies can be beneficial, we characterized the same 7q11.23 band through different vocabularies. From the perspective of dysmorphology, through vocabulary *LDDB*, the highest ranking concept is *supravalvular aortic stenosis*. Other cardiovascular concepts occur, together with anxiety and mental retardation, suggesting involvement of the central nervous system. The latter is confirmed through use of the disease related vocabulary, OMIM, linking the genomic location to the Williams-Beuren syndrome. To elucidate an underlying molecular function for this anomaly, the same query was submitted with the GO Molecular Function vocabulary. The highest ranking concept, *elastin*, is assigned a near zero p-value. Indeed, the majority of Williams-Beuren syndrome (WBS) patients have been shown to have a micro-deletion within 7q11.2 including the elastin gene, leading to disorganized pre-elastic and mature elastic fibers [194]. Through this brief discussion we have illustrated how different domain vocabularies each provide a specific view towards a genotype-phenotype association.

NIH data set — Genes and Disease

The online NIH book *Genes and Disease* (`www.ncbi.nlm.nih.gov/books/`) discusses a set of genes and the diseases that they are known to cause. With each genetic disorder, the underlying mutations are discussed, along with clinical features and links to key web sites. Over 80 genetic disorders have been summarized in this resource, which we use as positive controls in the validation of our method.

concept	BC	B	p-value
valve	5	51	8.23e-7
connective tissue	6	96	2.64e-6
aorta	5	70	5.43e-6
metencephalon	1	2	3.92e-5
heart	20	1324	1.12e-4
hepatocyte	3	79	1.58e-3
carotid artery	1	10	1.71e-3
pons	1	13	2.92e-3
tonsil	1	14	3.40e-3
artery	3	120	7.06e-3
penis	1	22	8.34e-3
cardiovascular system	1	22	8.34e-3
brain	23	2267	9.16e-3
skeletal muscle	9	664	9.78e-3
midbrain	1	24	9.88e-3

Table 3.6: Highly significant hits (p-value < 0.01) for query 7q11.23 on vocabulary CBIL. The band 7q11.23 has a total of 301 documents associated to it. The four columns show the hit, the number of documents that are linked to both band and concept, the number of documents linked to the concept (hit), and the p-value.

For Chromosome 1, results are shown in Table 3.7. The first two columns show the Gene name and disease as they occur in the NIH book. The disease name is the search term that was used to test our method. In some cases, spelling variants were used. Further columns indicate whether (**H**) the method assigned a highly significant p-value ($p < 0.01$) to the band to which the disease is actually associated, (**S**) whether it assigned a significant p-value ($p < 0.05$), (**P**) whether it delineated the band precisely; i.e., at the maximum level of karyotype resolution (4p16.1 is more precise than 4p16), and (**T**) whether it rated the band as the most significant candidate for this disease, ranking higher or as high as all other bands.

Gene	Disease / Concept	H	S	P	T	NIH	Top	p-value
UROD	porphyria cutanea tarda	1	1	0	1	1p34.1	1p34	0.70E-4
GBA	Gaucher disease	1	1	1	1	1q21	1q21	2.41E-22
GLC1A	glaucoma	1	1	1	0	1q24.3	1q24	2.21E-26
HPC1	prostate cancer	1	1	1	0	1q25.3	8p22	0.00E-0
PS2	Alzheimer disease	0	1	1	0	1q42.13	3q42.1	0.24E-2

Table 3.7: NIH Book Validation for Chromosome 1. On this chromosome, 5 disease genes are annotated. Further columns indicate whether (**H**) the method assigned a highly significant p-value (< 0.01) to the band to which the disease is actually associated, (**S**) whether it assigned a significant p-value (< 0.05), (**P**) whether it delineated the band at the maximum level of karyotype resolution, and (**T**) whether it rated the band as the most significant candidate for this disease, ranking higher or as high as all other bands.

A validation of our method with the disease related genes on other chromosomes is available as supplementary material to the publication that describes this work [253].

An overview of the entire validation is shown in Table 3.8.

Our method assigns a significant p-value ($p < 0.05$) to 84 out of 93 (over 90%) gene linked diseases discussed in the NIH book data set. Of these, 80 (or 86%) are assigned a highly significant p-value ($p < 0.01$). For 57 (or 61%) of these genetic diseases, the cytogenetic band containing the causative gene was reported with the most significant p-value of all reported bands. These results can be verified through the supplementary material or reproduced through the aBandApart web application.

Eight diseases were not significantly linked to the band containing the causative gene. Most of these misses are explained by the fact that the concept is not in any of the domain vocabularies (6 of 9 misses). This occurs with complex or overly detailed concepts (*e.g.*, *gyrate atrophy of the choroid and retina*) or chemical compounds (*e.g.*, *steroid 5-alpha reductase*, *alpha-1-antitrypsin deficiency*). Although the concept *multiple endocrine neoplasia* does not occur in any of the vocabularies, the NIH band for this disease does show a relatively high number of cancer related concepts.

Secondly, misses can also be explained by the fact that there exists no literature in the MEDLINE corpus associating a concept or any of its synonyms to the band in question. This is the case for the CKN1 gene, where no abstracts link the Cockayne syndrome to 5q12, and for the Zellweger syndrome, where no literature links it to 12p13.3.

Finally, although a band is found, it is sometimes not assigned a significant p-value. This is the case for *diabetes*, which our method only weakly links to 7p13. Diabetes has putative causative links to many genomic regions.

Congenital malformations

To further validate our methodology, we evaluate its agreement with a chromosome map of autosomal deletions composed by Brewer *et al.* [34, 35]. In this work, clinical and cytogenetic information from the Human Cytogenetics Database was used to associate different congenital malformations to non-mosaic single contiguous autosomal deletions and duplications. We have assembled a list of 63 malformation-to-band associations that the authors deemed statistically highly significant. Brewer *et al.* classified malformations in 7 categories: cardiac, central nervous system, craniofacial, gastrointestinal, genitourinary, ocular, and skeletal and limb malformations. Table 3.9 shows results for congenital malformation.

Out of 63 malformation associated bands deemed significant by Brewer *et al.*, 44 were assigned a significant p-value by our method (70%), 35 were given a highly significant p-value (56%). Five associations were detected but not given a significant p-value. Of the 14 associations made by Brewer *et al.* that were not detected by our method, one was missed because of different phrasing of *agenesis of corpus callosum* in literature, and 13 were missed because no abstracts were found linking band and malformation.

3.4.5 Discussion

aBandApart links phenotype information to genomic aberrations at the level of cytogenetic bands. We assessed that significant p-values yielded by the method are supported by known cytogenetic aberrations and by published malformations and diseases.

With regard to our text-mining methodology, one point worth noting is that MEDLINE abstracts are used instead of the full text of the corresponding article. Although full text articles are increasingly made available through centralized repositories and open access initiatives, harvesting full text is not possible for all publications because

of technical and legal restrictions. Although the potential difference in information present in full text must be kept in mind (for example, the surplus of sequence-related data reported in full text vs. abstract was proved to be significant in an earlier study [268]), the use of abstracts alone is justified because they summarize the key information from a paper (for example, as keywords [217]).

Regarding the statistical methodology, it is again worth stressing that the hypergeometric approach can yield small p-values for associations that not necessarily deserve to be marked as meaningful. This is the case for very small numbers of concepts and documents. For example, associations of *4p16.3* to both *broad nasal tip* and *microcephaly* are flagged as significant by this method; the first based on 1 co-citation in 2 documents, the latter on 11 co-citations in 322 documents. Even though both resulting p-values correctly represent the probability of observing this or a higher number of co-citations in a statistical sense, it is clear that the p-value in the first situation should be regarded with more caution. One option could be to use a regularized estimator that penalizes more strongly associations involving few documents. We decided against this choice because such association can be meaningful. In the case of association through few documents, individual abstracts must be reviewed to confirm the potential associations and avoid overreliance on the p-value. To allow an informed decision on the actual significance of an association between a band and a concept, the web application also indicates the actual counts that were used to calculate the p-value. This raw count information is crucial to the interpretation of results from the web tool: p-values must always be evaluated in the light of the counts mentioned in the "Links" column directly to the left. The caption of the result table explicitly mentions the meaning of each field.

Our method for associating biomedical concepts to cytogenetic bands provides diagnostics support to clinicians looking to identify chromosomal regions containing genes involved in disease processes, and to determine clinical entities linked to genomic aberrations in patients. It supports genetic counseling and an educated follow-up of clinical cases. It also aids cytogeneticists to generate refined accounts on cytogenetical findings they interpret and report to medical professionals (such as gynecologists, pediatricians, psychiatrists, or genetic counselors) and to the patient's family.

For researchers, the generation of a phenotypic genome map based on text-mining will ease the identification of genes involved in disease processes and could delineate novel clinically recognizable entities. Through our controlled vocabularies, their research can be focused on specific knowledge domains. Additionally, the tool provides non-cytogeneticists an accessible bridge to the cytogenetic literature.

The databases can support curation of chromosomal aberrations catalogues. They do not render case report catalogues obsolete, rather, they aim at complementing these resources by offering a publicly available, free, online and searchable resource that is kept up to date through regular automated updates.

3.5 aGeneApart: mapping genes to human dysmorphology phenotypes by literature mining

The aBandApart approach has proven a useful tool for automatically generating genotype-phenotype maps based on a public corpus of biomedical literature. Cytogenetic bands are, however, rather coarse genomic regions, and can contain many genes.

Hence, a logical next step is to extend the approach by applying it to genes as central loci for genotype-phenotype correlation. Finding genes in biomedical literature, however, is not as straightforward as identifying bands: banding pattern descriptions can be complex, but follow a standard nomenclature (as described in Section 3.4.3). Just like the International System for Human Cytogenetic Nomenclature (ISCN) provides universal terminology for the description of chromosomal anomalies based on cytogenetic staining techniques [216] that can be recognized by a set of regular expressions as is done by the process underlying aBandApart, a number of gene name nomenclatures exist as well, of which probably HUGO is best known: for each known human gene, the HGNC or HUGO Gene Nomenclature Committee [7] approve a gene name and symbol (short-form abbreviation). All approved gene symbols are stored in the HGNC Database.

The HGNC database keeps these primary gene symbols unique and ensures that each gene is only given one approved gene symbol. It is necessary to provide a unique symbol for each gene so that researchers and geneticists can talk about them, but it also facilitates electronic data retrieval from publications. However, genes in literature are not only referred to using their HGNC gene symbol: sometimes, gene names are abbreviated differently, sometimes old or obsolete symbols are used, sometimes synonyms are used – and for some synonyms it is not possible to disambiguate between two different genes the same synonym maps onto. Names for genes and proteins are highly ambiguous and there are usually multiple names for the same gene or protein [137].

In some cases, it is impossible to discern gene symbols from regular English words. Table 3.10 illustrates this; it is clear that in this context it is no longer possible to use basic matching algorithms against a vocabulary of strings representing gene names and synonyms. We have already seen how similar problems need to be addressed in the context of tagging phenotypes and disease names. For gene identifiers, the problem is at least equally hard. Luckily, a host of different methods and algorithms have been developed specifically to address the issue of finding gene and protein identifiers in free text [48, 93, 145, 157, 161, 165, 192, 240]. An example of an interesting abstract mentioning both phenotype information as gene names is shown in Image 3.2.

3.5.1 A pipeline to extract relevant concepts from literature

As explained above, the text processing pipeline behind aGeneApart is more complex than that for aBandApart. In the following paragraphs, the system will be explained in detail, following the schematic overview in Figure 3.3.

Abstract selection In a first step, the full body of PUBMED abstracts licensed from the US National Library of Medicine are unpacked and fed through a filter that splits the archives into individual abstracts. These individual documents are checked to verify that they were published in a selection of journals relevant to the field of human medical genetics. Then, articles published over 20 years ago are removed to reduce noise, unless they were flagged as highly relevant due to being present in curated gene-to-document sets such as the Entrez Gene database.

[7] `www.genenames.org`



Figure 3.2: Example of a PUBMED abstract containing relevant concepts: phenotypes are highlighted in blue, genes in red.

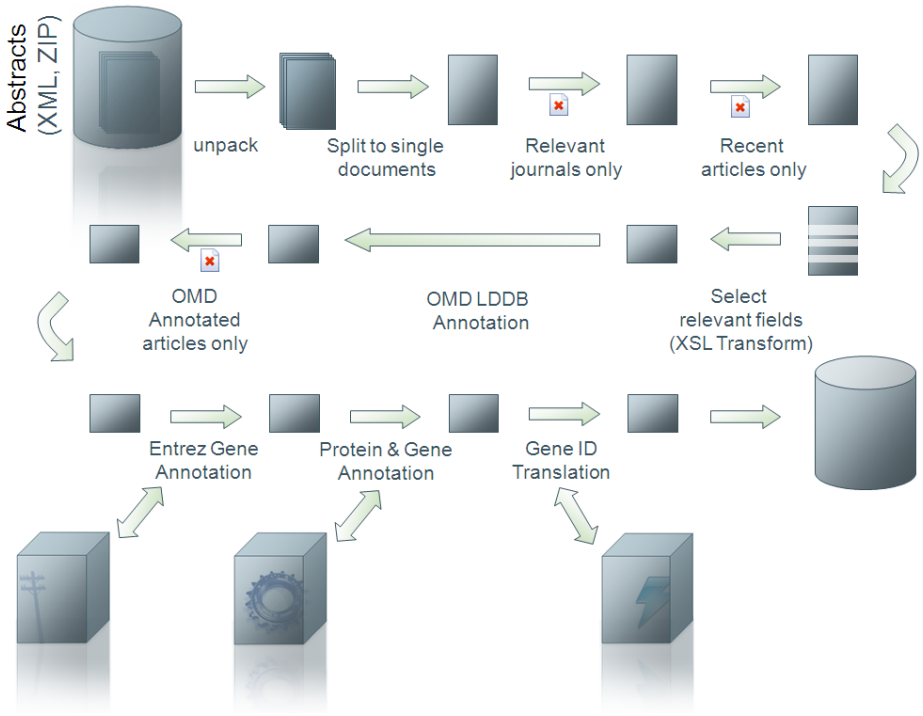


Figure 3.3: Flow Chart of the text processing pipeline behind aGeneApart. The schema shows abstract filtering, information extraction, named entity recognition, identifier translation and finally, composition of the gene-to-phenotype database.

Field selection From the remaining abstracts, relevant fields are selected. An abstract is received in XML format and contains much more than just the abstract text, author list, title, and publishing information. Specifically, an abstract is typically annotated with gene and protein names, chemical compounds, MeSH (Medical Subject Headings) classifiers, and much more. If this information is present, it is withheld and stored for further processing.

Phenotype annotation The selected text from title and abstract, together with any other relevant fields such as MeSH annotations and chemical compounds, are fed through a phenotype annotation pipeline. This pipeline uses a sliding window technique to detect phenotype strings in blocks of text. It is based on a manually annotated list of a human dysmorphology vocabulary, the Oxford Medical Dictionary LNDB database, that has been discussed earlier. It was manually annotated in the sense that alternative spellings (*tumor* vs. *tumour*), synonyms (*small head circumference* vs. *microcephaly*) and alternative ways of writing ('patient *fingers* were *small*' vs. 'patient had *small fingers*') do not stand in the way of accurate phenotype detection, since such phrases are identified by sorting individual terms alphabetically within a sliding window of multiple words. This filter annotates all text with XML tags mentioning the unique LNDB identifiers that correspond to the phenotype that was detected. At the end of this filtering step, all abstracts that do not show any phenotype annotation are removed, as they can not be used to link a gene to a phenotype.

Gene name annotation The remaining abstracts are fed through a gene annotation filter. This filter will first annotate genes to the abstract at hand based on its PMID (PubMed identifier): if the identifier in question occurs in Entrez Gene, a database that links genes and abstracts together based on manual curation, that gene is annotated to the abstract under investigation. The rationale is that Entrez Gene marks this abstract as particularly relevant for this gene, and hence, that the link between the gene and the abstract should in no circumstances be missed (*e.g.* by a failing automated gene annotation). Then, an automated gene annotation kicks in, passing the abstract and all other selected data fields through a web service that was specifically built for the purpose of tagging gene and protein names in free text, at the EBI text-mining Team. This web service, *Whatizit* [188], is offered as a public service through the EBI informatics team to both outside and inside researchers.

Whatizit is a text processing system that allows to perform text-mining tasks defined by pipelines offered by the EBI. *Whatizit* is great at identifying molecular biology terms and linking them to publicly available databases. Identified terms are wrapped with XML tags that carry additional information, such as the primary keys to the databases where all the relevant information is kept. The wrapping XML is translated into HTML hypertext links. This service is highly appreciated by people who are reading or automatically parsing literature and need to quickly find more information about a particular term, such as UniProt identifiers.

Whatizit is available as a Web Service, a streamed servlet, and a web interface. The web service allows to enrich content within an external pipeline such as the one set up in the context of aGeneApart. *Whatizit* is good at identifying gene names and serving UniProt identifiers for them. The EBI text-mining team uses statistical models for disambiguation of terms that can be confused with regular English words bearing different meaning. Interestingly, several vocabularies can be integrated in a single pipeline, as is the case for the pipeline that tags both SwissProt gene or protein

names, and Gene Ontology terms. It is this filter, the *WhatizitSwissprotGo* pipeline, that is used by the aGeneApart annotation pipeline.

Normalization of identifiers In a second to last step, all identifiers, whether it be Entrez Gene id's, SwissProt and UniProt id's, or others, are translated to a single unique gene identifier. The EnsEMBL Gene ID was chosen as a central identifier, because it eases linking out to the EnsEMBL database, and makes the pipeline immediately compatible with other tools developed at both ESAT and the Sanger institute.

Statistical over-representation In a last step, the full body of annotated literature will be processed to generate count tables. Four count tables are needed: (a) the total number of abstracts found, (b) a count table with all genes and the number of times they occur in an abstract (for example, Gene ENSG00000012048, or *BRCA1* occurring 12.348 times), (c) a count table with all phenotypes and the number of times they, or one of their synonyms, occur in an abstract (for example, Dysmorphy 17.02.00, or *breast abnormalities*, occurring 18.796 times), and (d) a link count table, containing a co-occurrence count of every gene co-cited with a phenotype (for example, ENSG00000012048 and 17.02.00 co-occurring 1.583 times). These four bits of information are exactly what we need to apply the very same statistical over-representation calculation as described in 3.4.3, leading to a large table of gene-to-phenotype links and corresponding p-values providing a weight or confidence for that link.

3.5.2 Validation efforts

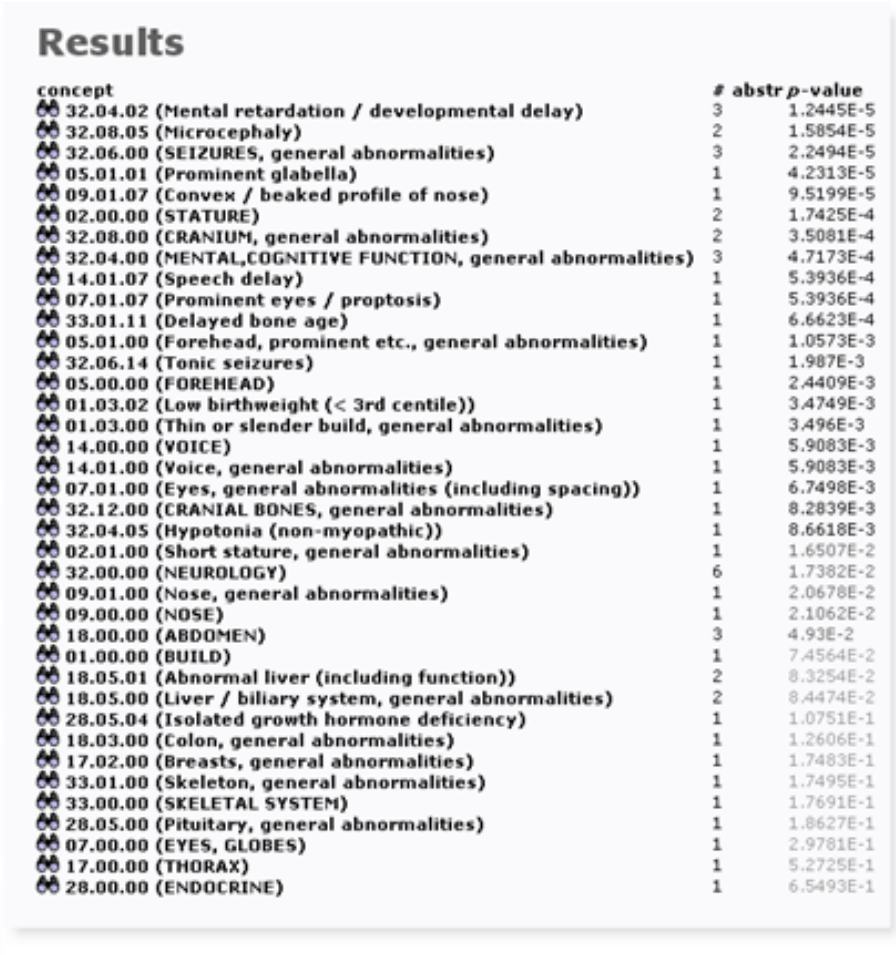
Well known genes

As a first validation effort, well known single gene disorders extracted form the DECIPHER syndrome list have been evaluated. Each time the correct phenotypes are top ranked when querying the system with the gene identifier corresponding to the monogenic disorder; vice versa, for the main phenotypic trait of the disorder, the gene ranks in the top 10 p-value listing. In Tables 3.11 and 3.12, results are shown for two phenotype traits: Supravalvular Aortic Stenosis (LNDB taxon 17.5.4), and Cleft Palate (LNDB taxon 12.4.1). In Image 3.4, the highest ranking phenotype traits are shown for gene WHSC2 (Wolf-Hirschhorn syndrome candidate 2).

aGeneApart as a research tool for other labs

While the results of the aGeneApart pipeline and the corresponding web interface have not been published, the web tool itself has been made available to the research public via word of mouth and conference presentations. At the time of writing, there is one example of a publication where aGeneApart has been used by another research group to perform gene-to-phenotype correlation in the context of a study on the characterization of a 7.6 megabase germline deletion encompassing the NF1 locus [173].

In this patient, a large germline deletion removing the NF1 locus was identified in an 8-year-old French girl with a particularly severe NF1 contiguous gene syndrome, a disorder due to deletion of multiple gene loci that are adjacent to one another. Contiguous gene syndromes are characterized by multiple, apparently unrelated, clinical features caused by deletion of the multiple adjacent genes. Each of the individual genes within a contiguous region, when touched, gives rise to a distinct feature.



Results

concept	# abstr	*p*-value
32.04.02 (Mental retardation / developmental delay)	3	1.2445E-5
32.08.05 (Microcephaly)	2	1.5854E-5
32.06.00 (SEIZURES, general abnormalities)	3	2.2494E-5
05.01.01 (Prominent glabella)	1	4.2313E-5
09.01.07 (Convex / beaked profile of nose)	1	9.5199E-5
02.00.00 (STATURE)	2	1.7425E-4
32.08.00 (CRANIUM, general abnormalities)	2	3.5081E-4
32.04.00 (MENTAL,COGNITIVE FUNCTION, general abnormalities)	3	4.7173E-4
14.01.07 (Speech delay)	1	5.3936E-4
07.01.07 (Prominent eyes / proptosis)	1	5.3936E-4
33.01.11 (Delayed bone age)	1	6.6623E-4
05.01.00 (Forehead, prominent etc., general abnormalities)	1	1.0573E-3
32.06.14 (Tonic seizures)	1	1.987E-3
05.00.00 (FOREHEAD)	1	2.4409E-3
01.03.02 (Low birthweight (< 3rd centile))	1	3.4749E-3
01.03.00 (Thin or slender build, general abnormalities)	1	3.496E-3
14.00.00 (VOICE)	1	5.9083E-3
14.01.00 (Voice, general abnormalities)	1	5.9083E-3
07.01.00 (Eyes, general abnormalities (including spacing))	1	6.7498E-3
32.12.00 (CRANIAL BONES, general abnormalities)	1	8.2839E-3
32.04.05 (Hypotonia (non-myopathic))	1	8.6618E-3
02.01.00 (Short stature, general abnormalities)	1	1.6507E-2
32.00.00 (NEUROLOGY)	6	1.7382E-2
09.01.00 (Nose, general abnormalities)	1	2.0678E-2
09.00.00 (NOSE)	1	2.1062E-2
18.00.00 (ABDOMEN)	3	4.93E-2
01.00.00 (BUILD)	1	7.4564E-2
18.05.01 (Abnormal liver (including function))	2	8.3254E-2
18.05.00 (Liver / biliary system, general abnormalities)	2	8.4474E-2
28.05.04 (Isolated growth hormone deficiency)	1	1.0751E-1
18.03.00 (Colon, general abnormalities)	1	1.2606E-1
17.02.00 (Breasts, general abnormalities)	1	1.7483E-1
33.01.00 (Skeleton, general abnormalities)	1	1.7495E-1
33.00.00 (SKELETAL SYSTEM)	1	1.7691E-1
28.05.00 (Pituitary, general abnormalities)	1	1.8627E-1
07.00.00 (EYES, GLOBES)	1	2.9781E-1
17.00.00 (THORAX)	1	5.2725E-1
28.00.00 (ENDOCRINE)	1	6.5493E-1

Figure 3.4: Output of phenotype query for gene WHSC2 (Wolf-Hirschhorn syndrome candidate 2). The main phenotypes of this syndrome rank highest. This image also illustrates the result output interface: *p*-values are color-coded according to significance levels, individual terms can be clicked to review synonyms.

The deletion breakpoints were pinpointed and the aberrant region was located on chromosome arm 17q and is exactly 7.586.986 bp long. It encompasses the entire NF1 locus and about 100 other genes, including numerous chemokine genes and four micro-RNA genes. Through the use of aGeneApart, Pasmant *et al.* could *in silico*-select a cerebrally expressed candidate gene, designated NUFIP2, for nuclear fragile X mental retardation protein interacting protein 2. This is a gene whose deregulation might be involved in the reported patients mental retardation, being statistically associated with mental retardation in MEDLINE abstracts through aGeneApart with a p value under 10^{-5}. This gene encodes an 82-kD protein (called 82-FIP) distributed in various regions of the brain and interacting with FMR1 (encoded by FMR1 for fragile-X mental retardation 1), the absence of which causes the most common monogenic form of mental retardation, the fragile-X syndrome [17].

3.5.3 Availability through a web interface

The pipeline we have described, results in a large database of links between genes and phenotypes, each one of them annotated with a p-value indicating a link's strength, and with pointers back to abstracts of papers that serve as corroborating evidence. This database is only useful with either a direct interface on top to search and access the genotype-phenotype links, or through integration in other tools.

The web interface is available through `www.esat.kuleuven.be/ageneapart`, and is kept fast and interactive through the use of AJAX technology (Asynchronous JavaScript and XML). It finds correct EnsEMBL identifiers for genes as well as LNDB identifiers for phenotypes as the user types part of their names or definitions in a query box. Once the user has identified the gene or phenotype of interest, he/she can query the database in two directions. When querying the database with an EnsEMBL gene identifier, aGeneApart will list related phenotypes ordered by significance, show a definition, the number of abstracts annotated to the gene that mention the phenotype at hand or one of its synonyms, and a p-value indicating its significance. These synonyms can be viewed with a single click. Vice versa, when querying with a phenotype identifier, the database will list all genes ordered by significance, show the band the gene is in, display the number of abstracts that corroborate the link between the gene and phenotype, and a p-value. Extra information such as full gene name, description, and other identifiers are available through a single click.

3.5.4 Integration in other tools

The web interface described above is fast, simple and straightforward, but not powerful. For example, it does not allow querying with more than one gene or phenotype at the same time. The use of such an extension is clear: when a query with more than one gene combined with a list of more than one phenotype would be possible, this would provide a simple but powerful text based gene prioritization approach. With this in mind, the database underlying the aGeneApart pipeline has been included in two cytogenetic database tools: DECIPHER and CGHGate/Bench.

DECIPHER

DECIPHER, the DatabasE of Chromosomal Imbalance and Phenotype in Humans using EnsEMBL Resources (`www.sanger.ac.uk/PostGenomics/decipher`) has been

inspired by the need to distinguish clinically significant imbalances from transmitted imbalances or polymorphisms detected using microarrays. One of the aims of this project is facilitating research on genetics in human development and health. The database collects information about clinical cases of submicroscopic chromosomal imbalances. Submitted clinical and genetic information is mapped onto the human genome through the EnsEMBL Genome Browser. DECIPHER has already supported the identification of new syndromes, such as a 17q21.3 microdeletion syndrome associated with developmental delay and learning disability [223] and a 14q11.2 microdeletion syndrome associated with mental retardation and other minor anomalies [273]. The addition of Gene Prioritization within the tool, through incorporating the aGeneApart pipeline, was done during a three month visit at the Sanger Institute. A manuscript on DECIPHER which discusses in part the incorporation of this pipeline has been published [65]. A thorough discussion of DECIPHER will follow in the next chapter.

CGHGate/Bench

A second cytogenetic database that incorporates the results of the aGeneApart pipeline for text based gene prioritization, is BENCH. Formerly called CGHGate, it is a software and database tool set that has been written as a research tool initially, and has grown into a robust platform ready for use in clinical laboratories with a high diagnostic throughput. In this context, the platform is subject of an intellectual property transfer agreement into Cartagenia, a spin-off company of the K.U.Leuven. It will be discussed in great detail in the next chapter of this dissertation.

3.6 Gene prioritization through genomic data fusion

While biomedical text is a useful resource of knowledge for both supervised and unsupervised analysis of candidate gene lists in the context of human dysmorphy phenotype traits, it is not the only possible data source. Within the ESAT bio-informatics research group, it has been a long standing effort to investigate ways to use different data sources and combine them statistically through an approach called 'data fusion'. Endeavour, originally implemented as a thin client Java application [5] and now available through a web interface (`www.esat.kuleuven.be/endeavourweb`), is an online resource for the prioritization of candidate genes.

Using a training set of genes known to be involved in a biological process of interest, its approach consists of (i) inferring several models (based on various genomic data sources), (ii) applying each model to the candidate genes to rank those candidates against the profile of the known genes and (iii) merging the several rankings into a global ranking of the candidate genes. The training, ranking and fusion steps are shown in Figure 3.5, the simplified web interface process is depicted in Figure 3.6. In a recent update article, the latest developments of Endeavour are described [245]. This manuscript describes the web-based user interface, announces the support for multiple species, and introduces the use of additional data sources, including numerous databases: ontologies and annotations, protein – protein interactions, cis-regulatory information, gene expression data sets, sequence information and text-mining data.

The paper describes the performance of the current incarnation of Endeavour on 32 recent disease gene associations from the literature, and describes a number of

recent independent studies that made use of Endeavour to prioritize candidate genes for obesity and Type II diabetes, cleft lip and cleft palate, and pulmonary fibrosis. For a detailed review of this system, we refer to the manuscripts [5, 245]. In the context of this dissertation, we will only focus on part of this system, where biomedical text is used as a data source for gene prioritization. While the use of text data has undergone a series of changes and optimizations during the life and development of Endeavour, initially, the data sources for the text models of Endeavour were directly incorporated from TXTGate, a text-mining framework that will be discussed in Chapter 5. As TXTGate evolved, so did the underlying text models, who in turn are used by the Endeavour platform. The optimizations of the text models, including a comparison of structured vocabularies, different text representations and ranking algorithms for gene prioritization by text-mining have been discussed in a co-authored manuscript by Yu *et al* [272], and results will not be repeated here.

3.6.1 Data fusion for candidate gene prioritization

With the recent improvements in high-throughput technologies, many organisms have seen their genomes sequenced and, more importantly, annotated. This process leads to the generation of a large amount of genomic data and the creation and maintenance of corresponding databases. However, converting genomic data into biological knowledge to identify genes involved in a particular process or disease remains a major challenge. Nevertheless, there is much evidence to suggest that functionally related genes often cause similar phenotypes [228, 87, 115, 169]. To identify which genes are responsible for which phenotype, association studies and linkage analyses are often used, resulting in large lists of candidate genes. In many cases, the list of candidates can be narrowed down to a few dozen. However, it is generally too expensive and time-consuming to perform experimental validation for all these candidates. Therefore, these candidates may be prioritized to first validate the best ones. Given the amount of genomic data publicly available, it is often prohibitive to perform the prioritization manually, and consequently, there is a need for computational approaches.

During the past 5 years, the bio-informatics community has developed several strategies to address this question, and several tools are available online [278, 170]. To our knowledge, all the tools use the concept of similarity. It is based on the assumption that similar phenotypes are caused by genes with similar or related functions [228, 87, 115, 169]. However, the tools differ by the strategy they adopt in calculating the similarity (either between the candidate genes and the phenotypes or between the candidate genes and the training genes) and by the data sources they use. The most commonly used data sources are text-mining data, gene expression data and sequence information. Additionally, phenotypic data, protein–protein interactions, ontologies and cis-regulatory information are sometimes included. However, most of the existing approaches mainly focus on the combination of few data sources. For instance, the method of combining gene expression and protein interaction data proposed by Ma *et al.* [147] combines expression and interaction data. Several methods only rely on literature and ontologies: BITOLA [103], POCUS [246], Gentrepid [78], G2D [176], and the method defined by Tiffin *et al.* [243]. In contrast, systems that use more data sources have recently been designed, such as CAESAR [77], GeneSeeker [250], SUSPECTS [2], TOM [198], and Endeavour [5, 245]. For a more detailed description of the available tools, see Section 3.2.3, and the reviews by Oti and Brunner [169] or by Zhu and Zhao [278].

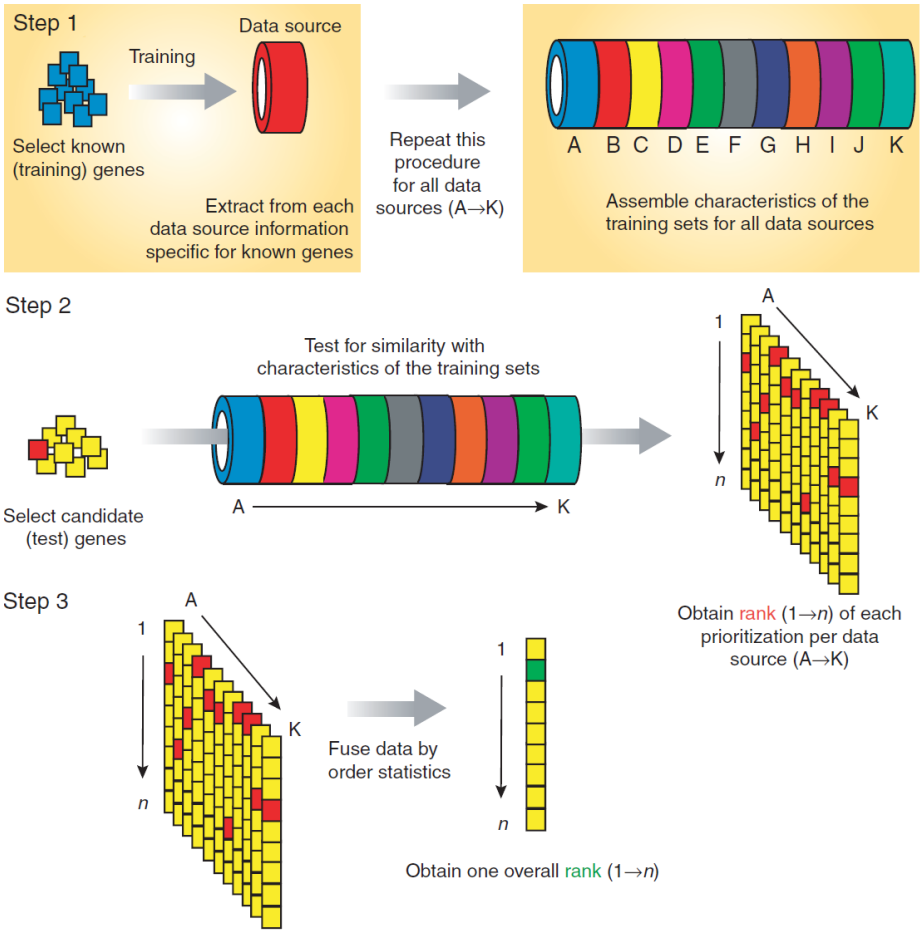


Figure 3.5: Concept of prioritization by Endeavour. Candidate test genes are ranked based on their similarity with a set of known training genes in a three-step analysis. In the first step (upper panel), information about a disease or pathway is gathered from a set of known training genes by consulting various data sources: A, literature (abstracts in Entrez Gene); B, functional annotation (Gene Ontology); C, microarray expression (Atlas gene expression); D, EST expression (EST data from EnsEMBL); E, protein domains (InterPro); F, protein-protein interactions (BIND); G, pathway membership (KEGG); H, cis-regulatory modules (TOUCAN); I, transcriptional motifs (TRANSFAC); J, sequence similarity (BLAST); K, additional data sources. In the second step (middle panel), a set of test genes is loaded. These are then ranked based on their similarity with the training properties obtained in the first step, which results in one prioritized list for each data source. Finally, in the third step (lower panel), Endeavour fuses each of these rankings from the separate data sources into a single ranking and provides an overall prioritization for each test gene.

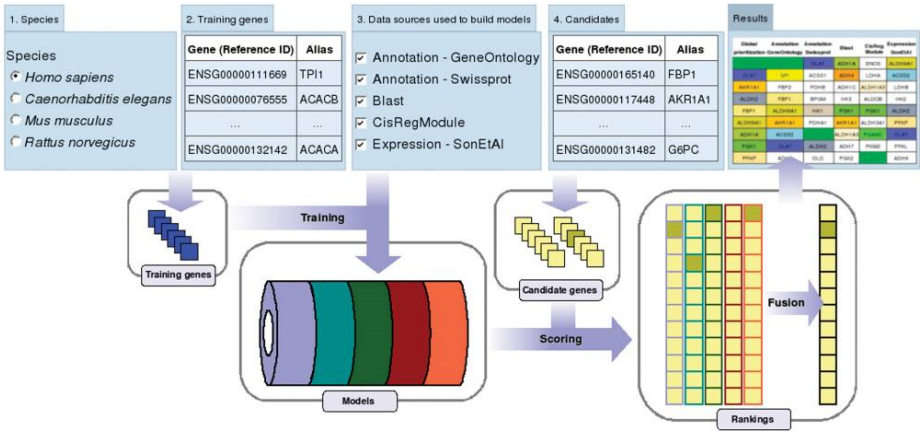


Figure 3.6: The algorithm behind Endeavour web. Once the organism of interest is chosen (Step 1), the user can specify the training genes (Step 2). Step 3 lets the user select the data sources that will be used to build the models. The models summarize the training gene information. The candidate genes specified by the user in Step 4 are then scored against the model. This produces one ranking per data source plus one global ranking obtained by fusion of the rankings per data source. The global ranking together with the rankings per data source are returned to the application and can be viewed in the 'Results' panel.

3.6.2 Principle of Endeavour

Endeavour requires two inputs: the training genes, already known to be involved in the process under study, and the candidate genes to prioritize. Endeavour produces one output: the prioritized list of candidate genes, along with the rankings per data source. The algorithm is made up of three stages, called the training, scoring and fusion stages. In the training stage, Endeavour uses the training genes provided by the user to infer several models, one per data source. For example, with ontology-based data sources, genes are annotated with several terms and reciprocally one term can be associated to several genes. The algorithm selects only the significant terms, the ones that are over-represented in the training sets compared to the complete genome. Hence, the model consists of these significant terms together with their corresponding p-values that reflect the significance of the enrichment. In the scoring stage, the model is used to score the candidate genes and rank them according to their score. For ontologies, the algorithm scores each candidate independently by combining the p-values of its associated terms that are, at the same time, present in the model. The scores are then used to rank the candidates based on this one data source. In the final stage, the rankings per data source are fused into one global ranking using order statistics. Among the existing methods, the order statistics has the advantage of avoiding penalizing genes that are absent from a given data source. Indeed, the genomic data sources are almost always incomplete. For instance, some genes do not have any ontology annotations, while other genes do not have their corresponding probes spotted on the microarray platform for which data is available. The order statistics allow combination of the rankings per data source, taking missing values

into account. Thus, the use of unbiased data sources (*e.g.* gene expression data, cis-regulatory motifs and protein sequences), together with the use of the order statistics, allows us to obtain results that are not overly biased towards the most studied genes. The use of several data sources is indeed an important strength of our approach: combining two data sources, although possibly incomplete, can be more powerful than either individual data source, as shown by published validation experiments [5]. The fact that the Endeavour approach does not rely only on a single data source also reinforces its robustness to noisy data sources like microarray data.

3.6.3 A text model for Endeavour

The data source based on literature mining relies on the TXTGate framework [86], discussed in great detail in Chapter 5. The strategy is to screen the abstracts from PubMed [266] with a manually curated vocabulary based on Gene Ontology. Similarly to the ontologies described above, it provides more information on the molecular functions and biological processes of the genes.

we refer to the full manuscript [5] for a detailed discussion of Endeavour, and validation efforts of the underlying data fusion approach on both disease data sets and biological pathway sets, mono- and polygenic diseases, and regulatory genes of myeloid differentiation. In this section, we will limit platform and validation discussions to the influence of text (biomedical literature) as a data source.

Validation of the text model for Endeavour

For the full manuscript, a validation of Endeavour by individual data source was performed: for each individual data source, it was assessed whether the approach is capable of prioritizing genes known to be involved in specific diseases or receptor signaling pathways, through a large-scale leave-one-out cross-validation. In each validation run, one gene, termed the 'defector' gene, was deleted from a set of training genes and added to 99 randomly selected test genes. The software then determined the ranking of this defector gene for every data source separately. There were 627 training genes used, ordered in 29 training sets of particular diseases automatically selected from the Online Mendelian Inheritance In Man (OMIM) database. For pathway genes, three sets of training genes were compiled, involved in the WNT (43 genes), NOTCH (18 genes) and epidermal growth factor (15 genes) pathways. As a negative control for training genes, 10 sets of 20 randomly selected genes were assembled. This lead to a total of 903 prioritizations performed for each data source. From these, sensitivity and specificity values were calculated. Sensitivity refers to the frequency (% of all prioritizations) of defector genes that are ranked above a particular threshold position. Specificity refers to the percentage of genes ranked below this threshold. For instance, a sensitivity/specificity value of 70/90 would indicate that the correct disease gene was ranked among the best-scoring 10% of genes in 70% of the prioritizations.

The results of these validations allow comparison between data sources and are visualized as rank receiver operating characteristic (ROC) curves, from which sensitivity/specificity values can be easily deduced. The area under this curve (AUC) is a standard measure of the performance of this algorithm. For instance, an AUC-value of 100% indicates that every defector gene ranked first, whereas a value of 50% means that the defector genes ranked randomly.

For every single data source, Endeavour reached a higher AUC score for disease and pathway genes than for randomly selected genes, indicating that it was sensitive

and specific in ranking the defector gene, regardless of the type of data source consulted (Figure 3.7). Not surprisingly, the data sources differed in their usefulness and suitability to rank genes. From the AUC values plotted in Figure 3.7, it is clear that Literature as a data source performs well for both the disease validation (best performance of all data sources) and pathway validation (fourth place of all data sources). Interestingly, these results show that for the disease validation, Literature as a data source outperforms the global ranking based on fusion of all data sources. However, for a number of validations, use of the literature as the only data source was not very efficient, still, inclusion of all the other data sources yielded a high rank, clearly pointing out the benefit of the data fusion approach underlying Endeavour .

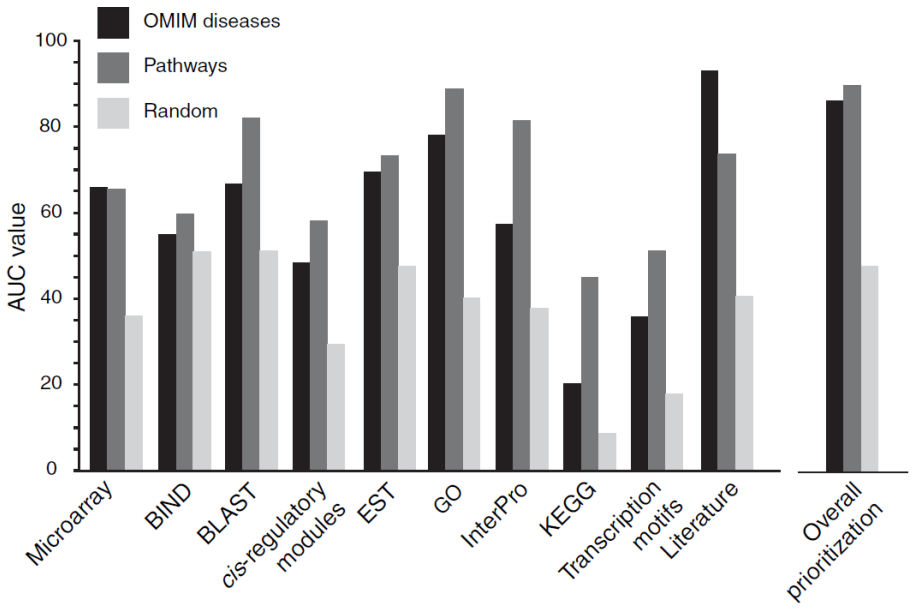


Figure 3.7: Cross-validation results. The AUC values obtained for all individual data sources are shown for disease prioritizations (black), pathway prioritizations (dark gray) and random prioritizations (light gray). The AUC values from the overall prioritization obtained after fusing all individual prioritizations are also shown.

For details on this validation and for a series of specific caveats regarding data sources leading to incompletely annotated genes, bias towards better-studied genes versus unknown genes, influence of the number of data sources used, bias on validation towards information already known through literature, we refer to the published manuscript [5] on Endeavour.

Cross-validation results

To finalize the discussion on the position of Literature as a data source within the Endeavour framework, we review the actual ROC curves showing cross-validation results for both the disease and the pathway validation, shown in Figure 3.8. On top

of these cross-validation results, a control ROC curve is shown in red. This control was obtained after prioritization with randomly constructed training sets and by using all data sources. Hence, this red line serves as a reference, corresponding to an uninformed random prioritization. For all other ROC curves, disease or pathway-specific training sets were generated. The data sources used to construct every ROC curve are indicated on the figure. The main point here is a comparison between the dotted lines and the solid lines. In both panels, the dotted line depicts the use of all data sources except literature, and the solid line of the same color indicates prioritization including literature. These ROC curves illustrate the contribution and importance of literature as a data source.

Note that Endeavour does not rely on literature-derived data alone. For each OMIM gene used in the disease validation, a mutation causing the disease had previously been reported in a landmark study. Because the inclusion of these publications may artificially increase the relative contribution of the literature data source in the overall performance of this algorithm, a separate validation for disease training sets was run with exclusion of the entire literature database, and for the same reason, the GO, KEGG and literature data sources were excluded from the pathway validation. Even under these conditions, the overall performance of the algorithm was only negligibly affected: the performance dropped by 6.1% for disease genes (from 86.6% to 80.5%, as seen in Figure 3.8, left panel) and by only 2.3% for pathway genes (from 89.9% to 87.6%, as seen in Figure 3.8, right panel). Thus, the diversity of data sources used by Endeavour enables meaningful prioritizations, even without the use of literature information.

3.7 Setting the scene for intelligent databases

Array CGH is increasingly being used to define candidate regions for putative genes responsible for human genetic diseases. The increase in gene identification through molecular karyotyping will be driven by building, operating, extending, and disclosing genotype-phenotype databases, by integration of these databases and by making them interoperable, searchable, and amenable to large scale data mining initiatives. Ontologies and standardization of data can support these efforts.

Currently, there is a gap between existing candidate gene prioritization tools and existing case report and genotype-phenotype correlation databases. It can be expected that future prioritization tools will increasingly make use of publicly available case report repositories, and that database efforts in turn will move towards offering tools for intelligent search, clustering, candidate gene prioritization and data mining in general. In this context, there is a clear case for collaborative, interactive tools that support both research and diagnostics. A high diagnostic throughput of experimental and phenotype data can populate large patient repositories that will serve as a valuable resource for research on copy number variability of the human genome.

Standardization of ontologies, conventions on storage and annotation of raw experiment data to make them available to the community in a useful way (such as the MGED (`www.mged.org`) initiative MIAME (Minimum information about a microarray experiment) [32, 14]) and the use of novel data mining algorithms for data integration will improve the automated gene annotation processes of chromosomal aberrations and the delineation of novel and complex clinical entities. The tools and databases being developed to identify the functional relevance of genes will not only deepen our understanding of human development but will, in addition, aid in the data interpretation

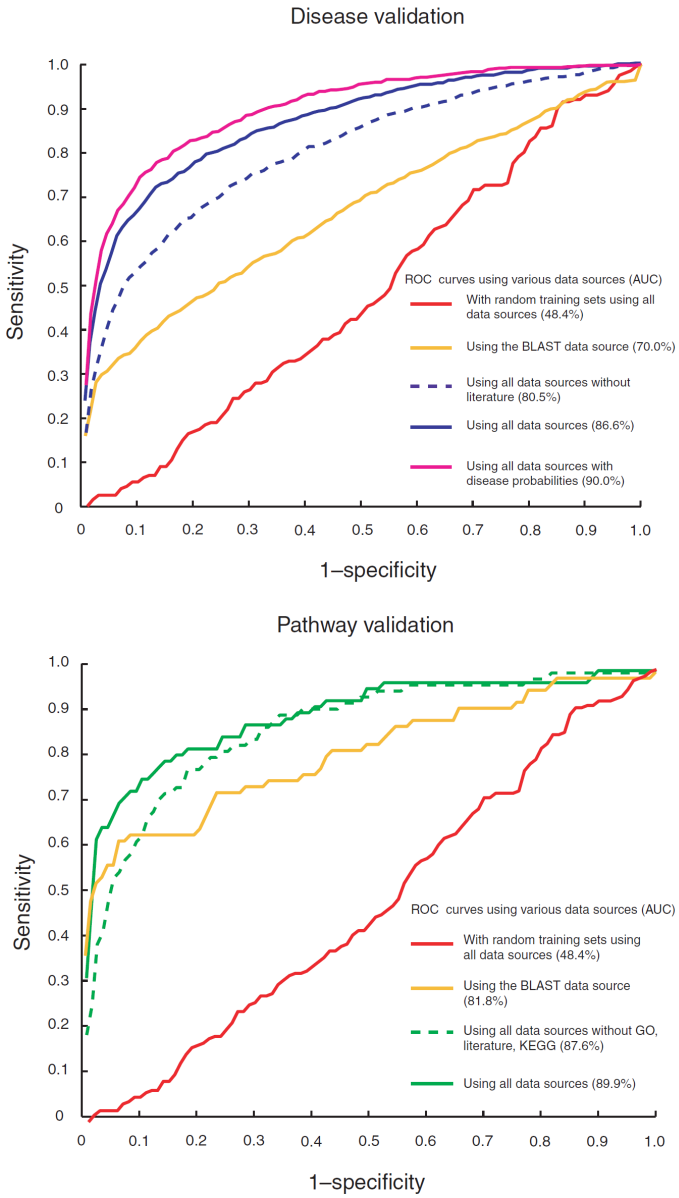


Figure 3.8: Cross-validation results. **Top:** Rank ROC curves obtained for the disease validation. **Bottom:** Rank ROC curves obtained for the pathway validation. In both figures, the control ROC curve (red line) was obtained after prioritization with randomly constructed training sets and by using all data sources. For all other ROC curves, disease or pathway-specific training sets were generated. The data sources used to construct every ROC curve are indicated on the figure.

and improve genetic counseling.

With these remarks in mind, we can move on to Chapter 4, which will describe intelligent databases that take the leap from stand-alone tools that each address only part of the analysis and mining tasks involved with current human genetics research and diagnostics, to human genomics platforms that will address all issues at hand in an encompassing framework that is ready for current and future *in silico* genetic analysis needs.

Chr.	Gene names	Genes	H	S	P	T
1	UROD GBA GLC1A HPC1 PS2	5	4	5	4	5
2	ETM2 MSH2 MSH6 PAX3	4	4	4	4	2
3	VHL MLH1 SCLC1 ETM1	4	4	4	4	2
4	EVC HD FGFR3 NRCLP SNCA FOP	6	6	6	6	4
5	SRD5A1 CKN1 SMN1 Asthma DTD	5	3	3	3	2
6	SCA1 HFE CYP21A IDDM1 EPM2A	5	5	5	4	2
7	GCK ELN CFTR Pendrin OB	5	4	4	4	3
8	WRN MYC	2	2	2	1	2
9	CDKN2 FRDA ABC1 TSC1 ABL	5	5	5	4	5
10	PAHX OAT	2	0	0	0	0
11	HRAS IDDM2 LQT VMD2 MEN1 ATM	6	4	5	5	2
12	PXR1 PAH	2	1	1	1	1
13	BRCA2 CX26 RB1 ATP7B	4	4	4	3	3
14	PS1 SERPINA1	2	1	1	1	1
15	SNRPN UBE3A FBN1 HEXA	4	4	4	4	4
16	HBA1, HBA2 PKD1 FMF Crohn's Disease	4	4	4	4	4
17	P53 CMT1A BRCA1	3	2	2	2	2
18	NPC1 DPC4 / SMAD4	2	2	2	2	1
19	Jak3 BCKDHA APOE DMPK	4	3	4	4	2
20	ADA	1	0	1	1	0
21	SOD1 APS1	2	2	2	2	2
22	SGLT1 DGS NF2 BCR	4	4	4	4	3
X	PIG-A DMD ATP7A COL4A5 IL2RG TNFSF5 HPRT1 FMR1 ALD MECP2 HEMA	11	11	11	11	8
Y	SRY	1	1	1	1	0

Table 3.8: NIH Validation summary. Columns indicate whether (**H**) the method assigned a highly significant p-value (< 0.01) to the band to which the disease is actually associated, (**S**) whether it assigned a significant p-value (< 0.05), (**P**) whether it delineated the band precisely, at the maximum level of karyotype resolution (4p16.1 is more precise than 4p16), and (**T**) whether it rated the band as the most significant candidate for this disease, ranking higher or as high as all other bands.

malformation	band	type	p-value < 0.01	p-value < 0.05
aortic stenosis	11q23-24	del		
hypoplastic left heart	11q23-25	del	✓	✓
hypoplastic left heart	16q11-12	dup		
patent ductus arteriosus	16q22	dup	✓	✓
pulmonary stenosis	20p13-11	del	✓	✓
pulmonary stenosis	22q11	del	✓	✓
pulmonary stenosis	8q22-24	dup		
tetralogy of fallot	8q22-24	dup	✓	✓
truncus arteriosus	22q11	del	✓	✓
truncus arteriosus	2q22	del		
ventricular septal defect	22q11	del	✓	✓
ventricular septal defect	4q31	del		✓
ventricular septal defect	8q24	dup		✓

Table 3.9: Congenital malformation validation. All 13 cardiac anomalies discussed by Brewer *et al.* are shown. Check marks indicate the significance with which our method associated band and concept.

gene/protein	UniProt ID	Frequency in BNC
Had	Q9WVK7	402433
who	O01367	194312
how	O01367	94873
Last	Q9D032	73553
put	Q51841	59248
Yes	Q04736	58451
Set	Q63945	48232
Great	Q91ZZ5	45250
car	Q9Y1I2	34155

Table 3.10: Gene and protein names that correspond to frequently used English words. Column 1 shows the gene name, column 2 the identifier for this gene in the UniProt database, and column 3 shows the frequency of the word in the British National Corpus, a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written.(`www.natcorp.ox.ac.uk`). `Source: EBI Rebholz text-mining Team.`

score	p-value	EnsEMBL Gene ID	Validated	Gene name	Gene description
129	$1.464757 \cdot 10^{-129}$	ENSG00000049540	+	ELN	elastin (supravalvular aortic...
12	$3.120289 \cdot 10^{-12}$	ENSG00000071462	+	WBSCR22	Williams Beuren syndrome ...
12	$1.810960 \cdot 10^{-12}$	ENSG00000077809	+	GTF2I	general transcription factor II, i
11	$6.238573 \cdot 10^{-11}$	ENSG00000106638	+	TBL2	transducin (beta)-like 2
11	$4.967726 \cdot 10^{-11}$	ENSG00000006704	+	GTF2IRD1	GTF2I repeat domain containing 1
10	$9.880883 \cdot 10^{-10}$	ENSG00000168542	+	COL3A1	collagen, type III, alpha 1 ...
09	$5.514204 \cdot 10^{-09}$	ENSG00000009950	+	MLXIPL	MLX interacting protein-like
08	$2.764015 \cdot 10^{-08}$	ENSG00000077800	+	FKBP6	FK506 binding protein 6, 36kDa
08	$1.398369 \cdot 10^{-08}$	ENSG00000107438	+	PDLIM1	PDZ and LIM domain 1 (elfin)
07	$4.414078 \cdot 10^{-07}$	ENSG00000165119	+	HNRPK	heterogeneous nuclear ...

Table 3.11: The most strongly linked genes for phenotype trait Supravalvular Aortic Stenosis (LNDB taxon 17.5.4). p-values are shown in the second column; all genes were confirmed to be linked to this phenotype by literature and database searches, implying that in the top ten ranking genes, no false positives come up.

score	p-value	EnsEMBL Gene ID	Validated	Gene name	Gene description
19	$1.926391 \cdot 10^{-19}$	ENSG00000184058	+	TBX1	T-box 1
18	$4.051166 \cdot 10^{-18}$	ENSG00000102271	+	KLHL4	kelch-like 4 (Drosophila)
15	$9.455555 \cdot 10^{-15}$	ENSG00000119042	+	SATB2	SATB family member 2
15	$3.299856 \cdot 10^{-15}$	ENSG00000101871	+	MID1	midline 1 (Opitz/BBB syndrome)
14	$1.279493 \cdot 10^{-14}$	ENSG00000069399	-	BCL3	B-cell CLL/lymphoma 3
13	$1.520270 \cdot 10^{-13}$	ENSG00000204248	+	COL11A2	collagen, type XI, alpha 2
12	$1.585302 \cdot 10^{-12}$	ENSG00000099889	+	ARVCF	armadillo repeat gene in VCF
10	$5.880943 \cdot 10^{-10}$	ENSG00000060718	+	COL11A1	collagen, type XI, alpha 1
10	$2.295810 \cdot 10^{-10}$	ENSG00000172893	+	DHCR7	7-dehydrocholesterol reductase
10	$1.824248 \cdot 10^{-10}$	ENSG00000163513	+	TGFBR2	transforming growth factor, ...

Table 3.12: The most strongly linked genes for phenotype trait Cleft Palate (LNDB taxon 12.4.1). p-values are shown in the second column; all genes were confirmed to be linked to this phenotype by literature and database searches, apart from BCL3.

Chapter 4

Intelligent databases for constitutional cytogenetics

"Did you really name your son `Robert'); DROP TABLE Students;--`*?"*
"Oh, yes. And I hope you've learned to sanitize your database inputs."

xkcd (2009)

4.1 Biomedical interpretation, collaborative research and diagnostic support

Genome-wide array comparative genomic hybridization screening is uncovering pathogenic submicroscopic chromosomal imbalances in patients with developmental disorders. In those patients, imbalances appear to be scattered across the whole genome, and most patients carry different chromosomal anomalies. Screening patients with developmental disorders can be considered a forward functional genome screen. Imbalances pinpoint the location of genes that are involved in human development. Because most imbalances encompass regions harboring multiple genes, the challenge is to (1) identify those genes responsible for the specific phenotype and (2) disentangle the role of the different genes located in an imbalanced region.

It is clear that for the aetiology of complex genetic disorders to be elucidated, much more data will be needed than for, for example, regular full penetrance single gene disorders. This data will include knowledge from external databases, such as copy number variation and haplotype data, genotype to phenotype correlation maps, information extracted from the body of available biomedical literature, results from large scale case control studies in normal individuals, as well as large volumes of historic patient data – not only from within the research lab, but resulting from large scale or even global patient registry efforts.

A powerful idea in the hunt for large scale patient repositories is ensuring that the steady flow of results from a diagnostic routine pipeline are captured, annotated and aggregated in a registry amenable to search, to supervised or unsupervised data mining, to reporting, and to strong interfaces and visualization possibilities.

In this chapter, relevant novel cytogenetics databases are discussed that have recently been developed to aid the gene discovery process and the diagnostics and research work flow in general. Research on all aspects of the field of complex genetics will not only deepen our understanding of human development but will, additionally, aid in the data interpretation and improve genetic counseling.

In the following sections, a number of public databases and case repositories are discussed. For the remainder of the chapter, two such databases are singled out and discussed in greater detail. One is DECIPHER, an international consortium effort to aggregate case reports over a consortium of hospitals and research institutes. DECIPHER features text based gene prioritization in context of annotated patient phenotypes. This feature has been developed during an internship at the Sanger institute in the context of this PhD work. A second database for cytogenetics is Store+Bench, formerly called CGHGate, which is a similar effort that goes a lot further in supporting the diagnostic work flow itself, and not just the research aspects, through adding specific visualization and data mining functions, integration with existing biomedical knowledge, and a strong focus on automated reporting.

4.2 Public databases

A group of phenotypically related cases can be used to delineate a minimal genomic region that segregates with a clearly defined common part of the phenotype. Through such correlation of components of a phenotype with the loci or genes within the affected chromosomal region, novel clinical entities can be defined. For this to be possible on a large scale, tools and databases are needed. Databases need to be extensive and publicly accessible, and computational approaches need to be compatible with these databases. Both are necessary tools in large-scale studies for association of phenotypic information with genomic data.

Collaborative databases of case reports have been set up in support of discovering new clinical entities such as deletion and duplication syndromes and correlating aberrant genotypes with phenotypes. Both global and local case repositories exist; some initiatives are closed or consortium based, and others are public. Although these initiatives differ in approach and setup, they share the common goal of supporting association studies and efforts in delineating novel syndromes by aggregating patient case reports and, in most cases, encouraging data exchange. DECIPHER and ECARUCA are considered the two most important databases for constitutional cytogenetics. An overview of chromosomal aberration databases is given in Table 4.1. Usually, the data mining facilities of these databases are limited to search and retrieval. Features such as clustering and gene prioritization are planned in future releases of at least some of these tools.

DECIPHER (DatabasE of Chromosomal Imbalance and Phenotype in Humans using ENSEMBL Resources [1]) has been inspired by the need to distinguish clinically significant imbalances from transmitted imbalances or polymorphisms detected using microarrays. One of the aims of this project is facilitating research on genetics in human development and health. The database collects information about clinical cases of submicroscopic chromosomal imbalances. Submitted clinical and genetic information

[1] `www.sanger.ac.uk/PostGenomics/decipher`

is mapped onto the human genome through the Ensembl Genome Browser. DECIPHER has already supported the identification of new syndromes, such as a 17q21.3 microdeletion syndrome associated with developmental delay and learning disability [223] and a 14q11.2 microdeletion syndrome associated with mental retardation and other minor anomalies [273].

ECARUCA (European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations footnote`www.ecaruca.net`) is a European database that covers both common and rare chromosome aberrations. It contains details of thousands of published cytogenetic imbalances and is prospectively gathering rare cytogenetic and molecular cytogenetic aberrations, bringing together cytogenetic, molecular, and clinical data [61].

4.3 Store and Bench

A number of challenges that surpass the offerings of most tools described in Chapter 3, are addressed by encompassing platforms that combine content-aware databases with tool sets aimed specifically at the requirements of a state of the art routine diagnostic work flow in genetic testing laboratories, both with an interest as well as a strong mission in genetics research. We begin this section by identifying these challenges. Next, we discuss two such platforms, Store+Bench and DECIPHER, and show how they each address the issues raised below.

4.3.1 Challenges for genomics IT platforms

Challenge 1 - clinical interpretation

Some genes in the human genome have known or predicted functions in human development and disease, but the function of the vast majority remains unknown. Genomic microarray analysis has the potential to identify small aberrations containing only a few genes. Some aberrations are novel or extremely rare making clinical interpretation difficult. Defining the gene content and comparing clinical features from patients with overlapping rearrangements offers an opportunity to define the phenotype resulting from altered dosage of these genes: often, patients suffering from developmental disorders harbor sub-microscopic deletions or duplications which by affecting the copy number of dosage-sensitive genes or disrupting normal gene expression lead to the disease. However, many aberrations are novel or extremely rare making clinical interpretation problematic and genotype-phenotype correlations uncertain. By identification of several patients sharing a region of genomic deletion or duplication and with phenotypic features in common, greater certainty can be given to the pathogenic nature of the rearrangement and new syndromes defined.

Challenge 2 - to chart common genetic variation

Although linked with the complexity of clinical interpretation, charting common variation in the human genome merits separate mention. Copy number changes which are common in the normal population should be taken into account and excluded from the list of aberrant genomic regions in a patient, in order to be able to perform an educated analysis on the impact of aberrant genes or genomic loci on the patient phenotype, diagnosis, prognosis, treatment and counseling. Such filtering enables the

Table 4.1: An overview of chromosomal aberration databases. These efforts aim at aggregating chromosomal aberration information at various levels of resolution and annotate the genome with case reports, congenital anomalies or phenotypes.

Database	Description
Catalogue of Unbalanced Chromosome Aberration in Man	Albert Schinzel's comprehensive catalogue of chromosomal aberrations in man in book form. The catalogue is a standard reference for clinicians treating patients with autosomal chromosome aberrations and for physicians and biologists working in cytogenetic laboratories and human genetic institutes [207].
Chromosome Abnormality Database (`www.ukcad.org.uk/cocoon/ukcad/`)	The UK Association of Clinical Cytogeneticists (ACC) Chromosome Abnormality Database (CAD) is a collection of both constitutional and acquired abnormal karyotypes reported by UK Regional Cytogenetics Centers. It is open to all Genetics Professionals, and available for searches on different abnormalities and karyotypes in both a clinical context as for medical research.
Chromosome Anomaly Collection (`www.ngrl.org.uk/Wessex/collection.htm`)	a catalogue of unbalanced structural chromosome abnormalities (USCA) without phenotypic effect. The Collection also includes the cytogenetically visible euchromatic variants as part of the continuum of copy number variation in the human genome.
DECIPHER (`www.sanger.ac.uk/PostGenomics/decipher`)	DatabasE of Chromosomal Imbalance and Phenotype in Humans using ENSEMBL Resources, see text
Database of Chromosome Aberrations in Cancer (`cgap.nci.nih.gov/Chromosomes/Mitelman`)	the Mitelman catalog for cancer cytogeneticists is a standard reference database that compiles information on chromosome changes identified in human neoplasms. The electronic version supports searches by karyotype, reference, tumor type, and location.
ECARUCA (`www.ecaruca.net`)	European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations, see text
The Human Phenome Project	a proposed international effort to create comprehensive phenomic databases of systematically collected phenotypic information, and to develop approaches for analyzing such phenotypic data [71].
Mendelian Cytogenetics Network DataBase	(`www.mcndb.org`) - an online database on disease associated balanced chromosomal rearrangements, containing information on breakpoints and clinical features and disease potential. Aims at initiating collaborative studies of specific disorders.
Online Database of Chromosomal variation in man (`www.wiley.com/legacy/products/subject/life/borgaonkar/access.html`)	a systematic collection of important citations from the world's literature reporting on all common and rare chromosomal alterations, phenotypes, and abnormalities in humans. The database is organized by variations and anomalies, numerical anomalies, and chromosomal breakage syndromes [31].
Progenetix (`www.progenetix.de`)	a database of published cytogenetic abnormalities in human malignancies, mostly from CGH experiments

identification of those changes which are novel and potentially pathogenic and aids differentiation between pathogenic and polymorphic copy number changes.

As noted in Section 4.1, public and private initiatives have emerged and are being developed to chart common variation. The database of Genomic Variants and the Redon Copy Number Variation sets are two examples. Also, many individual cytogenetics groups or local consortia are developing maps of genomic variability for their own use or for publication. It is important to acknowledge that genomic variation can vary locally by sub-population, a phenomenon that makes interpretation and use of such data collections much more delicate.

Challenge 3 - to link back to known genetic disease

Tapping into the knowledge of known genetic syndromes is important. This can be done through linking to external resources such as OMIM, but equally by putting patients in perspective of a local repository of known diseases. Another important aspect aside from perusing public case report repositories, disease and case control databases, and curated literature, is enabling users to go through a historic trail of patient records in their own local database.

Challenge 4 - flexible search and information retrieval

So as to be able to review the historic trail of patient reports, flexible search tools are needed, which can retrieve case reports through genotype or loci searches, phenotype traits, full text search, and importantly, similarity. Through defining similarity measures for genotypes, phenotypes, clinical features, ... for patients, not only search is facilitated, but also clustering, automated suggestion of relevant cases and facts, and so on.

Challenge 5 - to facilitate collaboration

Intelligent next generation genomic platforms need to support collaboration on at least three levels. Firstly, they need to enable collaboration between clinical geneticists, lab technicians, molecular cytogeneticists, pediatricians, ... within the genetic testing center or research group. Ideally, this interfaces with a flexible project management and access rights system. Secondly, these platforms need to support local collaborations and consortia, where regional hospitals that work on the same array platform, can exchange patient information and divide responsibilities within the consortium, whether it be local, regional, national or international. This requirement makes a strong case for a web-based approach, but leads to specific security and storage issues that need to be addressed. Thirdly, there is a great need for international collaboration to report and catalogue genotype-phenotype correlations such that clusters of individuals sharing similar genomic rearrangements and phenotypes can be identified. Ideally, platforms can communicate patient information with each other, while taking into account regulatory aspects, informed consent, and privacy issues.

Challenge 6 - uniform annotation of phenotypes

This subject has been thoroughly discussed in Sections 3.3 and 3.5.1, and caveats involved in using ontologies for annotating patient phenotypic traits of patients were discussed there as well. We have noted the complex form and semantics of patient phenotypes [231] and the rich context around it [177] is well represented and analyzed.

Consistent evaluation of phenotypes by multiple clinical geneticists, uniform phenotype classification criteria, keeping phenotype and clinical data be amenable to large scale data integration, exchange, and mining, requires formalization and rigid use of well built classification schemes such as dedicated phenotype taxonomies and ontologies. This requirement does not only hold for dysmorphic features but also for genetic anomalies such as copy number alterations themselves.

Challenge 7 - flexible reporting

Regardless of which platforms are used in a clinical routine or research setting, the need for flexible reporting is clear. Especially in routine diagnostics, speed of patient throughput is important and cost of manual labor spent due to review and reporting is high. Therefore, a lab work-flow benefits from a flexible and automated reporting system.

Challenge 8 - interacting with external resources

Contextualizing local data to outside information is crucial and can be achieved through superficial linking with external databases (by providing links to resources like OMIM, PUBMED, Ensembl), temporary integration of data for manual review (viewing these same resources in line with patient information, integrating them in reporting systems, or feeding internal patient data to external genome viewers through the DAS protocol through a principle called distributed genome annotation), and actual transfer and sharing of case reports (through construction of data bridges between platforms, with the added advantage that patient records needn't be entered multiple times across different databases).

In a similar fashion, there is benefit in making platforms compatible with other databases and software systems, be it for extraction of data to external tools (Bench's export to Microsoft Excel, importing data from software tools of array CGH technology providers, exporting reports to existing LIMS systems and databases, transfer of genotype information between ArrayCGHBase and Bench, transfer of patient records from Bench to DECIPHER, and so on).

Also, these platforms need to be able to put all data in perspective of the most recent genome assembly. This does not just imply accurate mapping on raw sequence, but also presenting correct, complete and up to date gene lists, functional annotations, and so on.

Challenge 9 - management of raw assay data

Some regulatory bodies require that patient data is kept in storage for a minimum period of 10 years. Apart from this requirement, the management of actual raw assay data, but also of the heterogenous information with which both experiments as well as patients are annotated, is a central issue to genomics IT platforms. These platforms need to manage all assay related information, from raw data over called copy number variant regions, polymorphisms, or other genomic feats of interest, to the actual patient record with clinical and phenotype information. This is especially important in a high throughput context, and has clear data mining advantages, as illustrated in Section 2.4.3. Ideally, platforms will support a routine clinical diagnostics work flow from raw assay to final report, as well as biomedical research projects.

Challenge 10 - storage and computing capacity requirements

As these platforms become central repositories for patient related data, security and privacy requirements cannot be underestimated. Databases need to be robust and reliable (running on a distributed infrastructure to guarantee uptime and safe fail over when systems issues arise), scalable (allowing for growth and the adoption of new assay technologies as they arise), secure (providing encryption, authorization, authentication, audit trails) and safe (enabling or providing regular back-ups of data). Clearly, with the growth of data volumes, scalability is not limited to the storage needs for data volumes that stem from the increasing resolution of array platforms and the advent of next generation sequencing technologies, but also to processing power supporting algorithms processing raw data or huge numbers of patient annotations as case report databases grow in size.

While some of these challenges are especially delicate in the context of web-based platforms (the internet is an insecure environment), there are advantages to be noted. Specifically, this context provides a case for so-called Cloud Computing, a concept that points to using the Internet as an environment over which to use computer technology – It is a style of computing in which IT-related capabilities are provided as a service, allowing users to access technology-enabled services from the Internet ("in the cloud") without knowledge of, expertise with, or control over the technology infrastructure that supports them. Cloud computing is a general concept that incorporates software as a service (SaaS), Web 2.0 and other recent, well-known technology trends, in which the common theme is reliance on the Internet for satisfying the computing needs of the users. This concept is being introduced not just for former desktop applications (Google Apps and GMail) but is gaining influence as a middle-ware for very high traffic social web sites, and as a back bone for services that have very high computing power and/or storage needs. In this paradigm, applications are provided online and are accessed from a web browser, while the software and data are stored on a very large grid of distributed servers that communicate over the internet.

Challenge 11 - contextual analysis of assay data

The genome of a patient relates directly to that of his/her parents, siblings or other family members and even to his/her subpopulation. Therefore, genomics platforms need to be able to deal with familial links, with the analysis of Trios as the most clear cut example. Analysis of DNA samples from one or both parents needs to be put in relation to copy number changes found in the patient. This enables the user to see immediately whether a genomic aberration identified in a patient is of parental origin or has occurred as a *de novo* event. Other situations occur where assays are confounded and do not represent a single patient. A clear example of this are alternative designs such as loop designs, of which a 3-way loop design is a specific case, as discussed in Section 2.3. Also, some technology platforms provide result files where multiple assays are grouped into single result files. Also, in the context of oncology, multiple tissues can be treated at the same time and give rise to added complexity in analysis.

Challenge 12 - quality assurance

Data quality is imperative for a correct interpretation. There is a clear need for a system of objectively assessing laboratory performance by an outside agency. QA schemes exist both within consortia as in public initiatives. Tested DNA is sent out in a blinded fashion to independent laboratories to re-analyze on various array platforms

and results are fed back to the referring laboratory. The main objective of external quality assessment is to establish inter-laboratory compatibility. External Quality Assessment processes aim at continuous improvement of processes and do not merely involve wet lab and work flow processes, of which QA should be supported by IT platforms, but should involve the IT processes themselves as well.

Challenge 13 - identifying loci responsible for features of the phenotype

Tools that aid analysis and can serve to identify which gene or genes in a deleted or duplicated region may be responsible for features of the phenotype (a process that, when performed manually, involves considerable literature and database searching) can greatly improve a speedy, correct and informed diagnosis, provide a basis for sound genetic counseling, and improve our understanding of gene function and disease. Many individual tools are geared at addressing this specific challenge. They have been discussed in great detail in Section 3.2.3. Their integration in genomics platforms is highly desirable.

4.4 DECIPHER as an Array-CGH genomics platform

To facilitate the analysis of rare clinical genetics events, an interactive web-based database was developed, called DECIPHER [2] - (**D**atabas**E** of **C**hromosome **I**mbalance and **P**henotype in **H**umans using **E**nsembl **R**esources) - with a suite of tools designed to facilitate the interpretation of submicroscopic chromosomal imbalances, inversions and translocations. When DECIPHER is used in conjunction with the Database of Genomic Variants[3] or the Copy Number Variation Project[4] at the Wellcome Trust Sanger Institute, copy number changes which are common in the normal population can be excluded, thus enabling the identification of those changes which are novel and potentially pathogenic. DECIPHER facilitates genetic counseling by retrieving relevant information from a variety of bio-informatics resources. Known and predicted genes within a given aberration are listed in the DECIPHER patient report and genes of recognized clinical importance are highlighted.

DECIPHER enables clinical scientists worldwide to maintain records of phenotype and chromosome rearrangement for their patients within the database and, with informed consent, share this information with the wider clinical research community through display in the genome browser Ensembl. By sharing cases worldwide, clusters of rare cases having phenotype and structural rearrangement in common can be identified, thus leading to the definition of new syndromes and furthering understanding of gene function.

4.5 DECIPHER and copy number changes

Advances in molecular cytogenetic techniques and sequencing of the human genome now enable chromosome rearrangements to be analyzed at an unprecedented level

[2] `decipher.sanger.ac.uk`
[3] Database of Genomic Variants - `projects.tcag.ca/variation/`
[4] Copy Number Variation Project - `www.sanger.ac.uk/humgen/cnv/`

of accuracy. The resolution of conventional Giemsa-banded chromosome analysis is 5-10Mb, whereas that of a high-density genomic array is 100kb, resulting in an approximate 50-100-fold improvement in sensitivity. Using high resolution genomic array analysis, disorders of chromosome imbalance (copy number change) can now be identified with such precision that mapping the rearrangements onto the reference sequence of the human genome becomes a realistic proposition.

4.5.1 Copy number changes and disease

As discussed earlier in Chapter 2, this approach is proving particularly effective for the investigation of patients with developmental delay, learning disability, dysmorphic features and/or congenital anomalies and is generally identifying the underlying cause of the disease phenotype in approximately 15% of previously undiagnosed cases [260, 224, 254, 234, 140]. Moreover, a genomic basis to several later-onset disorders has now been defined, *e.g.* Charcot-Marie Tooth disease (CMT) [143], early-onset Alzheimer disease with amyloid angiopathy (EOAD) [199] and adult onset autosomal dominant leukodystrophy [172]. However, copy number changes in many patients are novel or extremely rare such that uncertainty remains as to whether the aberration is pathogenic or simply a benign variant. Identification of additional patients sharing a region of genomic deletion or duplication and with phenotypic features in common allows greater certainty to be given to the pathogenic nature of the rearrangement and new syndromes defined.

4.5.2 Normal copy number variation

Recently, several studies highlighted the presence of large numbers of deletions, insertions and duplications in the normal population, ranging from a few kilobases to several megabases in size [109, 213]. For example, Redon *et al.* [190] using the complementary technologies of single-nucleotide polymorphism (SNP) genotyping arrays, and clone-based comparative genomic hybridization generated a global map of copy number variable regions (CNVRs) in the human genome and found 1447 CNVRs covering approximately 12 % of the human genome which were variable in copy number. As more studies are completed and as the resolution of genomic array analysis increases more copy number variable regions are being discovered. There are currently over 8000 CNVRs listed in the Database of Genomic Variants [5]. These regions contain several hundreds of genes, disease loci and segmental duplication regions [220]. As patients also have these normal copy number variants, with the use of high resolution genomic arrays, it has become a challenge to identify which changes are normal variants and which are likely to be associated with the phenotype.

4.5.3 The DECIPHER project

While many clinical centers are now applying genomic microarray technology to investigate patients with developmental delay, learning disability and congenital malformation [257], the sporadic nature and rarity of the majority of these cases limits the ability of the individual clinician to interpret the molecular findings from genome-wide array analysis. Thus, there is a great need for international collaboration to report

[5] DGV - `projects.tcag.ca/variation/`

and catalogue genotype-phenotype correlations such that clusters of individuals sharing similar genomic rearrangements and phenotypes can be identified. This will not only facilitate diagnosis and genetic counseling but also improve our understanding of gene function and disease.

The DECIPHER project was initiated in 2004 with the general aim of providing a clinical and research tool to:

- Aid in the interpretation of data from genomic microarray analysis *e.g.* the differentiation between pathogenic and polymorphic copy number changes.
- Use the human genome map via the Ensembl [6] genome browser to define genes involved in a specific microdeletion, microduplication, translocation or inversion.
- Facilitate collaboration between clinical geneticists and molecular cytogeneticists to accelerate progress in the delineation of new syndromes and of gene function.

The DECIPHER projects builds on several independent but interlinked developments:

- **The Human Genome Mapping project.** The publication of the finished human genome sequence in October 2004 [131, 46] provided a reference sequence, or assembly map, onto which chromosome rearrangements and the order and position of genes and predicted genes can be placed.
- **Bio-informatics.** Crucial to the operation of DECIPHER was the development of the Ensembl genome browser [107] which allows users to select and view an annotated segment of the human genome map and connect to other relevant resources via hyperlinks.
- **Molecular cytogenetics.** The clone resources developed in the Human Genome Mapping Project, the SNP Consortium Project and indeed the sequence itself have provided cytogeneticists with a new suite of molecular tools with which to analyze chromosome rearrangements. The use of fluorescence in situ hybridization (FISH) and in particular the introduction of genomic array analysis ('molecular karyotyping') using microarray-based comparative genome hybridization (array-CGH) or high-density single-nucleotide polymorphism (SNP) genotyping has revolutionized the identification of subtle (hitherto unidentifiable) cytogenetic imbalances.

4.5.4 The DECIPHER consortium

Contributing to the DECIPHER database is a Consortium. The DECIPHER Consortium is an international network of academic departments of Clinical Genetics comprising more than 100 centers as of September 2008. Each contributing center has a nominated Clinical Geneticist (with expertise in dysmorphology) and a nominated Molecular Cytogeneticist who are jointly responsible for data entry for their center. Each center maintains control of its own patient data (which is password protected within the center's own DECIPHER project) until patient consent is given to allow anonymized genomic and phenotypic data to become viewable within Ensembl [7]. Once data is shared, consortium members are able to gain access to the patient report and contact each other to discuss patients of mutual interest thus facilitating the

[6] Ensembl - `www.ensembl.org`
[7] Ensembl - `www.ensembl.org`

delineation of new microdeletion/microduplication syndromes. With patient consent, positional genomic information together with a brief description of the associated phenotype becomes viewable without password protection via the DECIPHER track in Ensembl. This is of benefit to clinicians advising patients with similar findings and also of benefit to researchers working on the specific disorders seen in that patient (*e.g.* Congenital heart disease or cleft lip) or the role of genes contained within the aberration.

4.5.5 Key features of DECIPHER

DECIPHER integrates seamlessly with the Ensembl genome browser

DECIPHER interrogates the current version of the human genome assembly displayed in the Ensembl genome browser. Since DECIPHER is a dynamic system, each time it is interrogated the most recent data with regard to gene content is returned into the Patient Report. Ensembl itself builds all its annotated genes and associated features onto the most recent version of the human genome assembly which is created by the NCBI. The current assembly is NCBI-36. The annotation of genes and their features are continually being reviewed, refined and updated, culminating bi-monthly in a new version of the annotation, released onto the Ensembl web site. New features and types of data are continually added, routinely using the distributed annotation system (DAS) which allows a set of data to be dropped into place, and viewed across the genome as a new feature, *e.g.* copy number variation (CNV) data. This approach makes the whole process very fluid and allows rapid updates to data to be made available.

DECIPHER links to other genetics databases

DECIPHER is linked to other genetic and medical databases including HUGO Gene Nomenclature Committee (HGNC)[8], On-line Mendelian Inheritance in Man (OMIM)[9], PubMed, GeneReviews[10], Ensembl genes, Swiss-Prot and a frequently updated list of emerging bio-informatics databases. DECIPHER not only displays detected aberrations in relation to the genome sequence, but has been configured to allow the clinician or researcher to rapidly obtain clinically relevant information (publications) about the genes located in these regions.

DECIPHER is configurable to work with any position data mapped onto the reference sequence

For DECIPHER to interact directly with the Ensembl genome browser, positional data regarding the array must first be mapped onto the reference sequence. DECIPHER can be configured to utilize any molecular method (*e.g.* FISH, MAPH, MLPA, PCR, array-CGH, SNP-genotyping) where the probes or SNPs are positioned on the reference sequence. Any structural rearrangement defined in this way can be catalogued within DECIPHER including copy number changes, translocations and inversions.

[8]HUGO Gene Nomenclature Committee - `www.gene.ucl.ac.uk/nomenclature/`
[9]Online Mendelian Inheritance in Man - `www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM`
[10]GeneReviews- `www.geneclinics.org/`

DECIPHER uses a restricted ontology of phenotype terms

DECIPHER uses the well-known hierarchy of phenotype terms developed for the The Baraitser-Winter Neurogenetics Database [16]. In this way, clinicians in different centers use a common set of terms which facilitates consistent phenotype description and enables efficient data sharing and database searching.

Copy number change categorization

A major problem in the identification of potentially pathogenic changes is the large number of normal variants which are identified particularly as array resolution increases. DECIPHER include a feature graph tool which displays patient copy number changes together with data from the Database of Genomic Variants [11], as well as variants identified in a selection of the major studies of copy number variation in normal individuals. Together with the analysis of parental samples (analysis of so called trios), each copy number change in the patient can then be classified as *de novo*, CNV (normal variant) or as a familial variant. Only *de novo* changes and familial variants are then displayed, with consent, in Ensembl.

Gene lists

The gene lists and gene displays generated by DECIPHER are passed via HGNC to eliminate redundancy of genes caused by synonyms and to ensure that DECIPHER is using currently approved gene names. Genes included in the OMIM Morbid database, *i.e.* those of known importance in human disease are denoted with the suffix 'M' so that the clinician is immediately alerted to explore their potential clinical significance. Similarly, imprinted genes listed at the web-site GeneImprint[12] are denoted by the suffix 'I' since in this instance, an inherited deletion/duplication, although present in a normal parent, may be potentially pathogenic.

Search

In order to ensure its availability to the wider medical and scientific community, all of the consented data held within the DECIPHER database is searchable. Features which can be queried include chromosomal band, genomic position, chromosomes and phenotypes or a combination of these. DECIPHER members can also use this search function to interrogate non-consented data within their own project group.

Analysis of trios

DECIPHER has developed novel software to assist in the analysis of trios (DNA samples from a patient and both parents). DECIPHER collects information from array analysis of both parents and displays this in relation to copy number changes found in the patient. This enables the user to see immediately whether a genomic aberration identified in a patient is of parental origin or has occurred as a *de novo* event.

[11] DGV - `projects.tcag.ca/variation/`
[12] GeneImprint - `www.geneimprint.com/`

Quality assurance (QA)

Data quality is imperative for a correct interpretation of the data. It is crucial that members of the consortium as well as the public have confidence in the data quality submitted to DECIPHER. To address this concern, a QA scheme has been introduced to DECIPHER members. Tested DNA is sent out on an annual basis to all participating laboratories, which each analyze the samples on their respective array platforms and feed back the results to the reference laboratory.

Reports

Once all the phenotype and genotype data of a patient have been entered into DECIPHER, two types of printable reports are directly available from the web site. The first report comprises a summary of the full clinical phenotypic description and an overview of genes affected by the aberration. This report can be used as a clinical report and includes the karyotype and ideogram of the affected chromosome(s). It is possible to add patient images to this report. Alternatively, a simplified version of the full clinical report can be printed out for counseling patients and families.

Confidentiality

In order to protect patient privacy, data in DECIPHER is served over an encrypted SSL (secure socket layer) connection, similar to a security level employed by banks and financial institutions. Clinical photographs held in DECIPHER are password protected so that they are only accessible to logged-in members of the consortium. Photographs are also digitally watermarked for additional security.

Syndromes

(xi) DECIPHER Syndromes - The Syndromes pages within DECIPHER provide a single, curated resource of information and web-links. DECIPHER Syndromes is supported by a panel of Expert Advisors, each responsible for reviewing and updating the entry for a specific syndrome on an annual basis. For each syndrome entity DECIPHER provides:

- a brief clinical synopsis
- information regarding the size and origin of a deletion/duplication
- an ideogram of the location of the deletion/duplication on the relevant chromosome
- a list of the genes contained within the aberrant interval
- an up-to-date publication reference list
- a direct clickable link to visualize the deletion/duplication in Ensembl

Gene prioritization

Identifying which gene or genes in a deleted or duplicated region may be responsible for features of the phenotype involves considerable literature and database searching. DECIPHER includes an interface to the results of the text-mining pipeline that was discussed in detail in Section 3.5 of Chapter 3. The gene prioritization interface orders genes in the aberrant regions of a patient according to their likelihood of being

associated in the literature with a phenotypic feature or group of features that are annotated to this patient.

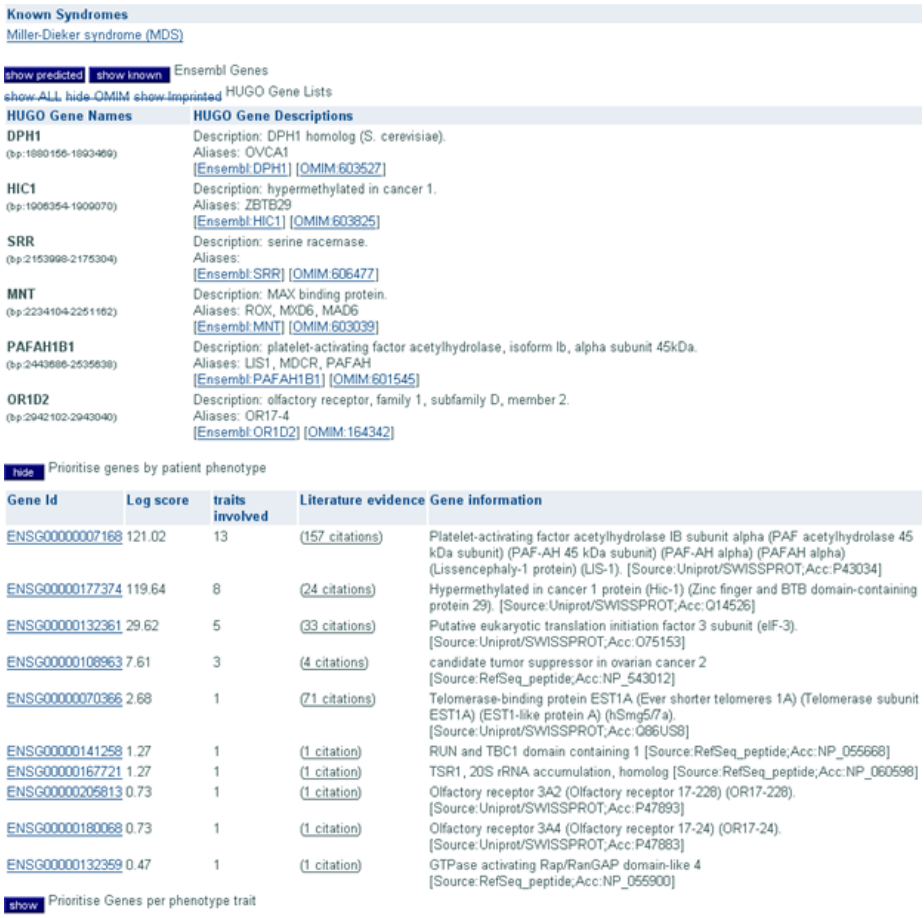
A DECIPHER Patient report also includes literature based gene prioritization: all genes present in a region of copy number variation for a patient report are ranked according to their relation to the patient phenotype. In this way, candidate genes that are causal to all or part of the phenotype can be identified. Genes in copy number variable regions can be prioritized either by the entire patient phenotype or by individual traits that make up the phenotype.

To support this feature, DECIPHER contains a large database of gene-to-phenotype associations that are collected from biomedical literature by text-mining following the process described in Section 3.5. These associations bear p-values that express the strength of the link between gene and phenotype, and are mined from a selection of 848625 MEDLINE abstracts that are relevant to the field of constitutional genetics. Abstracts are selected for analysis when they are deemed relevant (ie. are linked to a human gene in Entrez Gene or are published in a journal that publishes on human genetics), published within the last 20 years, and contain information on both genes and phenotypes. This set is passed through the public EBI Web Services that underly EBIMed [189] to identify gene and protein names as well as any associated abbreviations and alternative symbols. Also, all phenotypes described in the London Neurogenetics database (LNDB)[13] as well as their synonyms and variants are identified. This annotated text corpus is then used to construct the genotype-phenotype database that underlies DECIPHER, using a method described earlier [253]. For a detailed explanation of the text-mining pipeline, see Section 3.5. As of August 2007, the database contains 496038 gene-to-phenotype associations, for 14972 unique gene identifiers and 1164 unique LNDB phenotype terms. Counts may vary with each update of analysis pipeline and text corpus.

The prioritized gene list that is generated for a DECIPHER patient report, scores the relation between each gene and phenotype, and provides links to the literature evidence that was used to infer the association between a gene from the copy variant region and the patient phenotype. A screen shot of a Decipher patient report is shown in Figure 4.1. This figure shows a report for a patient that has an aberration in the region known to be associated to the Miller-Dieker syndrome (MDS). The report is only shown partially. Above, the HUGO names of affected genes in the aberrant region are listed, together with gene descriptions. Below, the prioritization of these genes by the patient phenotype is shown. In this table, genes are ranked according to their likelihood of being involved in the patient phenotype. In the first column, it shows the same affected genes, with links to Ensembl in the first column. The second column shows a score on how strong the link is between this gene and the patient phenotype. The third column provides a count on how many of the individual phenotype traits that make up the full patient phenotype have actually been associated to the gene in that row. The second to last column shows the number of citations that corroborate this gene-to-phenotype correlation. This number can be clicked to make the table fold open and show a list of ten links to relevant PUBMED abstracts on this gene and one or more of the phenotypes. The last column provides additional information on the gene.

The calculation of the score is straightforward. The score represents how well a single gene correlates with the complete patient phenotype. It is the sum of all scores for that gene with each individual phenotype trait that makes up part of the patient

[13]Baraitser-Winter Neurogenetics database (2007). Version 1.0.11 - `www.lmdatabases.com`



Known Syndromes

Miller-Dieker syndrome (MDS)

show predicted | show known Ensembl Genes

~~show ALL~~ ~~hide OMIM~~ ~~show Imprinted~~ HUGO Gene Lists

HUGO Gene Names	HUGO Gene Descriptions
DPH1 (bp:1880156-1893469)	Description: DPH1 homolog (S. cerevisiae). Aliases: OVCA1 [Ensembl:DPH1] [OMIM:603527]
HIC1 (bp:1906354-1909070)	Description: hypermethylated in cancer 1. Aliases: ZBTB29 [Ensembl:HIC1] [OMIM:603825]
SRR (bp:2153998-2175304)	Description: serine racemase. Aliases: [Ensembl:SRR] [OMIM:606477]
MNT (bp:2234104-2251162)	Description: MAX binding protein. Aliases: ROX, MXD6, MAD6 [Ensembl:MNT] [OMIM:603039]
PAFAH1B1 (bp:2443686-2535638)	Description: platelet-activating factor acetylhydrolase, isoform Ib, alpha subunit 45kDa. Aliases: LIS1, MDCR, PAFAH [Ensembl:PAFAH1B1] [OMIM:601545]
OR1D2 (bp:2942102-2943040)	Description: olfactory receptor, family 1, subfamily D, member 2. Aliases: OR17-4 [Ensembl:OR1D2] [OMIM:164342]

hide Prioritise genes by patient phenotype

Gene Id	Log score	traits involved	Literature evidence	Gene information
ENSG00000007168	121.02	13	(157 citations)	Platelet-activating factor acetylhydrolase IB subunit alpha (PAF acetylhydrolase 45 kDa subunit) (PAF-AH 45 kDa subunit) (PAF-AH alpha) (PAFAH alpha) (Lissencephaly-1 protein) (LIS-1). [Source:Uniprot/SWISSPROT;Acc:P43034]
ENSG00000177374	119.64	8	(24 citations)	Hypermethylated in cancer 1 protein (Hic-1) (Zinc finger and BTB domain-containing protein 29). [Source:Uniprot/SWISSPROT;Acc:Q14526]
ENSG00000132361	29.62	5	(33 citations)	Putative eukaryotic translation initiation factor 3 subunit (eIF-3). [Source:Uniprot/SWISSPROT;Acc:O75153]
ENSG00000108963	7.61	3	(4 citations)	candidate tumor suppressor in ovarian cancer 2 [Source:RefSeq_peptide;Acc:NP_543012]
ENSG00000070366	2.68	1	(71 citations)	Telomerase-binding protein EST1A (Ever shorter telomeres 1A) (Telomerase subunit EST1A) (EST1-like protein A) (hSmg5/7a). [Source:Uniprot/SWISSPROT;Acc:Q86US8]
ENSG00000141258	1.27	1	(1 citation)	RUN and TBC1 domain containing 1 [Source:RefSeq_peptide;Acc:NP_055668]
ENSG00000167721	1.27	1	(1 citation)	TSR1, 20S rRNA accumulation, homolog [Source:RefSeq_peptide;Acc:NP_060598]
ENSG00000205813	0.73	1	(1 citation)	Olfactory receptor 3A2 (Olfactory receptor 17-228) (OR17-228). [Source:Uniprot/SWISSPROT;Acc:P47893]
ENSG00000180068	0.73	1	(1 citation)	Olfactory receptor 3A4 (Olfactory receptor 17-24) (OR17-24). [Source:Uniprot/SWISSPROT;Acc:P47883]
ENSG00000132359	0.47	1	(1 citation)	GTPase activating Rap/RanGAP domain-like 4 [Source:RefSeq_peptide;Acc:NP_055900]

show Prioritise Genes per phenotype trait

Figure 4.1: This figure shows a report for a patient that has an aberration in the region known to be associated to the Miller-Dieker syndrome (MDS). The report is only shown partially. Above, the HUGO names of affected genes in the aberrant region are listed, together with gene descriptions. Below, the prioritization of these genes by the patient phenotype is shown. In this table, genes are ranked according to their likelihood of being involved in the patient phenotype. In the first column, it shows the same affected genes, with links to Ensembl in the first column. The second column shows a score on how strong the link is between this gene and the patient phenotype. The third column provides a count on how many of the individual phenotype traits that make up the full patient phenotype have actually been associated to the gene in that row. The second to last column shows the number of citations that corroborate this gene-to-phenotype correlation. This number can be clicked to make the table fold open and show a list of ten links to relevant PUBMED abstracts on this gene and one or more of the phenotypes. The last column provides additional information on the gene.

phenotype. These individual scores, in turn, are calculated as $-log_{10}(p)$, where p are the very same p-values as discussed in Section 3.5. This scoring scheme is illustrated in Figure 4.2.

Phenotype	PAFAH1B1	HIC1	eIF-3	...
CRANIUM, general abnormalities	12.58	-	-	...
Fingers, general abnormalities	1.38	41.95	0.98	...
Lissencephaly/pachygyria	28,54	16.01	-	...
Tonic/clonic seizures	29.41	-	-	...
Mouth	1.45	2.61	8.10	...
...	...	...	...	...
Sum of log scores	**121.02**	**119.64**	**29.62**	...

Lissencephaly / pachygyria: LNDB 32.32.09
PAFAH1B1 (LIS1): ENSG00000007168
***p*-value** 2.8951×10^{-29}
- log(*p*): 28,54

Figure 4.2: The DECIPHER text based gene prioritization scoring scheme. The score represents how well a single gene correlates with the complete patient phenotype. It is the sum of all scores for that gene with each individual phenotype trait that makes up part of the patient phenotype. These individual scores, in turn, are calculated as $-log_{10}(p)$, where p are the very same p-values as discussed in Section 3.5.

As DECIPHER is directly linked to Ensembl, genes which are located in affected regions of the genome can be identified and queried by accessing gene specific databases. Furthermore, additional relevant information from the literature held in PubMed can be interrogated as DECIPHER indexes all relevant journals in PubMed, including their abstracts.

4.5.6 Using DECIPHER

Cases are entered into DECIPHER using standard web browsers and the data is mapped directly to the latest human genome assembly via the Ensembl genome browser. For consented patients, the aberrant region is displayed together with the phenotype along with any other similar cases within the database and within the context of other genomic features such as gene content, segmental duplication and regions of normal copy number variation.

A manuscript on DECIPHER has been published [65]. The manuscript does not cover the gene prioritization as thoroughly as the past sections, but does discuss some patient case studies in great detail.

4.5.7 Main benefits of DECIPHER

DECIPHER as a clinical tool

Genomic microarray analysis brings the opportunity for much higher diagnostic yield than conventional chromosome imbalance. It also brings challenges for the clinical team in the interpretation of unfamiliar findings. DECIPHER enables clinicians to evaluate an unfamiliar aberration by:

1. Comparing the finding against the copy number variation tracks to determine whether it is a variant that has been seen before in the normal population.
2. Showing whether it overlies a known Syndrome. Whereas clinicians are familiar with the small number of well-recognized microdeletion disorders which predate genomic microarray analysis, the rapid proliferation of new syndromes over the past year, *e.g.* 17q21.3 [223, 124, 219] 15q24 [219, 221], TAR syndrome susceptibility locus [122], means that most clinicians will struggle to stay abreast of this knowledge.
3. Displaying whether the aberration has been seen before by another member of the DECIPHER consortium and if so how the phenotypes of the patients compare.
4. Determining the gene content of the aberration. For sub-microscopic imbalances, the size of the deletion is a poor guide to its gene content since the density of genes in the human genome is very uneven. Some regions are particularly gene rich, whereas other regions are virtual gene deserts. Gene content plays an important role in assessing the possible pathogenicity of a novel aberration.
5. Establishing whether the aberration contains genes of known clinical importance and enabling the clinician to access the literature to evaluate the extent to which these genes may explain the phenotype seen in the patient. This enables clinicians to be alerted to any genes of potential clinical importance *e.g.* tumor suppressor gene which when deleted have important implications for patient management.

Currently our understanding of the clinical consequences of copy number change is very incomplete. When evaluating a novel copy number change, a large number of different potential mechanisms of inheritance need to be borne in mind before dismissing a change as not linked to the phenotype. These include imprinting, unmasking of a recessive mutation on the other allele, deletion/duplication of regulatory elements, complex inheritance where the phenotype results from the interplay of more than one genomic alteration and/or SNPs and the possibility that regions that are recognized copy number variants in the normal population may have a threshold effect whereby normal variation tolerates copy number variation within certain parameters, but beyond or below a given threshold, copy number change becomes pathogenic.

These factors, together with awareness that the human genome map continues to evolve towards a more complete reference sequence with consequent repositioning of genes and other elements, need to be considered [144, 142] as they necessarily imply a degree of caution in interpreting and using data in DECIPHER.

DECIPHER as a research tool

Many genes in the human genome are still without known function in human development and disease. For other genes, the function may be predicted by cross-species

comparison, but the disease correlates in humans are not yet known. Dominantly inherited disorders (where the phenotype does not abolish reproductive fitness) can be mapped by linkage analysis. Likewise, homozygosity mapping can refine the position of genes responsible for recessively inherited disorders in consanguineous families. However, mapping genes for severe *de novo* dominant disorders which are reproductively lethal typically relies upon identifying chromosomal aberrations to obtain positional information about gene location. The importance of genomic rearrangement as a cause of sporadic traits has only recently been recognized [142]. Indeed, the de novo rate for genomic rearrangement may be orders of magnitude greater than for point mutation.

Chromosome rearrangements have historically been important in identifying disease genes using a variety of approaches:

1. the coincidence of a sporadic occurrence of the disorder with an apparently balanced *de novo* autosomal reciprocal translocation *e.g.* the identification of SATB2 as a cleft-palate gene [221] and SOX2 gene as an anophthalmia gene [122], or
2. the segregation of a phenotype with a translocation *e.g.* the identification of EOMES as a cause of severe neurological malformation (microcephaly with polymicrogyria and agenesis of the corpus callosum) [66], or
3. the fine mapping of visible deletions *e.g.* the identification of the TSC2 gene [59]. Because high resolution genomic microarray analysis has the potential to identify very small aberrations which may contain only a few genes, it offers an enhanced opportunity for the discovery of gene function [11] *e.g.* identification of CHD7 as the CHARGE syndrome gene [45].

Since DECIPHER brings together genomic and phenotypic data about submicroscopic chromosome aberrations and translocations within the Ensembl genome browser it can facilitate this process.

Reverse dysmorphology genomics

DECIPHER enables clinical scientists worldwide to maintain records of phenotype and chromosome rearrangement for their patients within the database and, with informed consent, share this information with the wider clinical research community through display in Ensembl. By sharing cases worldwide, critical numbers of rare cases with phenotypes and structural rearrangements in common can be identified. In classical genetics, individuals with common phenotypes are identified and a clinical syndrome is delineated *e.g.* Supravalvular aortic stenosis and characteristic facies and neurodevelopmental profile in Williams-Beuren syndrome (WBS) [262]. In due course a common genetic basis to the clinical syndrome is recognized. In the case of WBS, this is a recurrent microdeletion on 7q11.23 [261].

Importantly, genomic microarray analysis is enabling the delineation of new syndromes which have less distinctive clinical features by a process which can be termed 'reverse dysmorphology' or 'reverse genomics'. Here, the classical practice of phenotype-genotype correlation is inverted. In 'reverse genomics' clusters of cases are first recognized by their overlapping copy number changes and subsequently the phenotypes of the individuals in the cluster are compared. This new approach permits the recognition of common clinical features which were initially too subtle or too variable (when seen in patients of different age, sex and ethnicity) to enable a new syndrome to be identified solely on clinical grounds. This approach may be particularly valuable for

the investigation of disorders with high locus heterogeneity such as learning disability or mental retardation.

4.6 Store+Bench

We have seen how DECIPHER addresses many of the challenges listed in Section 4.3.1. Now, we'll discuss how Store+Bench fulfills this task. Functionally, DECIPHER and Store+Bench show a significant overlap. However, where DECIPHER is specifically geared at the consortium for which it is the IT platform, Store+Bench has stronger support for the routine diagnostic side. We will discuss a number of features found in Store+Bench and describe them in detail in the following sections, and illustrate them with screen shots and examples when relevant. Rather than to provide a complete overview of the system, this section aims at illustrating the main points of the Store+Bench Framework. Important features are singled out to illustrate the challenges outlined in Section 4.3.1. The global architecture of Store+Bench is built in Java [14], relies heavily on the Spring Application Framework [15], uses MySQL as the underlying relational database management system [16], and relies on the DAS protocol for data exchange [17]. Store+Bench runs within an application server such as Apache Tomcat [18]. A high level overview of the architecture is shown in Figure 4.3.

4.6.1 Main feature 1 - managing Array CGH related information

Store+Bench manages all Array-CGH related information, from raw data over called aberrations to the patient record. It supports a high throughput diagnostics work flow from array scan to final report, as well as ongoing biomedical research projects. It does this as a web-based application which stores, manages, searches, mines, analyzes, reports and visualizes Array CGH-related data. It was developed in context of this PhD and is a joint effort of the bio-informatics group at ESAT and the Constitutional Cytogenetics group at the Center for Human Genetics of the University Hospital. It constitutes of three major parts: Store, Bench, and Toolbox. Store is the assay or experiment centric part of the platform. Bench is patient centric. Finally, Toolbox forms a platform that can be used to plug in a number of specific analysis algorithms written in R (The R Project for Statistical Computing [19], a free software environment for statistical computing and graphics) and using BioConductor [20], an open source and open development software project for the analysis and comprehension of genomic data. It runs on top of R. The main screen of the Bench part of the tool is shown in Figure 4.4. This figure shows that patient records are grouped in projects. Access rights can be granted to individual user accounts on these projects. The interface is rendered in Dutch, French, German and English, depending on the browser configuration. The main report overview listing shows the project a patient belongs to, a number of identifiers, comments, and what type of information the patient is currently annotated

[14] `java.sun.com`
[15] `www.springframework.org`
[16] `www.mysql.com`
[17] `www.biodas.org`
[18] `tomcat.apache.org`
[19] `www.r-project.org`
[20] `www.bioconductor.org`

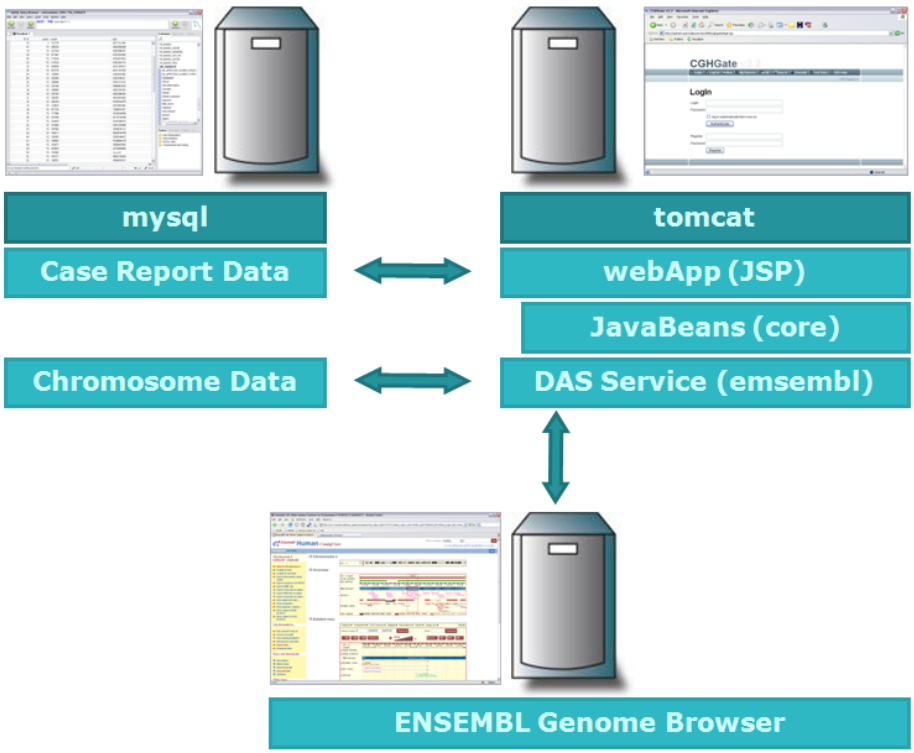


Figure 4.3: The global architecture of Store+Bench is built in Java, relies heavily on the Spring Application Framework, uses MySQL as the underlying relational database management system, and relies on the DAS protocol for data exchange. Store+Bench runs within an application server such as Apache Tomcat.

with. An action menu allows specific operations, such as drawing karyograms.

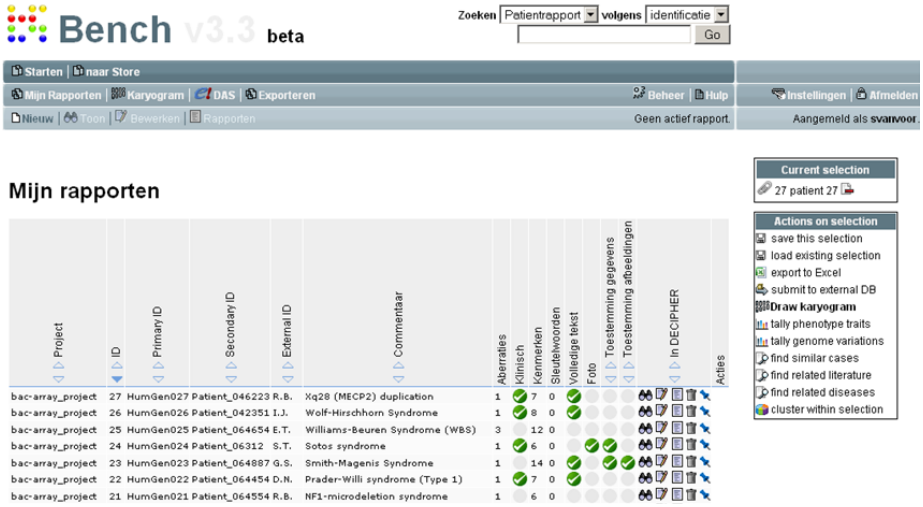


Figure 4.4: This screen shot shows the main screen of Bench. Patient records are grouped in projects. Access rights can be granted to individual user accounts on these projects. The interface is rendered in Dutch, French, German and English, depending on the browser configuration. The main report overview listing shows the project a patient belongs to, a number of identifiers, comments, and what type of information the patient is currently annotated with. An action menu allows specific operations, such as drawing karyograms.

Bench features karyogram drawings for individual case reports, groups or selections of case reports, and for the entire database. In Bench, users can add various annotation information on a patient case report. This varies from general information such as different identifiers and comments, over clinical data such as age and ethnicity to clinical features fully configurable to the needs of the laboratory through an interface that helps configure categories and labels within those categories. Apart from this structured annotation, free text and free keywords can be annotated to a patient record.

While free text annotations are possible, it is recommended to use the phenotype annotation module to describe a patient's phenotypic traits. The phenotype module can be populated with hierarchically organized phenotype taxonomies. Currently, LNDB is supported (see Section 3.3), as well as a coding scheme that describes heart related phenotypes and which is assembled and maintained by the association for European Paediatric Cardiology (AEPC [21]), and the POSSUM classification system [22] that is part of a database of multiple malformations, metabolic, teratogenic, chromosomal and skeletal syndromes and their images, syndrome entries, case commentaries and citations. The database is used for learning and diagnosis by many genetic testing centers and hospitals throughout the world. While currently only 3 classification

[21] `www.aepc.org/aepc/nid/European_Paediatric_Cardiac_Coding`
[22] `www.possum.net.au`

systems are present, the system is extensible and other classification schemes can be loaded, such as ICD10 (the International Statistical Classification of Diseases and Related Health Problems, revision 10, is a coding of diseases and signs, symptoms, abnormal findings, complaints, social circumstances and external causes of injury or diseases, as classified by the World Health Organization) or other ontologies present in UMLS (The Unified Medical Language System is a compendium of many controlled vocabularies in the biomedical sciences. It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems; it may also be viewed as a comprehensive thesaurus and ontology of biomedical concepts).

The phenotype trait annotation page is shown in Figure 4.5. The interface shows two parts. Above, phenotypes from the taxonomies immediately appear when the user starts typing part of their definition. Below, the full hierarchies are shown as trees that can be navigated by expanding nodes. The combination of an auto-completing form and a navigated tree allow a fast annotation of patient phenotypes, minimizing the time needed by geneticists for data entry.

With this system, patient phenotypes can be represented in a consistent way, ensuring easy case report retrieval. Additionally, this phenotype annotation enables data mining for detection of similar patients, and identification clusters of patients with similar phenotypes as putative novel syndromes.

The cytogenetic report view also shows chromosomal aberrations. It places all known aberrations of a patient in context: gene lists can be obtained for individual regions, and annotations and definitions of genes are available. Links to external databases such as Ensembl and OMIM for disease genes are present. All data can be automatically exported to Microsoft Excel.

4.6.2 Main feature 2 - flexible reporting

Bench automates the process of writing technical, clinical, and overview reporting on array and patient data. A flexible and extensible reporting system allows construction of custom report templates. Reports are editable after their were generated and are stored permanently. Reporting can be done in multiple languages and integrates information from external databases (ENSEMBL, disease databases, repositories of known CNV's, relevant biomedical literature, previous sightings of patient genotype and/or phenotype, and more).

The reporting module is based on a template system. Users can create templates for different reports, and decide what information should be present in a report. When a template (upper part of Figure 4.6) is executed for a specific case report, the template tags are replaced with the actual information for that patient (*e.g.* [sex] is replaced by 'Male') and a full report is rendered (lower part), approved and stored. Tags that can be used in the template system include case report identifiers, molecular karyotype (*e.g.*, XY arr cghXq28(RP11-812A23→XX-QLL2C9)x3), the array platform used (*e.g.*, Agilent Oligo aCGH 244A), affected regions, confirmation methods, cases with similar phenotype, cases with similar genotype, overlapping or nearby affected regions, relevant OMIM references, links to OMIM genes and diseases, and many more.

The reporting module is able to generate standard and customizable reports and automates the process of retrieving and gathering information from case reports. This facilitates retrieval of relevant public domain knowledge, and alleviates users form the error prone process of generating lists and structuring genotype descriptions that follow

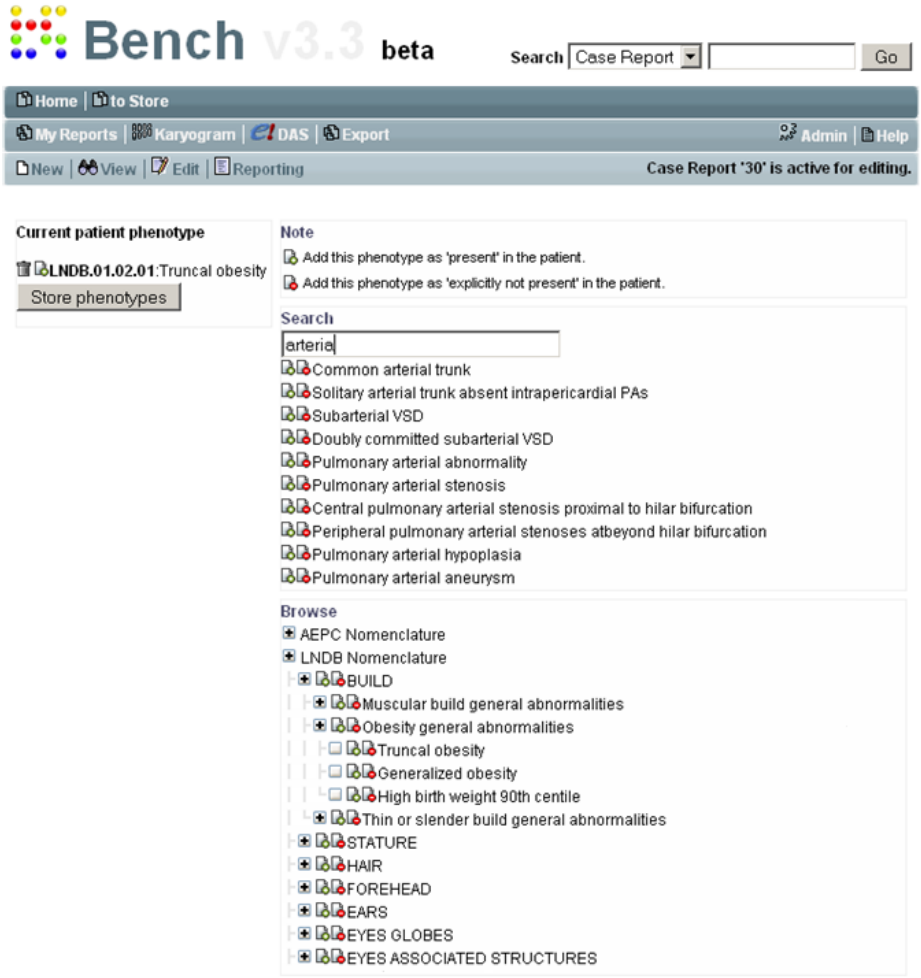


Figure 4.5: The phenotype annotation interface of Bench. The interface shows two parts. Above, phenotypes from the taxonomies immediately appear when the user starts typing part of their definition. Below, the full hierarchies are shown as trees that can be navigated by expanding nodes. The combination of an auto-completing form and a navigated tree allow a fast annotation of patient phenotypes, minimizing the time needed by geneticists for data entry.

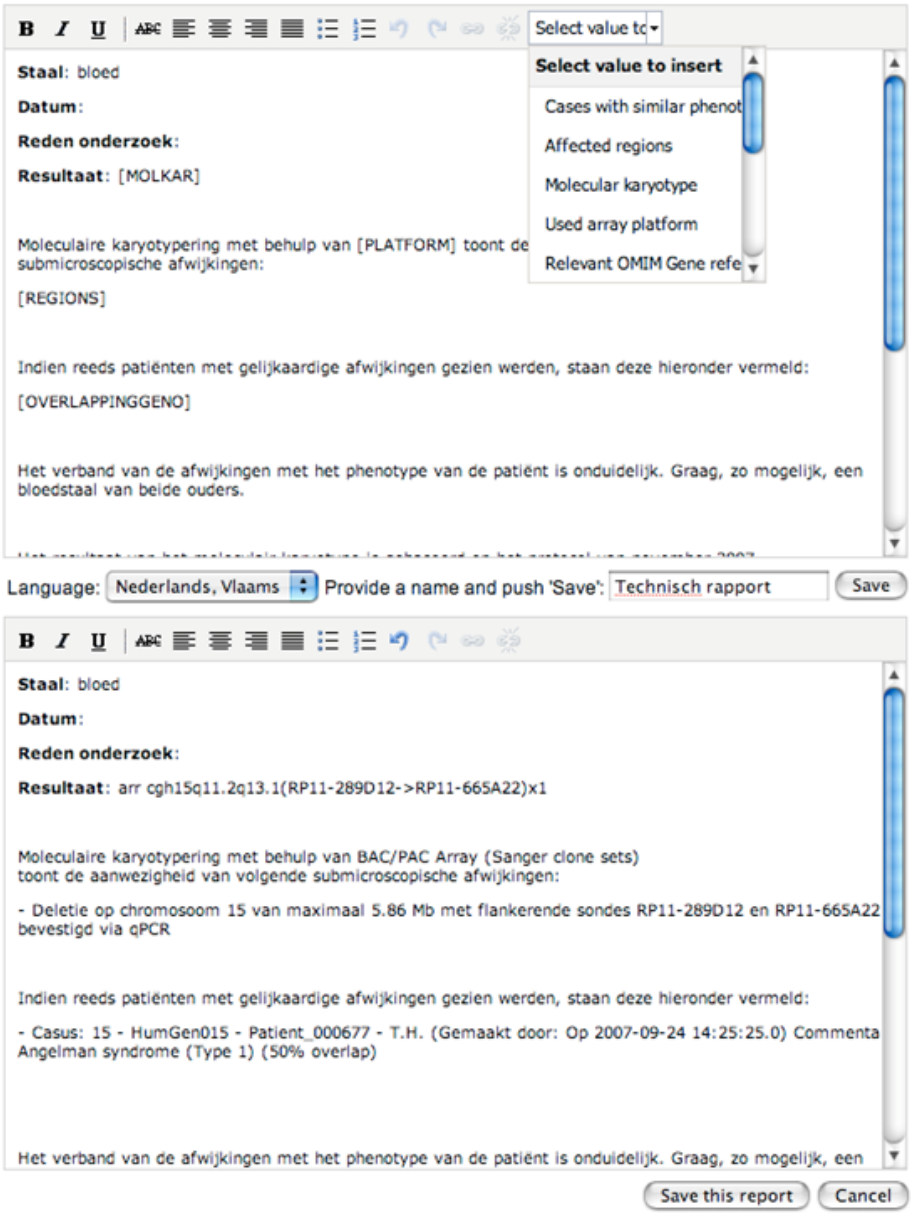


Figure 4.6: Bench reporting module. The reporting module is based on a template system. Users can create templates for different reports, and decide what information should be present in a report. When a template (above) is executed for a specific case report, the template tags are filled out and a full report is rendered (below), approved and stored.

the ISCN nomenclature (an International System for Human Cytogenetic Nomenclature [216]).

4.6.3 Main feature 3 - a data store

Store is the part of the platform where all array related data in the form of files can be stored safely, robustly, and for a long time. Store is format agnostic and serves as a repository for files regardless of array platform, file size, or file type (patient pictures, FISH images, raw array data, spot calls, reports from region calling software, text documents, Agilent- Affymetrix-, BlueFuse- and custom software reports...). Store features support for many file types, and automatic information extraction for most.

On top of this data store, based on the output of different parsers that run automatically when files are uploaded into Store, runs a filtering system that allows users to interpret raw assay results and lists of unfiltered genomic aberrations that stem from external region calling tools present in platform vendor array analysis packages or open source algorithms. Filters include a chromosome filter, a filter that selects aberrations based on length, a known CNV filter based on the DGV data set discussed earlier, a gene content filter, a known syndromes filter, a trio filter for recognizing inheritance patterns (inherited vs. *de novo* aberrations), a loop design filter, a manual review filter and a Gene Prioritization / Enrichment filter.

These filters all rely on the automated parsing of aberrant region call reports. Many file types are supported: Affymetrix Copy number CNT files, Agilent Interval-based region reports and clone-based reports, BlueGnome BlueFuse region reports, ArrayCGHBase extracts, and QuantiSNP reports based on Illumina bead assays. More parsers will be released as formats and array platforms become available.

4.6.4 Main feature 4 - integration and interaction with external resources

Throughout Store and Bench, links are available to and from other tools and databases, such as PUBMED, OMIM, CNV sets, Ensembl, and many more. One important feature is the use of the Distributed Annotation System for making patient information available in genome browsers such as the Ensembl web site (`www.ensembl.org/`). The technical details and advantages of this approach as well as a number of applications will be discussed in Chapter 5. In Figure 4.7, a screen shot is shown where Bench patients are visible as an Ensembl genome browser track.

4.6.5 Main feature 5 - gene prioritization

As explained in Section 3.5, the gene to phenotype correlation pipeline aGeneApart is not only available as a stand-alone tool but is also used by both DECIPHER and Bench. While the process of calculating correlations between genes and human dysmorphy phenotype traits is the same, the interfaces differ between the web layer on top of aGeneApart, DECIPHER, and Bench. While DECIPHER merely provides a prioritization for the full patient phenotype or for individual traits, the Bench interface is more flexible. From within Bench, the user can choose to prioritize the genes in aberrant regions either by the full phenotype, by single phenotype traits, or by a user defined sub-selection of phenotype traits. Secondly, the user has control over the minimum number of PUBMED abstracts that should be available in order for a link between a gene and a phenotype trait to be taken into account. Thirdly, Bench will

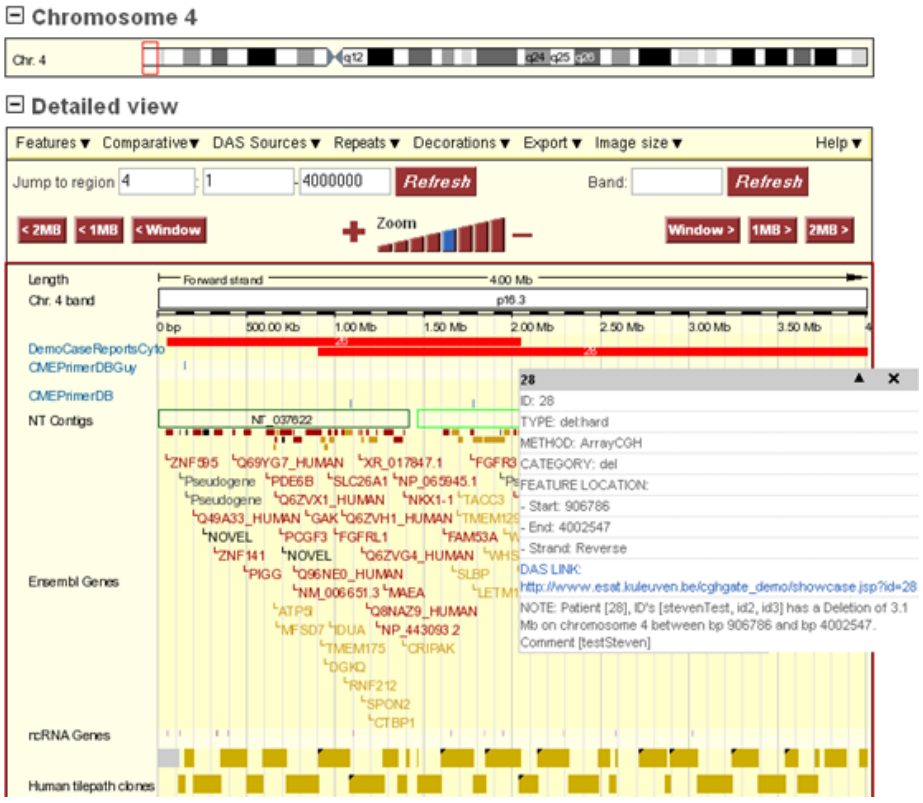


Figure 4.7: Bench DAS integration. This screen shot shows how Bench patients are visible as an Ensembl genome browser track.

allow a full genome prioritization, hence not limiting the analysis to the genes that lie in the regions that are aberrant in the patient. Fourthly, linking out to the actual literature is tightly integrated with the Entrez PUBMED web interface [23]. Fifth, additional information is available through the interface, such as the mapped location of the gene within the aberration.

The Bench gene prioritization interface is shown in Figure 4.8. In this figure, it is shown how a sub-selection of a patient phenotype of 3 out of 7 traits is made. The minimal number of abstracts for a link to be taken into account is set to 2, the default value. The top ten of genes in regions affected in the patient is shown. The first column shows the score, calculated in the same way as explained in Section 4.5.5. The second column shows the ENSEMBL identifier. The third column shows the exact position of the gene. This is useful when prioritizing genes for patients that have multiple aberrations. The fourth column shows gene descriptions and the last column provides a link to the actual abstracts that corroborate the gene to phenotype association (the number of abstracts is shown for each gene).

For a full discussion of the pipeline, information on how it was validated, and a case study, see Section 3.5.

4.6.6 Key applications of Store+Bench

As a conclusion on Store+Bench, we review the different needs this platform addresses.

A lab experiment management system

Bench is a Lab Experiment Management System for Array CGH in constitutional cytogenetics, supporting both diagnostics and research at large volumes of patient throughput. It features an intelligent database that is continuously updated in features and content, and serves as a robust, production level database tool.

A storage tool

Store+Bench contains useful tools for storage of Array CGH experiment data, whether it concerns results from in-house or commercial BAC/PAC assays based on Sanger clone sets, the BlueFuse CytoChip, oligo assays such as the Agilent platform, or Affymetrix and even Illumina assays. It provides storage of patient data, clinical and phenotype data, and every kind of raw data file, whether scans, log ratio files, reports, images, PDFs or other document formats.

A reporting tool

Clinical reports can be generated through a template system. Reports are printable and can be stored, the template system is extensible and includes visualizations such as karyograms. The tool provides genome browsing through ENSEMBL via DAS hence supports genome annotations. It contains many external data sources for feeding information on genes, function, known CNVs, patient information, diseases, and literature.

[23] `www.ncbi.nlm.nih.gov/pubmed/`

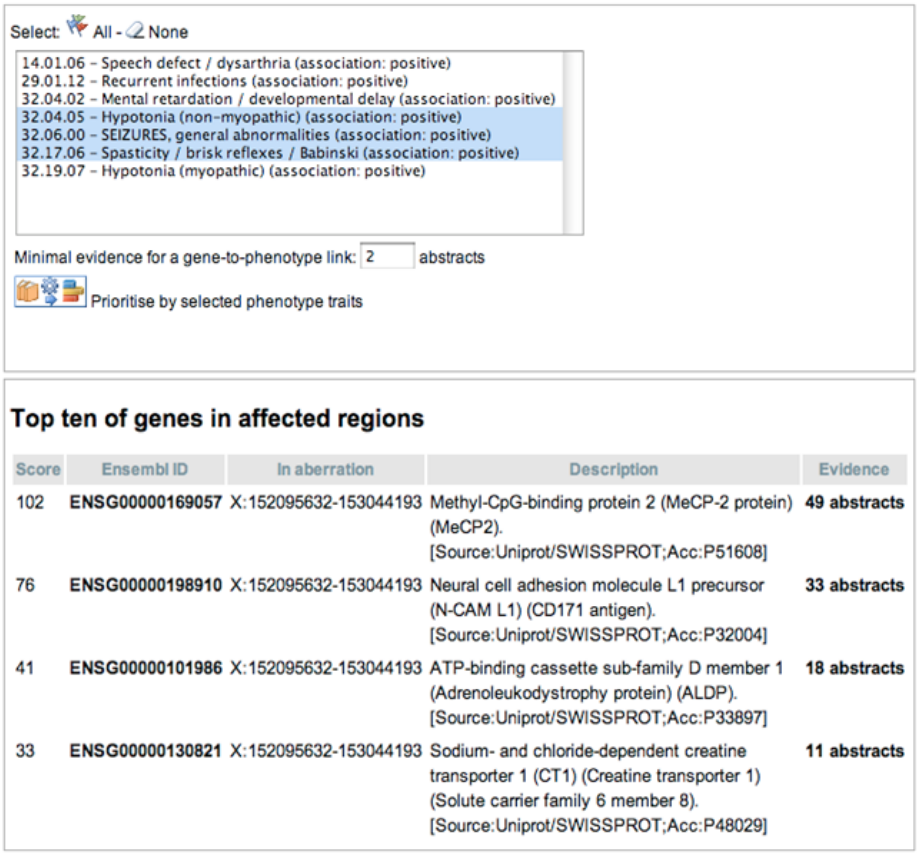


Figure 4.8: The Bench gene prioritization interface. On this screen shot, it is shown how a sub-selection of a patient phenotype of 3 out of 7 traits is made. The minimal number of abstracts for a link to be taken into account is set to 2, the default value. The top ten of genes in regions affected in the patient is shown. The first column shows the score, calculated in the same way as explained in Section 4.5.5. The second column shows the ENSEMBL identifier. The third column shows the exact position of the gene. This is useful when prioritizing genes for patients that have multiple aberrations. The fourth column shows gene descriptions and the last column provides a link to the actual abstracts that corroborate the gene to phenotype association (the number of abstracts is shown for each gene).

A lab collaboration tool

Multiple centers in a clinical or research collaboration can use Store+Bench to work together in a consortium. Within a lab, different types of users can collaborate and each take charge of part of the diagnostic work flow, whether as a lab technician, cytogeneticist, paediatrician or genetic counselor. The database provides security through authentication and access control at different levels (read, read/write, and administrative), allows managing projects and project access rights, and groups case reports as well as experiments into projects.

A data safe

Store relies on multiple servers and all data is backed up regularly. Secure access is provided through encryption protocols while the tool is accessible from anywhere over the web - from conferences, from home, from collaborating centers, or from the own lab. The tool offers full storage of experiment data as well as any annotations on experiments and patients.

An intelligent database

Bench offers various data mining capabilities, such as detection of similar patients by genotype or by phenotype, the identification of relevant literature through gene prioritization, automatic inference between aberrations and phenotypes, extended search, easy access to the growing body of patients in the repository, automated annotation through providing gene lists, linking to common CNVs and literature. It is platform aware as it can automatically import aberration calls from a host of analysis software packages, and contains an accounting module for clustering patients by genotype or phenotype and finding over-represented regions or phenotypes, a first step towards the detection of new syndromes.

A flexible data processing tool

The toolbox evolves around an R Bridge that will allow users to incorporate existing, well-known and proven algorithms, and algorithms of their own that were written in R using BioConductor. Hence, the R Bridge can be used as a test bed for new algorithms, or for fine tuning a lab data processing work flow. Like Bench and Store themselves, the R Bridge will allow input from all popular data processing tools, different platform or vendor specific analysis software, and links to external tools for output. A module that will allow automatic transfer of patient records to DECIPHER is under development.

4.7 Conclusions

The application of array-CGH to patients with developmental delay, learning disability, dysmorphic features and/or congenital anomalies is revolutionizing clinical genetic practice. Tools like DECIPHER and Store+Bench have the aim of being key tools for the clinician in harnessing this new high resolution cytogenetic data to the benefit of patients and their families. Tools like DECIPHER has already been instrumental in defining new syndromes and platforms like these will become more powerful as the number of records in the databases grows.

Chapter 5

Genome annotation

"The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them."

William Lawrence Bragg (1862 - 1942)

The completion of the human genome sequence [131, 256] is a landmark of historical importance, and gave scientists a first draft of the complete genome sequence of the Homo Sapiens organism. The determination of this sequence has been a daunting task, taking a substantial amount of funding and years of work. Determining the sequence of an organism, however, only marks the beginning of understanding the complexity of the genetic makeup of life. Finding meaning in sequence by elucidating the biological processes and physical functions it entails, is a daunting task in itself. There are regions that are difficult to sequence because of repetitive DNA. Even when a sequence is stable and as good as complete, there will still be errors and omissions because DNA sequencing is not a completely accurate process.

5.1 Determining sequence, and annotating it

Hence, charting a genome is more than determining its sequence. Starting from sequence, genomic features need to be found. Many features of a sequence, like the location of genes, proteins, retrotransposons, mRNA's, common copy number variations, CG-content, repeats, conserved sequence, orthologic and paralogic sequence, ORFs (open reading frames), coding regions, regulatory elements such as motifs, ... have been known for a long time, but with the advent of detailed sequence information, could finally be mapped onto the string of base pairs that forms a genome.

This task is not straightforward. Features can be missed. Features can be confused or misinterpreted. Finding features in a genomic sequence depends on the knowledge, expertise and also the interest of research labs and scientists. Also, a true reference genome is an illusion, a final genetic sequence of an organism is a moving target. Parts of the genome are sequenced and resequenced. Assembling a genome from a huge number of sequenced fragments through use of large scale IT platforms is an error prone process. Even now, eight years after the release of the first draft of the human genome [131, 256], periodical releases of updated sequences are very common. Even now, Ensembl (`www.ensembl.org`) is continuously refining and extending its gene

building process, calibrating it against regions of the genome that have been hand annotated and experimentally investigated, for a large number of genomes, including that of man.

We have established that obtaining a reference sequence is an ongoing effort and that the results are made available are in no way final. Also, annotating the genome is a continuous process. Annotations need updating to reflect the new information gained through rapid accumulation of genomic sequences. In the narrow sense, genome annotation is the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do. An annotation can be considered as note or commentary that is added on top of sequence information as an explanation. Once a genome is sequenced, it needs to be annotated to make sense of it. These commentaries are usually called 'features'.

In short, genome annotation is the process of attaching biological information to sequences. An important step in genome annotation is identifying candidate features. This task entails characterizing features, determining feature boundaries, classification of features, sorting them out in families (*e.g.* for proteins as features), finding related or consecutive features, and so on.

In the broad sense, genome annotation is linking knowledge on a sequence to the location, and hence, making a link between the world of logic, and the physical world. This knowledge can be very diverse. It can be information on expression of genes in tissues, gene and protein information, information on pathways, biochemical and biological function, involvement in high level processes such as disease, interactions between different entities, annotations regarding regulation, and so on. Annotation is the result of research and of thought processes. Especially complex annotations are done manually. In large scale annotation efforts, this process is supported by computers. Genome annotation is an important subfield in the realm of bio-informatics: from gene finding in newly sequenced organisms, to investigating synergies between and among new and existing annotations, within and between different species, and recently also across different individuals.

As mentioned above, determining a reference sequence is crucial, it is the scaffold along which annotations are made. With the discovery of inter-individual variability, and the advent of large scale projects that work towards determining the sequence of thousands of normal individuals, the complexity of determining a reference sequence is exponential. Another fact that needs to be taken into account is that initial sequencing efforts were based either on a single individual or a pool of a limited number of individuals, in order to obtain what is called a reference sequence.

5.2 A reference sequence for many genomes

An international research consortium has been formed to create the most detailed and medically useful picture to date of human genetic variation. The 1000 Genomes Project will involve sequencing the genomes of at least a thousand people from around the world. Drawing on the expertise of multidisciplinary research teams, the 1000 Genomes Project will develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a resolution unmatched by current resources. As with other major human genome reference projects, data from the 1000 Genomes Project will be made swiftly available to the worldwide scientific community through freely accessible public databases.

How will the output of the 1000 genomes project (`www.1000genomes.org`) be

represented? A linear string of G, A, C and T that is downloadable and can be stored in a flat file will soon no longer suffice: new machine and mathematical representations and ways to visualize genomes will have to emerge to be able to continue to think about and work with sequence information. Representations of genomes will have to be able to capture the structure of a genome, which we discover to be more complex with time.

5.3 Current approach to genome annotation

In this chapter, we will discuss genome annotation on two levels. Firstly, we will look at technical aspects of genome annotation. Genome annotation is not just about mapping knowledge to sequence, but also about knowledge distribution, information merging, and providing means of visualization that ease interpretation.

Secondly, we discuss an actual case study of generation of genome annotation. Typically, genome annotations result from research efforts that often involve significant manual curation or investigation, lab work, experiments, and thought. This process is complemented and supported by bio-informatics and statistics. We will discuss a specific example of annotating *cis*-regulatory elements of the genome based on a combination of manual curation and a text mining pipeline.

5.4 Genome annotation: technical aspects

DAS or Distributed Annotation System [54] is a protocol that allows integration of annotation efforts from various sources and facilitates visualization of annotations interactively in a browser (in the broad sense: genome browsers are not necessarily web based or ran in web browsers). It serves as a means for decentralization of annotation efforts and is used to share genome annotations with a community, whether publicly or closed.

5.4.1 What is DAS?

New large scale techniques in biology are producing a rapidly growing amount of publicly available data that are usually stored in centralized database resources and are often subject to frequent updates. This gives rise to challenges related to the exchange of these data with the community. The Distributed Annotation System (DAS) addresses these issues by providing a protocol, based on XML, to openly exchange biological annotations between distributed sites. It is a protocol for integrating biological data [113, 182]. Data distribution, performed by DAS servers (such as ProServer [64], Dazzle (`www.biojava.org/wiki/Dazzle`), the Lightweight Distributed Annotation Server (LDAS, `biodas.org/servers/LDAS.html`, and now also the Ensembl genome browser [106]), is separated from visualization, which is done by DAS clients (such as GBrowse [53], WormBase [237], the Ensembl genome browser [106], ProView (`www.sanger.ac.uk/perl/PostGenomics/proview/`), and Dasty [116]).

This separation of concerns is inherent to the client-server aspect of DAS and allows clients like Ensembl to integrate information from multiple servers. It allows a single machine to gather up genome annotation information from multiple distant web sites, collate the information, and display it to the user in a single view. Little coordination is needed among the various information providers. A big advantage is

that use of the DAS protocol by a genome browser allows many annotations from heterogenous sources to be provided within a single view or to be processed within a single system [105].

DAS is heavily used in the genome bio-informatics community. Over the past years, there has also been a growing acceptance in the protein sequence and structure communities, thanks to an extension of the DAS protocol to allow for annotation of the three-dimensional structure of proteins.

5.4.2 How does DAS work?

DAS reference servers typically provide the structure and sequence of a genome. Annotation servers provide features along the scaffold offered by a reference server. DAS provides a convention on how to encode a DNA or protein sequence and its annotated features into simple XML documents that are exchanged over the Internet. Regions on a chromosome (and also protein sequences) are considered to be the reference objects for which sequence information is provided by reference servers. Annotation servers offer features that are mapped to specific genomic sequences according to a coordinate system imposed on that sequence (such as annotation by gene, by sequence base pair position, by clone, by marker, and so on). Annotation servers provide the actual annotation data that are available for the reference objects. A public registry listing many publicly available DAS servers is available (`www.dasregistry.org`).

The DAS protocol is under active development, and people are defining specifications and setting up implementations for security, write back (using the protocol not just for querying and retrieving but also writing and updating annotations), among other new capabilities.

5.4.3 Distributed annotation: PrimerDB

As a first example to illustrate the basic capabilities of DAS, we discuss PrimerDB, an in house database that has not been published and that serves information on primers (short DNA sequences used to initiate the synthesis of DNA in a polymerase chain reaction) within CME at the University Hospital. This DAS server is based on an instance of ProServer [64] and serves local lab information on primers, such as the location of freezers in the building where they can be found, order status, and contact person. This practical example illustrates how far genome annotation can reach. Typically, the Ensembl genome browser will be used as a client. This means that both lab technicians and researchers at CME can point their genome browsers to a genomic location of interest and see whether primers are available.

In order for this to work, a primer database and BLAST pipeline was set up. BLAST (Basic Local Alignment Search Tool) is an algorithm for comparing sequence information, such as the amino-acid sequences of different proteins or, in this case the nucleotides of DNA or cDNA sequences. Through this pipeline, the actual location where primer sequences map to the genome are retrieved. From the moment the primers have an actual location in respect to the genome coordinate system (in this case: chromosome, start base pair and end base pair), they can be provided as annotations through a DAS annotation server.

The DAS protocol offers the possibility to specify style sheets, so that layout can be performed on the graphical representation of annotations. In this case, arrows are used, because primers are directional (forward and backward). Visualization of a specific region containing a number of CME primers is shown in Figure 5.1.

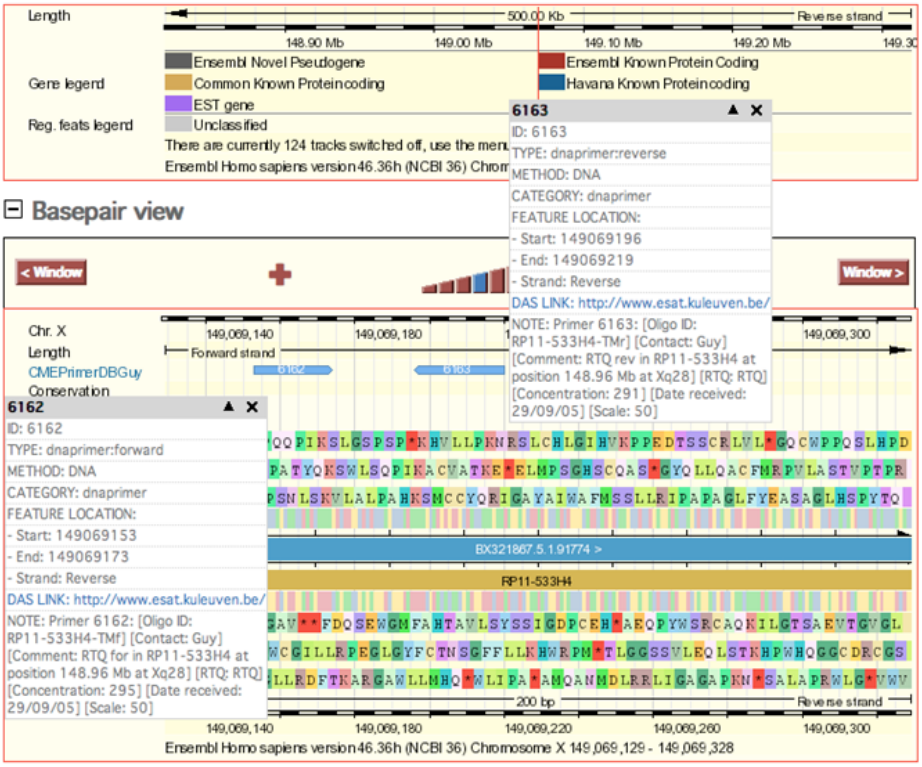


Figure 5.1: The DAS protocol offers the possibility to specify style sheets, so that layout can be performed on the graphical representation of annotations. In this case, arrows are used, because primers are directional (forward and backward). Visualization of a specific region containing a number of CME primers is shown in this figure.

5.4.4 Distributed annotation: Bench and DECIPHER

Both DECIPHER and Store+Bench are cytogenetic platforms that have their case report databases serve patient genotypes as DAS formatted features on the human genome reference sequence. This means that they can be queried by clients through the DAS protocol. Examples are the augmented Wiki (GeneWiki) that is able to incorporate patient information, and, importantly, the Ensembl genome browser, where annotation tracks can be activated. In Image 5.2, both the patients present in the CME instance of Store+Bench as well as the consented patients available within the DECIPHER consortium are shown in their genomic context, together with other annotations such as the Redon CNV loci, genes and contigs, chromosome banding patterns, and human tile path clones.

5.5 Case study: text-mining assisted regulatory annotation

Decoding transcriptional regulatory networks and the genomic *cis*-regulatory logic implemented in their control nodes are fundamental challenges in genome biology. High-throughput computational and experimental analyses of regulatory networks and sequences rely heavily on positive control data from prior small-scale experiments. However, the vast majority of previously discovered regulatory data remains locked in the biomedical literature. In this chapter, we report on the development of text-mining strategies to identify relevant publications and extract sequence information to assist the regulatory annotation process.

5.5.1 Unlocking transcription regulation information in biomedical literature

In what follows, it is demonstrated that document relevance ranking can assist the curation of transcriptional regulatory networks by using a vector space model (introduced in Section 6.1.4) to identify MEDLINE abstracts of papers likely to have high *cis*-regulatory content. Over 30,000 papers are estimated to harbor non-annotated *cis*-regulatory data. In a publication on this work, it is also shown that DNA sequences can be extracted from full-text articles with high *cis*-regulatory content and accurately mapped to genome sequences as a means of identifying the location, organism and target gene information that is critical to the *cis*-regulatory annotation process. The extraction of DNA sequence is out of the scope of this dissertation.

In this chapter, it is illustrated that text-mining technologies can be successfully integrated with genome annotation systems, thereby increasing the availability of annotated *cis*-regulatory data needed to catalyze advances in the field of gene regulation.

5.5.2 Background

The process of annotation is an essential first step in attributing biological information to genome sequences. Traditionally, the main focus of genome annotation has been the identification and annotation of protein-coding genes, RNA genes and repetitive DNA. Efforts to annotate these genomic features typically adopt one of several established annotation paradigms - the "museum," "jamboree," "cottage industry," or "factory" models of genome annotation [236, 58]. Other important functional regions

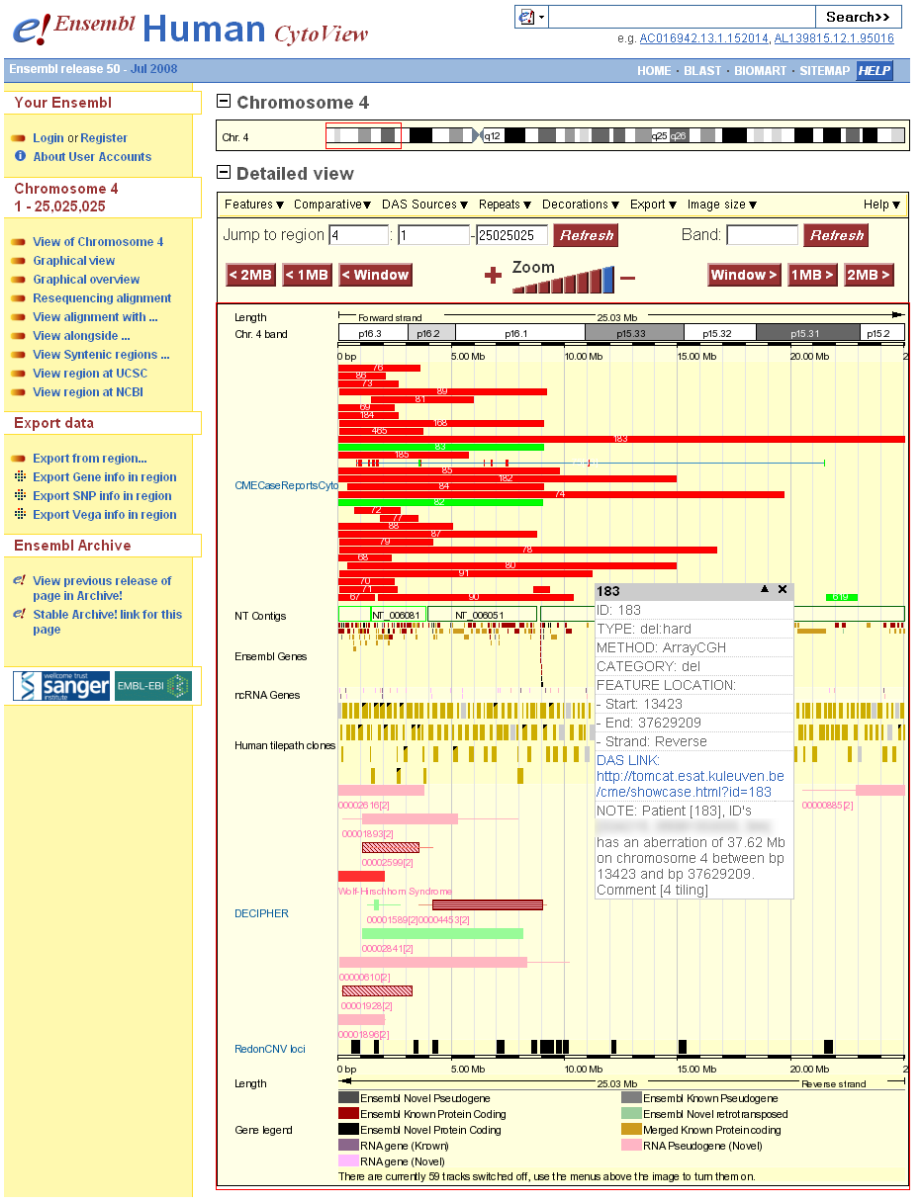


Figure 5.2: Screen shot of the Ensembl Genome browser pointed at the 4p terminal region of chromosome 4. In Ensembl, DAS protocol based genome annotation tracks can be activated. This screen shot shows both the patients present in the CME instance of Store+Bench as the consented patients available within the DECIPHER consortium visualized in human genomic context, together with other annotations such as the Redon CNV loci, genes and contigs, chromosome banding patterns, and human tile path clones.

of genomes that are more difficult to predict by *ab initio* or homology methods are often omitted from the standard genome annotation process, in particular the *cis*-regulatory sequences that control transcription. Instead *cis*-regulatory sequences are typically annotated by manual curation from the literature either under the museum model in the private domain [156] or under a "boutique" model [180] in the academic domain, whereby small teams curate organism- or process-specific data sets from the primary literature for short-term research purposes. Such decentralized resources are disseminated and maintained in *ad hoc* ways that are often not integrated with the major genome database resources, and present a bewildering array of choices to the computational or experimental end-user.

Recently, two efforts have been launched to develop integrated portals for *cis*-regulatory annotation - ORegAnno [163, 89] and PAZAR [180] - that aim to support research in *cis*-regulatory sequence and network analysis. Both ORegAnno and PAZAR provide principled, standardized technologies for the long-term, community-driven, open-access annotation of *cis*-regulatory data in the context of the major genome database resources (*e.g.* NCBI[1], Ensembl[2], UCSC[3]), and as such represent a new generation of resources for the annotation of *cis*-regulatory data. Despite these advances in infrastructure, many challenges still remain for the comprehensive community-based annotation of *cis*-regulatory data. First, as with all decentralized annotation efforts, community annotation of regulatory data from the literature requires systems to track the curation process, including triaging relevant and irrelevant articles and monitoring the curation status of papers. Second, the scale of the *cis*-regulatory annotation challenge remains unknown, and thus it is critical to identify and prioritize the set of documents with high *cis*-regulatory potential for curation. Third, with curation times currently on the order of 1-2 hours per paper, a major bottleneck remains in how to efficiently extract *cis*-regulatory data from primary text. Recently, rule-based information extraction systems have been developed to extract regulatory relations among pairs of genes and proteins [205, 204, 195], however many other data fields are necessary for comprehensive *cis*-regulatory annotation, such as the organism under investigation and, perhaps most importantly, the sequence and genomic location of *cis*-regulatory elements.

We have attempted to solve some of these challenges through the use of a community literature management system and application of text-mining techniques to retrieve and extract relevant documents and data for the annotation of *cis*-regulatory networks and sequences. These efforts were inspired by (and conducted in part through) the RegCreative Jamboree [4], a workshop that was held in late 2006 which attempted to explore the interface between regulatory bio-informatics and text-mining communities. As detailed elsewhere [89], a literature management system was recently developed for the regulatory annotation community, which warehouses the set of papers that are likely to contain *cis*-regulatory data and maintains information on their current curation status. This section reports on the development of a vector space model to identify MEDLINE abstracts of papers that are likely to have high *cis*-regulatory content, and how to use this model to demonstrate that document relevance ranking can assist the annotation of transcriptional regulatory networks and be used to estimate the scale of the regulatory curation challenge.

[1] `www.ncbi.nlm.nih.gov`
[2] `www.ensembl.org`
[3] `genome.ucsc.edu`
[4] The RegCreative Jamboree: `www.dmbr.ugent.be/bioit/contents/regcreative/`

In a complementary and simultaneous effort, a pipeline is set up for DNA sequences to be extracted from full-text articles and mapped to genome sequences as a means to identify the location, organism and target gene information that is critical to the *cis*-regulatory annotation process. A discussion of this effort is out of the scope of this dissertation. Collectively, both results demonstrate the utility (and the necessity) of employing text-mining approaches to accelerate the community-driven annotation of *cis*-regulatory sequences and networks that control transcription.

5.5.3 A literature management system for community annotation and text mining

Assembling the set of documents that are relevant for annotation and tracking the curatorial status of papers are major challenges in community annotation. To help overcome these issues, a literature management "queue" was developed for the ORegAnno database, which allows registered users to input papers with known or suspected *cis*-regulatory content as targets for curation using their PubMed identifiers (PMIDs). A full description of the ORegAnno Publication Queue and its features is detailed elsewhere [89]; here, we briefly describe its contents to aid interpretation of our text-mining results. The Queue was initially populated with expert entries obtained from the set of papers in ORegAnno plus existing sources of curated publications including the Drosophila DNase I Footprint Database [22], REDfly [75], a catalogue of regulatory elements for muscle-specific regulation of transcription [265, 99], ABS [28], TRED [277], ooTFD [83] and DBTGR [226]. Additionally, a large number of papers were added manually by individual ORegAnno users from literature searches and review articles. In the current work, we show how in addition to offering a powerful literature management system for community annotation, the ORegAnno Publication Queue offers a rich source of PMIDs for assessing information retrieval and information extraction techniques applied to biomedical text in the *cis*-regulatory domain.

5.5.4 A vector space model identifies MEDLINE abstracts with high cis-regulatory content

As a first step in employing text-mining to aid *cis*-regulatory annotation, we attempted to identify a set of full-text papers that could enter the curation process by using information retrieval technology. To do this, we implemented a vector space model [86] (introduced in Section 6.1.4) that scores the 16 million scientific abstracts from MEDLINE, each represented as a vector of index terms, against a model trained on a corpus of abstracts which *a priori* are known to have high *cis*-regulatory content. For initial model training purposes, 3626 abstracts retrieved with the PubMed query "transcription and regulation and "binding site" and (promoter or enhancer)" (see Section 5.5.4 for details) were first split into two equal parts that form a training set (*POS1*) and a validation set (*POS2*). *POS1* contains 3344 terms after stemming and stop-word removal, representing vocabulary *VOC1*. We compared ten different relevance rankings with *POS1* as query and either the complete *VOC1* or different subsets of *VOC1* as vocabulary. A vocabulary consisting of the 1000 terms with the highest frequency in the full corpus yielded the highest performance when applied to *POS2*. Similar results were obtained using a training set of 6306 abstracts from papers previously curated in ORegAnno [163], TRANSFAC [156], or FlyReg [22]. Thus we chose to develop our relevance ranking based on our "*cis*-regulatory" PubMed query to avoid biases

towards data type, species, or other unknown factors. This approach has the additional advantage that existing sets of curated papers can legitimately be used later as validation sets. To generate the final relevance ranking of MEDLINE used in further analyses we used a model based on the 1000 terms (from the 3626 training abstracts) with the highest corpus frequency as vocabulary. Figure 5.3 shows the distribution of the final similarity scores for all 16 million abstracts in MEDLINE, with an indication of the top 10K, top 50K and top 100K highest scoring abstracts in the distribution.

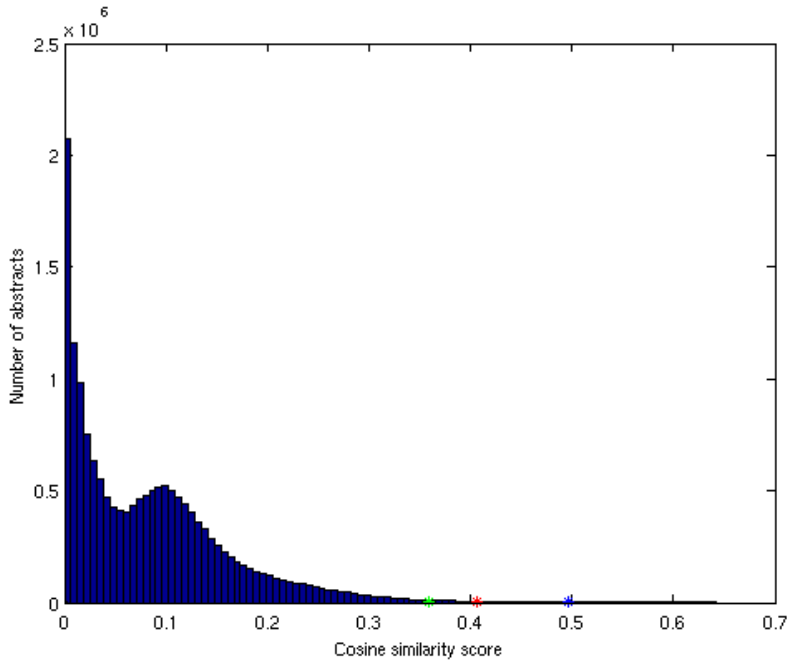


Figure 5.3: Distribution of cosine similarity scores between the query vector and each of the PubMed abstract vectors, indicating the $10,000^{th}$ (blue asterisk) $50,000^{th}$ (red) and $100,000^{th}$ (green) ranked abstract.

Implementation of a vector space model to identify MEDLINE abstracts with high cis-regulatory content

To identify papers with potential *cis*-regulatory data for community annotation, we used a vector space model that represents each of the 16 million scientific abstracts in MEDLINE as a vector of index terms. Each vector element is a weight that is proportional to the relative importance of the term in the abstract (using the inverse document frequency or IDF). Relevance ranking of the corpus is then achieved by calculating the similarity between each abstract and a query. This query can be represented by the same kind of vector as the documents, so that the similarities can be calculated by the cosine similarity measure between individual abstract vectors and

the composite query vector. In practice, a good query vector can be constructed from the average properties of a training set of true positive abstracts. In this study, we used a "*cis*-regulatory" PubMed query that yielded a very high amount of true positives to generate our training set, namely: *[transcription AND regulation AND "binding site" AND (promoter OR enhancer)]*.

Pseudo-curation of full-text articles

To evaluate the ability of our model to predict papers with high *cis*-regulatory content, we selected 344 papers from the top 100,000 scoring abstracts, of which 200 are uniformly distributed and 144 are related to the Drosophila transcription factor eve. Because the full curation of all 344 papers would require the organization of a second annotation jamboree, we opted for a distributed "pseudo-curation" procedure. Particularly, nine experienced curators examined whether these papers describe experimentally verified regulatory data, and if so, whether they also contain all the required data to allow genome annotation (*i.e.*, at minimum the species, the sequence and its genomic location, the transcription factor, and the target gene). A web application was created where the curators could open a pending PMID and score the full-text paper as success or failure. Failures could be of four types: (1) publication describes binding site or promoter but there is insufficient information to annotate it; (2) publication describes transcription factor (complex) but not a binding site or promoter; (3) publication describes consensus binding sites or a reference to a primary publication but is itself not the correct source for annotation; (4) publication does not describe a regulatory element. Regulatory interactions in the form of TF→TG were recorded as free text.

Support Vector Machine classification

Using a similarity-based ranking rather than a classification procedure is particularly useful for our task because it requires no negative training set and the similarity score allows a prioritization of documents for curation rather than a binary decision. To evaluate whether Support Vector Machines (SVM) are in agreement with our ranking, we classified the entire PubMed collection using the SVMlight package (`svmlight.joachims.org`). An SVM was trained using the same set of 3628 papers from our initial PubMed query as positives, and 3628 randomly selected PubMed abstracts as negatives. With the radial basis function kernel, we find that 169402 PubMed abstracts are classified as positive (1.07% of all PubMed abstracts). 95.57% of the top 100K abstracts identified by our cosine similarity method are called positive by the SVM approach. Cosine similarity values and SVM decision function values are furthermore highly correlated (Pearson correlation coefficient is 0.88). 78.42% of abstracts are shared by the top 100K when ranked by their cosine or SVM scores. Therefore, the cosine similarity and SVM methods both point to a very large but similar set of abstracts in PubMed as having high *cis*-regulatory potential.

5.5.5 Coverage

The coverage of several validation sets within the similarity-based ranking is shown in Table 5.1. Before calculating the sensitivity (recall) for each validation set, we removed all PubMed abstracts from these sets that were also part of the training set. As a first validation set we used TRANSFAC [156], a commercial database of manually

curated transcription factor binding sites (TFBSs). We collected all 5719 PMIDs from TRANSFAC (v10.4) that are linked to a curated TFBS. Of the set of 5183 independent TRANSFAC PMIDs (536 were part of the training set), 75.4% are found within the top 50K and 88.2% within the top 100K abstracts. This shows that our model is able to generalize and recover many true positive abstracts with high *cis*-regulatory content. In fact, the vector space model realizes an increase in the proportion of TRANSFAC PMIDs from 14.7% in the 3626 papers based on the initial PubMed query to 18.8% in the top 3626 publications after relevance ranking. Likewise, using a second validation set of 186 independent positive PMIDs from the FlyReg database of curated TFBSs in Drosophila, we find high sensitivities of 78.5% and 89.2% of FlyReg PMIDs in the top 50K and top 100K scoring abstracts in MEDLINE, respectively.

	TRANSFAC	**FlyReg**	**ORegAnno Queue**	**ORegAnno⋆**	**RegCreative Success**	**RegCreative Failure**
# PMIDs	5719	200	4145	376	260	218
# PMIDs (no training data)	5183	186	3687	340	228	212
# in top 10K	1390	38	1035	89	59	18
% in top 10K	26.8%	20.4%	28.1%	26.2%	25.9%	8.5%
# in top 50K	3908	146	2753	260	165	79
% in top 50K	75.4%	78.5%	74.7%	76.5%	72.4%	37.3%
# in top 100K	4572	166	3208	301	199	110
% in top 100K	88.2%	89.2%	87.0%	88.5%	87.3%	51.9%

Table 5.1: Coverage of validation sets (excluding PMIDs in the training set) within the top 10K, top 50K, and top 100K ranked abstracts for the vector space model relevance ranking. ⋆: ORegAnno prior to RegCreative

Next, we investigated the coverage of true positive abstracts using curated papers from the ORegAnno database [163], including those curated as a part of the RegCreative Jamboree[5]. Prior to the Publication Queue, ORegAnno contained 376 curated papers, of which 340 are not part of the training set in the vector space model. Of these, 88.5% ($n = 301$) are covered in the top 100K. Since the creation of the Publication Queue, curated papers are flagged with "failure" or "success", depending on whether they contained enough data to allow the creation of a full ORegAnno record (*i.e.*, either a regulatory region or a TFBS with all required fields, see above). Surprisingly, in a set of 478 papers from the ORegAnno Publication Queue (see above) that were known *a priori* to have a high likelihood of containing *cis*-regulatory data amenable to curation, only 54.4% ($n = 260$) were confirmed as "success" papers during the RegCreative Jamboree. The remaining 218 "failure" papers contained either no

[5]The RegCreative Jamboree: `www.dmbr.ugent.be/bioit/contents/regcreative/`

regulatory data, or one or more critical data fields were missing (*e.g.* the regulatory sequence could not be identified or unambiguously mapped to a target gene). Excluding training abstracts, 87.3% ($n = 199$) of the success papers are found in the top 100K but only 51.9% ($n = 110$) of the failure papers are found in the top 100K, indicating that our relevance ranking increases the likelihood that a paper has *cis*-regulatory data amenable to curation. Collectively, these experiments show that our vector space model successfully identifies and ranks papers with enriched *cis*-regulatory content based on MEDLINE abstracts, and that information retrieval techniques can be used to populate a larger ORegAnno Publication Queue to assist the community annotation of *cis*-regulatory data.

5.5.6 Estimating the size of the cis-regulatory corpus

Although the sensitivities of our vector space model on evaluation sets are high, the calculations were performed on large sets of PMIDs (10K, 50K or 100K), meaning that the majority of candidate papers do not fall into any of the existing sets of curated papers. To investigate the degree to which the additional predictions show high true positive rates, we conducted a validation experiment that also gives us an indication of the scale of the *cis*-regulatory annotation challenge. We constructed a sample of 200 PMIDs evenly spaced every 500 abstracts across the top 100K abstracts. Full-text papers for these 200 samples were subjected to a "pseudo-curation" procedure in which the paper was read by an expert and, instead of being fully curated, was only scored with respect to its "amenability to curation" for containing a TFBS (see Methods). This experiment allowed us to estimate how the proportion of true positives and false positives vary as a function of position in the ranked list of the top 100K scoring MEDLINE abstracts. Figure 5.4 shows the positive predictive value (PPV) for each threshold of the top 100K. The first 10 samples were all success papers, indicating that the top scoring 4501 papers are extremely likely to contain *cis*-regulatory data amenable for curation. From then onwards, the PPV starts to decrease but still remains above 30% for the entire top 100K scoring abstracts. This curve can be used to determine an optimal threshold for including papers in the ranked MEDLINE list into the ORegAnno Publication Queue. As noted above, the proportion of "success" papers from the expert-entry ORegAnno Queue was 54.4%. To achieve a similar curation success rate in the set of papers identified by the vector space model (namely PPV 50%), we would include the top 58.000 scoring abstracts. Therefore, we estimate that the scale of the full corpus with *cis*-regulatory data amenable to curation in MEDLINE is on the order of 30,000 papers. We note that this is a conservative measure because the success criteria are strict. Indeed, among the failure papers there are many that contain regulatory data or references to other potential success papers (see Figure 5.5). Based on these results, we added PMIDs and ranks for the top 58.000 scoring papers in MEDLINE as text-mining entries to the ORegAnno Publication Queue.

The positive predictive value (PPV or precision rate) is the proportion of documents with a positive test outcome that are correctly labeled so. In other words, it is the ratio of true positives over all positives. It is an important measure as it reflects the probability that a positive test reflects the underlying statistical condition being tested for.

$$PPV = \frac{\mathrm{TP}}{TP + FP} \tag{5.1}$$

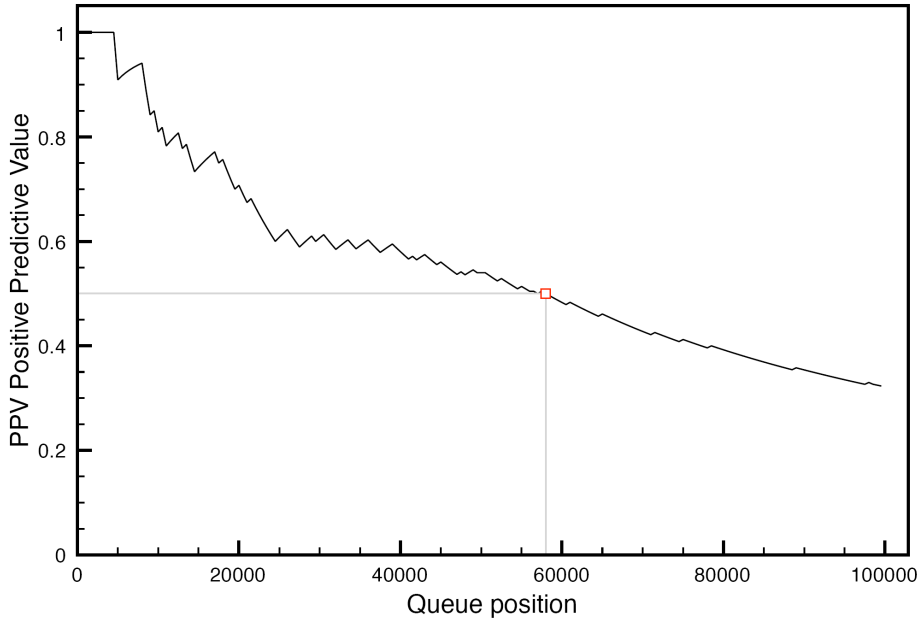


Figure 5.4: Positive predictive value (PPV) calculated for each threshold in the top100k of the final relevance ranking, using the pseudo-curation results of 200 evenly distributed samples. The length of the final queue is chosen at 58,000, to yield a PPV of 50%.

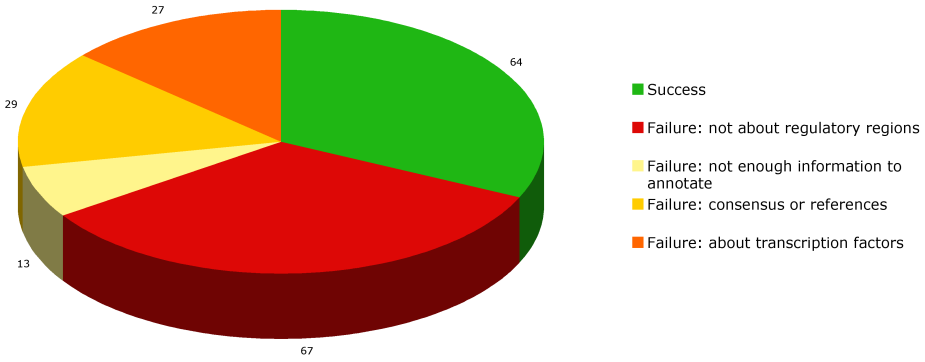


Figure 5.5: Results of the pseudo-curation procedure on 200 evenly distributed samples along the top 100K.

5.5.7 Abstract relevance ranking aids the construction of regulatory networks

To illustrate the utility of identifying papers with high *cis*-regulatory content, we queried the top 58K scoring abstracts for a particular transcription factor (TF), namely the Drosophila homeodomain-containing gene even-skipped (eve). Our goal was to use the set of papers enriched for *cis*-regulatory content to construct a literature-based transcriptional regulatory network focused on the upstream regulating factors and downstream target genes (TG) of eve, based on high-quality published TFBS data. For this experiment we started with the entire list of 664 references associated with eve in FlyBase [49], which also includes papers not related to *cis*-regulatory data (*e.g.* genetic interactions). We cross-referenced this list of all papers on eve with the top 58K list to filter for papers on eve that are likely to contain *cis*-regulatory data. Of the 664 eve papers, 88 are found in the top 58K list (147 are in the top 100K), and for 85 of those (144 for the top 100K) we retrieved the PDF full paper. We conducted a pseudo-curation analysis on these 85 papers to identify those that reported binary TF→TG relations. We classified 35 out of these 85 candidates as "success" papers, which revealed 43 unique binary TF→TG relationships (there were 47 relationships in total, including 4 relations that occurred twice), 20 of which involved eve either as TF or as TG. A summary of the identified regulatory interactions is presented in Figure 5.6 as a network constructed using Cytoscape [218]. By comparison with previously curated binary TF→TG relations for eve in the FlyReg database [22], our automated document retrieval process recovered 100% (12 of 12) of known upstream activating TFs, and 85% (6 of 7) of known downstream TGs. The only downstream TG curated in FlyReg that was missing in this analysis was Abdominal-A (Abd-A), which was omitted because it was not present in the original list of eve-related papers curated by FlyBase. These results show that cross-referencing general PMID lists for a given gene against our vector space model can enrich for papers that report direct *cis*-regulatory interactions for that gene, how transcriptional regulatory networks can be assembled from text-extracted binary TF→TG relationships [205, 204, 195, 10], and that TF→TG interactions may be extracted from text even when full curation of *cis*-regulatory sequences may not be possible.

5.5.8 Text Mining for cis-regulatory information: discussion

A principle aim of genome biology is to decode complete transcriptional networks, so as to better understand how the activation of specific subnetworks affect developmental processes or responses to the environment, and how variation in transcriptional networks can lead to functional diversity over evolutionary time. As with all grand challenges in interpreting genome sequences, solving this ultimate aim will require combining both computational and experimental approaches. As the reliability of predictive regulatory sequence bio-informatics is currently low [244], high-throughput experimental techniques currently prove to be the most efficient means of identifying regulatory sequences and assembling regulatory networks [94, 134]. The gold standard for evaluating both computational and high-throughput experimental techniques continues to be the sizable body of prior knowledge contained in small-scale experimental studies on *cis*-regulatory sequences, much of which remains locked in the biomedical literature. Here we have shown that application of text-mining technologies, including literature management, information retrieval and information extraction systems, can

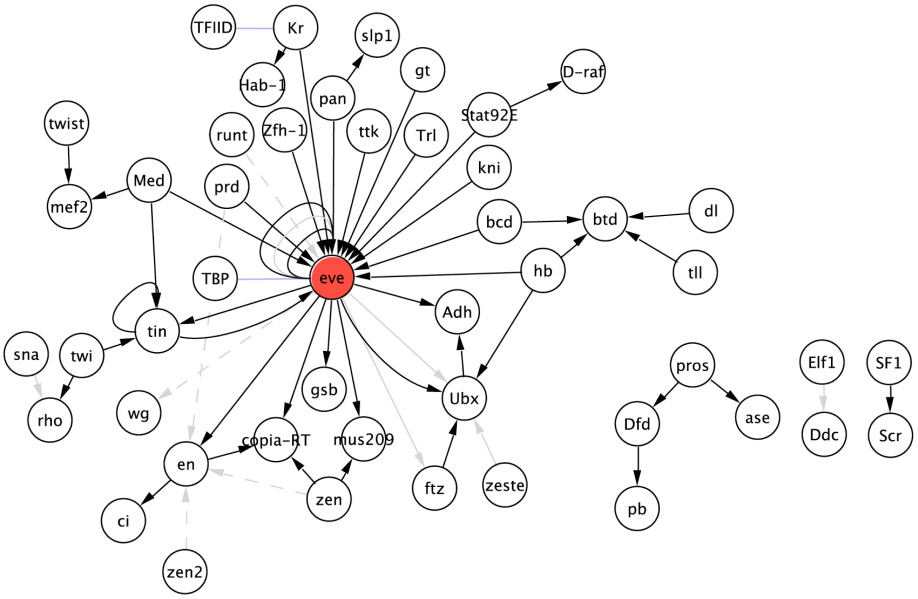


Figure 5.6: Transcriptional regulatory sub-network around the Drosophila transcription factor even-skipped (eve). All nodes and edges are retrieved from eve-related publications in the top 100K abstract list. Black edges are success papers (*i.e.*, publications fully amenable to curation); grey edges are failure papers that report regulatory data (*e.g.*, consensus sites) but are not the primary reference; grey dashed edges are failure papers that contain regulatory data that is not complete enough to allow full curation; blue edges are failures that report protein-protein interactions.

accelerate the community annotation of *cis*-regulatory networks and sequences. These advances should help generate the necessary training and test sets to improve the reliability of computational and high-throughput experimental methods in regulatory biology.

Previously it has been shown that manually-curated and automatically-extracted binary TF→TG interactions can be assembled into transcriptional regulatory networks [205, 204, 195, 10]. Here we show that abstract relevance ranking using a vector space model can be used to enhance the manual annotation of binary TF→TG interactions, and should likewise further improve the automated extraction of binary TF→TG interactions to construct regulatory networks. We have also shown that the binary TF→TG interactions that are central to the construction of transcriptional regulatory networks can be extracted from text even when a full curation of the *cis*-regulatory sequence responsible for this interaction may not be possible. Our vector space model also has allowed us to generate an enhanced "queue" of papers for annotation, and to gain a deeper insight into the size of the corpus of papers that may contain *cis*-regulatory sequences amenable to curation, which we estimate is on the order of 30,000 papers or more. At the rate of 1-2 hours curation time per paper, it would take a single person 15-30 years to curate and annotate this corpus manually. This estimate demonstrates the need for distributed community annotation systems and for computational tools that can assist the extraction of relevant *cis*-regulatory information.

As presaged by Lincoln Stein [236], our results demonstrate that it is indeed possible to leverage text-mining technologies to accelerate genome annotation. Our proof of principle in the field of regulatory annotation is only one potential application of text-based genome sequence annotation. The general combining of information retrieval systems with sequence extraction techniques [268], should allow researchers to enrich for any specific sub-domain of biomedical research and use sequence data reported in these corpora to directly annotate genomic regions of interest in a highly automated fashion. For example, the false positive mappings that correspond to coding sequences in our set of documents with high *cis*-regulatory content are likely to be mainly for proteins binding *cis*-regulatory sequences, and thus strategies similar to ours could accelerate the labor intensive identification of sequence specific transcription factors [3, 191]. Clearly, it is preferable that researchers deposit their sequences and annotations in databases as a condition for publication and thereby preclude the need for post-publication extraction of such valuable biological data. With established databases for general sequence submission (*e.g.* GenBank [21]) and specialized annotation [163, 180], researchers now have the necessary tools to deposit their *cis*-regulatory data. In the absence of direct database submission, we recommend that researchers report certain minimum information (absolute coordinates with genome build, sequence with sufficient flank, standard gene identifiers, official species name or id) to assist the regulatory annotation, both the human and automated, that is needed to help catalyze advances in the field of gene regulation.

Chapter 6

Data fusion on biomedical text and gene expression assays

"The capacity to blunder slightly is the real marvel of DNA. Without this special attribute, we would still be anaerobic bacteria and there would be no music."

Lewis Thomas (1913 - 1993)

In the last chapter of this dissertation, we will focus specifically on text-mining aspects of genome annotation. We have discussed text-mining applications throughout the dissertation: we have seen how literature is used as a data source for candidate gene prioritization in Endeavour (Section 3.6), we have seen how PubMed abstracts are mined to correlate human cytogenetic banding patterns to disease and phenotypes (Section 3.4.1), we have seen how literature is a source for gene-phenotype correlations in the aGeneApart pipeline (Section 3.5), and how the output from that process is used in intelligent genome analysis databases like Store+Bench (Section 4.6) and DECIPHER (Section 4.4). In Chapter 5, it was also discussed how literature is used to feed the human genome annotation process in the context of elucidating *cis*-regulatory elements.

In this final chapter, we will briefly introduce the main theoretical aspects of text-mining, and discuss two case studies of data fusion between text-mining and expression data. The first case we discuss is TXTGate, a platform that serves at profiling genes based on public literature. The system is aimed at the evaluation of groups of genes, that typically stem from microarray clustering experiments and helps investigate the coherence of groups of co-regulated genes, in order to support functional and pathway studies. This framework combines literature indices of selected public biological resources in a flexible text-mining system designed towards the analysis of groups of genes. By means of tailored vocabularies, term- as well as gene-centric views are offered on selected textual fields and MEDLINE abstracts used in Entrez Gene and the Saccharomyces Genome Database. Sub-clustering and links to external resources allow for in-depth analysis of the resulting term profiles.

The second case involves the integration of microarray expression data and text

data for elucidating regulatory networks. In this case study, prior information is constructed based on text to enhance the performance of predictive Bayesian networks that use expression data as input. Elucidating regulatory networks is an intensively studied topic in bio-informatics. Integration of different sources of information can facilitate this task. In this section, we illustrate how to incorporate different information sources in the structure prior of a Bayesian network. Two complimentary sources of information are under investigation: PUBMED abstracts combined with publicly available taxonomies or ontologies, and known protein-DNA interactions. These priors, either separately or combined, have the potential of reducing the complexity of reverse engineering regulatory networks while creating more robust and reliable models. Moreover this approach can easily be extended with other data sources. In such a way Bayesian networks provide a powerful framework for data integration and regulatory network modeling. In the current context, we look specifically at how text and expression array data can be used to improve the prognosis prediction of breast, lung and ovarian cancer patients.

6.1 Case 1: TXTGate – a platform for gene group profiling and expression array cluster validation

6.1.1 Introduction - a flexible framework

The following section discusses TXTGate [86], a text-mining framework that combines literature indices of selected public biological resources in a flexible web based system designed towards the analysis of groups of genes. The main aim of this framework is the analysis of clusters of co-expressed genes that stem from microarray expression assays. By means of tailored vocabularies, term- as well as gene-centric views are offered on selected textual fields and MEDLINE abstracts used in Entrez Gene[1] (formerly LocusLink [183]) and the Saccharomyces Genome Database. Sub-clustering and links to external resources allow for in-depth analysis of the resulting term profiles.

6.1.2 Rationale

High-throughput methods such as gene expression microarrays enable systematic testing of the functions of multiple genes, their interrelatedness and the controlled circumstances in which ensuing observations hold. As a result, scientific discoveries and hypotheses are stacking up, all primarily reported in the form of free text. However, as large amounts of raw textual data are hard to extract information from, various specialized databases have been implemented to provide a complementary resource for designing, performing or analyzing large-scale experiments.

Until recently, the fact that there is little difference between retrieving an abstract from MEDLINE and downloading an entry from a biological database has been largely overlooked[2]. The fading of the boundaries between text from a scientific article and a curated annotation of a gene entry in a database is readily illustrated by the GeneRIF

[1] NCBI Entrez Gene (`http://www.ncbi.nlm.nih.gov/sites/entrez`)

[2] Gerstein *et al.*, Blurring the boundaries between scientific papers and biological databases. Nature Online (`www.nature.com/nature/debates/e-access/articles/gernstein.html`)

feature in Entrez Gene (formerly LocusLink), where snippets of a relevant article pertaining to a gene's function are manually extracted and directly pasted as an attribute in the database. The broadening of biologists' scope of investigation, along with the growing amount of information, result in an increasing need to move from single gene or keyword-based queries to more refined schemes that allow comprehensive views of text-oriented databases.

As gene-expression studies typically output a list of dozens or hundreds of genes that are co-expressed, a researcher is faced with the assignment of biological meaning to such lists. Several text-mining approaches have been developed to this end.

Previous public efforts

Masys *et al.* [155] link groups of genes with relevant MEDLINE abstracts through the PubMed engine. Each cluster is characterized by a pool of keywords derived from both MeSH[3] and the UMLS[4] Ontology. Jenssen *et al.* [114] set up a pioneering online system to link co-expression information from a microarray experiment with the co-citation network they constructed. This literature network covers co-occurrence information of gene identifiers in more than 10 million MEDLINE abstracts. Their system characterizes co-expressed genes using the MeSH keywords attached to the abstracts about those genes. Shatkay *et al.* [222] link abstracts to genes in a probabilistic scheme that uses the EM[5] algorithm to estimate the parameters of the word distributions underlying a 'theme'. Genes are identified as similar when their corresponding gene-by-documents representations are close. Chaussabel and Sher [42] and Glenisson *et al.* [85] provide a proof of principle on how clustering of genes encoded in a keyword-based representation can further discern relevant sub-patterns. Finally, Raychaudhuri *et al.* [186] developed a method called neighborhood divergence, to quantify the functional coherence of a group of genes using a database that links genes to documents. The score is successfully applied to both gold-standard and expression data, but has the slight drawback that it does not give information on the actual function. Their method is indeed geared to the identification of biologically coherent groups, rather than their interpretation.

TXTGate scope

The system described here was built taking into account three main considerations, in an attempt to improve the quality and interpretability of term profiles. First, the construction of a sound linkage between genes and MEDLINE abstracts is often problem dependent and constitutes a research track on its own that requires advanced document-classification strategies as, for example, proposed by Leonard *et al.* [135] or Raychaudhuri *et al.* [185]. Despite some shortcomings, therefore, curated gene literature references are helpful resources to exploit.

Second, the information contained within curated gene references is sometimes diverse and can range from sequence to disease. In addition, the research questions that scientists are addressing when they scrutinize gene groups from high throughput assays are similarly diverse. Therefore, considering all the terms occurring in a large set of documents (that is, a general vocabulary) might be detrimental to the extraction of terms that are relevant to the question at hand. The construction of separate

[3]MeSH: Medical Subject Headings (`www.nlm.nih.gov/mesh/meshhome.html`)
[4]UMLS: Unified Medical Language System (`www.nlm.nih.gov/research/umls/`)
[5]Expectation Maximization

vocabularies according to gene name, disease and function seems a logical choice to provide increased insight.

Third, as mentioned previously, annotations offered by curated gene databases are often in semi-structured form and encompass keywords, sentences or paragraphs. To facilitate integration of such annotations with existing knowledge, controlled vocabularies that describe conceptual properties are of great value when constructing inter-operable and computer-parsable systems. A number of structured vocabularies have already arisen (Gene Ontology)[6], MeSH, eVOC [120]) and, slowly but surely, certain standards are systematically being adopted to store and represent biological information[7].

Armed with these insights, TXTGate[8] [86] was developed, a platform that offers multiple 'views' on vast amounts of gene based free-text information available in selected curated database entries and scientific publications. TXTGate enables detailed functional analysis of interesting gene groups by displaying key terms extracted from the associated literature and by offering options to link out to other resources or to sub-cluster the genes on the basis of text. This way, on the one hand, the need for easy means to validate gene clusters arising from, for instance, microarray experiments, is addressed. On the other hand, the platform offers an approach to the problem of using scientific literature in the form of free text as a source of functional information about genes. The strength of TXTGate is its use of tailored vocabularies to visualize only the information most relevant to the query at hand. TXTGate is implemented as a web application and is available for academic use.

6.1.3 Related software

This framework builds on the general ideas of textual profiling and clustering presented by Blaschke *et al.* [29] and Chaussabel and Sher [42], where the utility of literature indices for profiling gene groups in yeast and humans is proven. TXTGate implements the vector-space model for gene profiling [85] and provides indices for MEDLINE abstracts and selected functional annotations from two public databases. Various engineered domain-specific vocabularies (term- as well as gene centric) act as perspectives to the literature and the tool provides direct links to external resources. In what follows, TXTGate is compared to other reported biological text-mining software.

MedMiner[9] [239] retrieves relevant abstracts by formulating expanded queries to PubMed. It uses entries from the GeneCards database [187] to fish for additional relevant keywords to expand a query. The resulting filtered abstracts are summarized in keywords and sentences, and feedback loops are provided. Nevertheless, the system is directed at querying terms and specific gene-drug or gene-gene relationships, rather than at scrutinizing gene clusters.

MedMOLE[10] [39] is also a system to query MEDLINE more intelligently and detects Human Genome Organization (HUGO) names in abstracts via a natural language processing (NLP)-based gene-name extractor. The retrieved abstracts can be clustered, and top keywords are presented. However, the application scales less well, is not effective at profiling groups of genes, and the summaries provide much less detail than MedMiner and TXTGate.

[6] Gene Ontology Consortium (`www.geneontology.org`)
[7] Gene Ontology Annotation (`www.ebi.ac.uk/GOA`)
[8] TXTGate Portal (`www.esat.kuleuven.be/txtgate`)
[9] MedMiner (`discover.nci.nih.gov/textmining`)
[10] MedMOLE at CINECA (`www.cineca.it/HPSystems/Chimica/medmole`)

GEISHA [29][11] is a tool for profiling gene clusters with an emphasis on summarization within a shallow parsing framework. This system was implemented for Escherichia coli but is no longer updated.

PubGene [114][12] is a database containing gene co-occurrence and co-citation networks of human genes derived from the full MEDLINE database. For a given set of genes it reports the literature network they reside in, together with their high-scoring MeSH terms. As not all relevant information can be captured by gene symbols or MeSH terms, the functionalities offered by TXTGate provide complementary views to interpret groups of genes. Although the co-linkage feature in TXTGate, being a weaker form of co-occurrence that spans only the set of 73,152 MEDLINE abstracts used in Entrez Gene, is less elaborate than the possibilities offered by PubGene, its utility and added value will be shown through its integration in the broader TXTGate framework.

MedGene[13] [104] and G2D[14] [175] are specialized databases that, in contrast to TXTGate, are geared at ranking genes by disease. They accept user-defined queries scrutinizing gene-disease, disease-disease or gene-gene relationships extracted from the literature.

Finally, MeKE[15] [44] is an application listing gene functions extracted by an ontology based NLP system. Its current setup is directed more towards a functional knowledge base, rather than comprehensibly profiling information coming from groups of genes, as offered by the TXTGate software.

6.1.4 Representing text data

Vector Space Model

The Vector Space Model (VSM) is an algebraic model for representing text documents [202]. Documents are represented as vectors of index terms. The VSM equally allows objects in general to be represented: since a gene can be represented by documents that describe it, or a patient can be represented by a discharge report or piece of text describing his/her phenotype, these objects too can have representations in a vector space. The model is used for information filtering, information retrieval, indexing and relevance rankings. The order in which the terms appear in the document is lost in the vector space representation. For this reason, the model is sometimes referred to as 'bag of words'.

A document d is represented as a vector $\vec{d} = [t_1, t_2, \cdots, t_n]$ where t_i correspond to terms (words or phrases) from a vocabulary of size n. Each dimension corresponds to a separate term. If a term occurs in the document, its value t_i in the vector is non-zero. Several different ways of computing these values, also known as weighting schemes, have been developed.

Similarity

The VSM allows documents to be compared. Formally, the similarity between two documents d_1 and d_2 can be computed as:

[11] DNA Array Analysis with GEISHA (`www.pdg.cnb.uam.es/blaschke/cgi-bin/geisha`)
[12] PubGene Gene Database and Tools (`http://www.pubgene.org`)
[13] MedGene Database (`hipseq.med.harvard.edu/MEDGENE`)
[14] G2D Candidate Genes to Inherited Diseases (`www.bork.embl-heidelberg.de/g2d`)
[15] MeKE (Medical Knowledge Explorer) (`ismp.csie.ncku.edu.tw/~yuhc/meke`)

$$Sim(\vec{d_1}, \vec{d_2}) = cos(\widehat{\vec{d_1}\vec{d_2}}) = \frac{\vec{d_1} \cdot \vec{d_2}}{\|\vec{d_1}\|\|\vec{d_2}\|} = \frac{\sum_i w_{1,i} w_{2,i}}{\sqrt{\sum_i w_{1,i}^2}\sqrt{\sum_i w_{2,i}^2}}, \tag{6.1}$$

where $w_{j,i}$ is the weight of term t_i in document d_j, and $\cdot$ is the vector multiplication.

Indexing

A (usually large and sparse) matrix $\mathbf{W}$ with values $w_{j,i}$ representing the weight of term t_i in document d_j is called an index. The construction of an index can involve several pre-processing steps, such as stemming. Stemming entails normalization of inflections and alternative writings to a single base form, called a stem, so that 'microcephalic' and 'microcephaly' both map to the same term in $\mathbf{W}$. Another step involves applying a weighting scheme, of which TF-IDF (Term Frequency combined with Inverse Document Frequency) is one of the best known schemes.

For a detailed introduction to theoretic aspects of text-mining, consult the ample body of literature [12, 26, 25, 24, 84, 91, 149, 202].

The choice of indexing software throughout this PhD has evolved from the use of legacy C code available in the ESAT research group over an in house indexing framework written in Java via text processing scripts in Perl to the use of a robust, fast and reliable open source software that has become a *de facto* standard in the field: Lucene [96].

6.1.5 Application overview

A conceptual overview of the system is shown in Figure 6.1. Various literature indices were created based on selected annotation fields and linked MEDLINE information, both present in the curated repositories Entrez Gene and the Saccharomyces Genome Database (SGD). Several tailored vocabularies derived from public resources (GO, MeSH, Online Mendelian Inheritance in Man (OMIM), eVOC and HUGO) act as a perspective on the textual information. A user-defined query on any of these indices by providing a group of genes of interest results in a summary keyword profile which can be used for further query building for a variety of other databases. Currently, TXTGate smoothly accommodates queries of around 200 genes. Alternatively, the group can be sub-clustered on the basis of the selected textual information to discern substructures not apparent in the original summary profile. The operations that can be carried out are described below.

Combining multiple, linked documents into a single gene profile

When a given gene has several curated MEDLINE references associated to it, these abstracts are combined into an indexed gene entry by taking the mean profile. This operation is part of the off-line process.

Combining multiple gene profiles into a group profile

To summarize a cluster of genes and explore the most interesting terms they share, the mean and variance of the terms over the group are computed. Although simple,

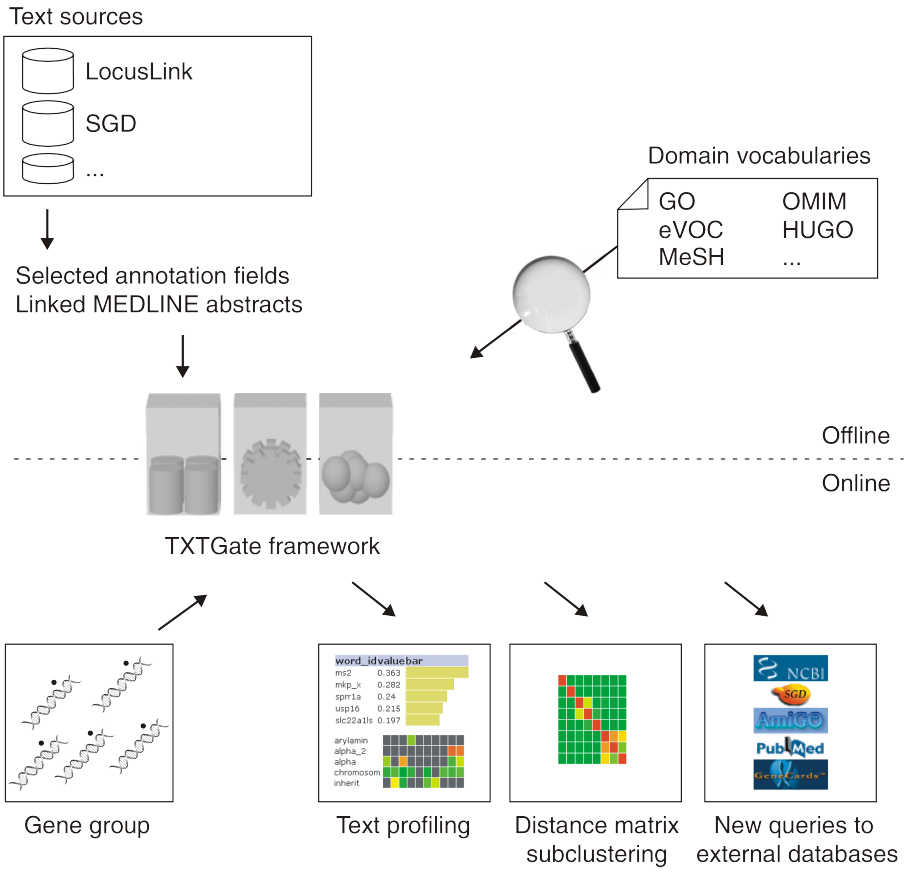


Figure 6.1: Conceptual overview of TXTGate. Two different sources of textual information about genes (Entrez Gene and SGD) were indexed using different domain vocabularies (off-line process). These indices are used online for textual gene profiling and clustering of interesting gene groups. TXTGate's link-out feature to external databases makes it possible to investigate the profiles in more detail.

Table 6.1: Overview of the indexed resources of textual information in TXTGate. The second column specifies which fields from the given resource were used. The third column lists the domain vocabularies with which the text fields were indexed.

Resource	Information fields	Domain vocabularies
Entrez Gene	Linked MEDLINE abstracts	GO, MeSH, eVOC, OMIM, HUGO gene symbols
	GeneRIF annotations	GO
	Functional summaries	GO
	GO annotations	GO
SGD	Linked MEDLINE abstracts	Pruned GO, SGD gene symbols
	GO annotations	Pruned GO

these statistics already reveal information on interesting terms characterizing the gene group. This is performed online.

Sub-clustering gene profiles

TXTGate offers the possibility of sub-clustering a group of a maximum of 200 genes online by means of hierarchical clustering. Ward's method was chosen because of its deterministic nature and the computational advantage of using the same solution when consecutively considering different numbers of clusters k. By varying the threshold at which to cut the tree, one can obtain an arbitrary number of clusters.

Text profiling, clustering and the supporting web interface are implemented as a Java web application that communicates with a MySQL database via Java Remote Method Invocation[16]. The literature indices are currently generated using Lucene[17] and a set of locally developed extensions.

6.1.6 Program development

Indexing

The indices are built using the vector-space model [12], where a textual entity is represented by a vector (or text profile) of which each component corresponds to a single (multi-word) term from the entire set of terms (the vocabulary) being used. For each component a value denotes the importance of a given term, represented by a weight. Indexing a document d_i is performed by the calculation of these weights:

$$d_i = (w_{i,1}, w_{i,2}, w_{i,3}, \ldots, w_{i,n}) \tag{6.2}$$

Each $w_{i,j}$ in the vector of document i is a weight for term j from the vocabulary of size N. This representation is often referred to as 'bag-of-words'. All textual information is stemmed using the Porter stemmer [181] (stemming is the automated conflation of related words, usually by reducing the words to a common root form) and indexed with a normalized inverse document frequency (IDF) weighting scheme, a reasonable choice for modeling pieces of text comprising up to 200 terms, as observed

[16] Java Remote Method Invocation (Java RMI) (`java.sun.com/products/jdk/rmi`)
[17] `lucene.apache.org/`

in database annotations and MEDLINE abstracts. With D the number of documents in the collection and D_t the number of documents containing term t, IDF is defined as:

$$IDF = \log D/D_t \tag{6.3}$$

All validations described in this chapter were performed at the time of release and publication of this framework, and are based on the indices present in the system at that time. These indices were constructed from the entire Entrez Gene (as of 8 April, 2003) and SGD (15 January, 2003) databases, from which a number of interesting fields were identified and selected (such as GO annotations and functional summaries). Although the validation in this chapter was performed on the data described, the actual data underlying the application is updated on a regular basis and hence, current results may differ.

Although indexing these database entries could have been performed on all fields at once, a preservation of selected parts of Entrez Gene's and SGD's logical field structure was deemed more appropriate for functional gene profiling. Not only the textual annotations were indexed but also the 73,152 MEDLINE abstracts referred to in all entries of Entrez Gene at the time of evaluation, as well as the 24,909 abstracts then linked to from SGD. Gene-specific indices were created by taking the average over all indices of MEDLINE abstracts annotated to a certain gene in Entrez Gene and SGD. The resulting indices are used in TXTGate as a basis for literature profiling and further query building of genes of interest. Table 6.1 provides an overview of the indexed resources of textual information and connects them to the domain vocabularies used.

6.1.7 Construction of domain vocabularies

Five different term-centric domain vocabularies were constructed and provide different views on the gene-specific information that was indexed. All vocabulary sources underwent parsing and pruning operations to obtain stemmed words and phrases, eliminating stop words (such as 'then', 'as', 'of', 'gene') from a handcrafted list. The Porter stemmer [181]) was applied to avoid information loss due to morphology and inflexion. Although stemming is not always desirable, for relatively small documents it has proven advantageous. Where applicable, phrases were derived directly from the vocabulary source through the use of a combination of regular expressions.

A first vocabulary was derived from the GO[18] and comprises 17,965 terms. GO is a dynamic controlled hierarchy of multi-word terms with a wide coverage of life-science literature, and genetics in particular. It can be considered an ideal source from which to extract a highly relevant and relatively noise-free domain vocabulary. All composite GO terms shorter than five tokens were kept as phrases. Longer terms containing brackets or commas were split to increase detection rate of relevant phrases. For the yeast indices, the vocabulary was pruned, retaining only those terms occurring at least twice and in less than 20% of all MEDLINE abstracts referred to in SGD[19], obtaining a new vocabulary of 3,867 terms.

Two other domain vocabularies are rather similar in scope but differ in size. One is based on the MeSH[20], the National Library of Medicine's controlled vocabulary

[18]Gene Ontology Consortium (`www.geneontology.org`)
[19]Saccharomyces Genome Database (`www.yeastgenome.org`)
[20]Medical Subject Headings (`www.nlm.nih.gov/mesh/meshhome.html`)

thesaurus, and counts 27,930 terms. The other is based on OMIM's Morbid Map[21]. This is a cytogenetic map location of all disease genes present in OMIM and their associated diseases. All disease terms were extracted to construct a 2,969-term vocabulary. A fifth domain vocabulary was drawn from eVOC [120], a thesaurus consisting of four orthogonal controlled vocabularies encompassing the domain of human gene-expression data. It includes terms related to anatomical system, cell type, pathology, and developmental stage.

In addition to these term-centric domain vocabularies, two gene-centric vocabularies were constructed with the screening of co-occurring and co-linked genes in mind. 'Co-occurrence' denotes the simultaneous presence of gene names within a single abstract, as described by Jenssen *et al.* [114]. Here, 'co-linkage' is defined as a weaker form of co-occurrence screening for the simultaneous presence of gene names (or terms, generally) in the pool of abstracts that is linked to a given group of genes.

From the HUGO database[22] a vocabulary was derived consisting of all uniquely defined human gene symbols and their synonyms. In total, this vocabulary consists of 26,511 gene symbols. The second vocabulary consists of all uniquely defined yeast gene symbols found in SGD and contains 11,319 terms. As these official gene symbols are frequently requested and used by scientists, journals and databases, it can be assumed they constitute a good first approximation to detect gene occurrence in MEDLINE abstracts. The domain vocabularies adopted here are listed in Table 6.2.

Table 6.2: Overview of the domain vocabularies in TXTGate. The vocabularies are named after the resource they stem from.

Domain vocabulary	**Number of terms**
Term-centric	
GO	17,965
GO-pruned (yeast)	3,867
MESH	27,930
OMIM	2,969
eVOC	1,553
Gene-centric	
HUGO gene symbols (human)	26,511
SGD gene symbols (yeast)	11,319

6.1.8 Online clustering

The online clustering is done with a Java implementation of Ward's method for hierarchical clustering [112]. Ward's method outperforms single, average or complete linkage. The similarity measure used is the cosine distance between two vector representations $\vec{d_i}$ and $\vec{d_j}$. The similarity between a newly formed cluster (r, s) (by linking two existing vectors/clusters) with $(n_r + n_s)$ elements and an existing cluster (t) with n_t elements is given by:

[21]OMIM - Online Mendelian Inheritance in Man `www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM`)

[22]HUGO Gene Nomenclature Committee (HGNC, `www.gene.ucl.ac.uk/nomenclature`)

$$\mathrm{d}\left[(t),(r,s)\right] = \alpha_r \mathrm{d}\left[(t),(r)\right] + \alpha_s \mathrm{d}\left[(t),(s)\right] + \beta \mathrm{d}\left[(r),(s)\right] \tag{6.4}$$

with

$$\alpha_r = \frac{n_r + n_t}{n_r + n_s + n_t}, \alpha_s = \frac{n_s + n_t}{n_r + n_s + n_t}, \text{ and } \beta = \frac{-n_t}{n_r + n_s + n_t} \tag{6.5}$$

Given the preferred number of clusters k, the linkage tree is cut at the appropriate level to yield k clusters.

6.1.9 Cluster coherence

As a measure of textual coherence, C_G, the median distance is calculated in term space from the profile of the group G of size n_G to the individual profiles, g_i, of all genes in that group:

$$C_G = \mathrm{med}_i\{\mathrm{d}(g_i, \hat{g})\}_{i=1..n_G} \tag{6.6}$$

with

$$\hat{g} = avg\{g_i)\}_{i=1..n_G} \tag{6.7}$$

In order to assess its significance, a background distribution from random gene clusters of different sizes is computed.

To demonstrate how Equation (6.7) scores groups of functionally related genes, its performance on 10 cell-cycle groups of Spellman *et al.* [233] is shown. These involve 126 genes in total, which are identified manually as well as by expression analysis. As can be seen in Table 6.3, all but the sporulation group display p-values below the 1-sided 0.025 threshold (that is, a gene group G is considered coherent if C_G is smaller than expected by chance).

This result corroborates the ability of Equation (6.7), and more importantly of the vector-space model that underlies TXTGate, to represent biologically relevant functional information. It provides a quantitative foundation that supports the underlying methodology of TXTGate.

6.2 TXTGate summarizes and identifies sub-clusters

TXTGate allows online sub-clustering and profiling of gene groups via terms extracted from MEDLINE. Below, two examples are described.

6.2.1 Yeast data

Starting from the reference data set from Eisen *et al.* [56], TXTGate was used to conduct a textual analysis similar to that of Blaschke *et al.* [29]. In Table 6.4 the text profiles of cluster E from Eisen *et al.* are shown in correspondence to a sub-clustering with $k = 2$. Although several of the text-mining settings in Blaschke *et al.* are different from the ones described here (because of the differences in MEDLINE corpus, textual analysis methodology, and the clustering algorithm used), a comparison of the term profiles in both analyses shows that TXTGate also identifies $E1$ as being related to glycerol, whereas $E2$ is more related to pyruvate metabolism and ethanol fermentation (for more details, see Blaschke *et al.* [29]). Detailed text profiles for each

Table 6.3: Significance of coherence score C_G. The significance is calculated with respect to 100-fold randomization for 10 cell-cycle related, functional groups selected from Figure 7 in Spellman *et al.* [233]. All groups are functionally coherent according to the score mentioned here, except for the sporulation group.

Gene groups	**Size**	**Coherence score**
Cell-cycle control	19	1.01E-167
DNA repair	3	3.91E-61
Fatty acids/lipids	25	4.28E-08
Glycosylation	7	6.29E-06
Methionine	5	9.88E-28
Mitotic exit	9	1.50E-82
Nutrition	19	1.76E-18
Pseudohyphae	10	2.79E-05
Secretion	13	1.11E-06
Sporulation	16	1.11E-01

of the clusters B, C, D, E, F, G, H, J, and K in Eisen *et al.* are referenced in the TXTGate publication [86].

6.2.2 Human data

To assess the quality of the indexed MEDLINE abstracts used in Entrez Gene, the output from TXTGate is compared to results presented in Chaussabel and Sher [42], where the authors describe, among other experiments, the profiling and clustering of nearly 200 genes involved in the 'common transcriptional program' induced in human macrophages upon bacterial infection. Here, the results are interpreted by retrieving the MEDLINE textual profiles of all genes in the clusters and TXTGate's best-scoring terms are then compared to the cluster terms in Chaussabel and Sher [42]. The results of the first four (non-overlapping) clusters (clusters a, b, c and d) can be found in Table 6.5. The terms 'adipose', 'metastasis' and 'NM' did not show up in the profiles from TXTGate because they are not contained in the GO domain vocabulary. For cluster e no common terms were found. Running TXTGate using the OMIM vocabulary, however, uncovers exactly those disease-associated terms that were retrieved by Chaussabel and Sher [42] by manually investigating genes from this cluster in the OMIM database. Table 6.6 highlights these terms in bold. As the set of diseases related to these genes is heterogeneous, the relevant terms display a high variance, rather than a high mean, a reason for also including a variance profile. Moreover, the fact that those disease terms are only retrieved by means of the OMIM vocabulary points out that the use of a variety of vocabularies in TXTGate leads to improved insights, a point discussed further in the next section. Note that all other cluster terms have a comparable equivalent in the TXTGate profiles.

Table 6.4: TXTGate profiling of cluster E from Eisen *et al.* [56]. Profiling by sub-clustering ($k = 2$). High-scoring terms are shown for each of the sub-clusters $E1$ and $E2$. Also shown are the terms (excluding gene names) resulting from a similar analysis conducted by Blaschke *et al.* [29]. ⋆ Terms that were labeled specifically to a sub-cluster by Blaschke *et al.* Although several of their settings are different from the ones described here (because of the differences in MEDLINE corpus, textual analysis and the cluster algorithm used), a comparison of the term profiles in both analyses shows that TXTGate also identifies $E1$ as related to glycerol, whereas $E2$ is more related to pyruvate metabolism and ethanol fermentation.

Gene symbol	**Cluster terms in Blaschke *et al.* [29]**	**Terms from TXTGate**
sub-cluster $E1$		
TPT1 FBA1	glyceraldehyde-3-phosphate⋆	glyceraldehyd_3_phosphat_dehydrogenas
GPM1 TKL1	glyceraldehyde-3-phosphate dehydrogenase⋆	glycolyt
PGK1 CDC19	phosphoglycerate kinase⋆	glucos
TDH3 HXK2	phosphoglycerate⋆	enzym
TDH2 TYE7	mutase⋆	glycolysi
ENO2 PFK1	dehydrogenase	carbon
TDH1 ACS2	enolase	pyruv_kinas
	glycerol-3-phosphate dehydrogenase	ethanol
	osmotic stress	phosphoglycer_kinas
	phosphoglycerate	growth
sub-cluster $E2$		
PDC5 PDC1	alcohol⋆	pyruv_decarboxylas
PDC6	transketolase⋆	pyruv
	catabolite repression	glucos
	decarboxylase	enzym
	ethanol	alcohol
	glucose	decarboxyl
	glucose repression	ethanol
	hexokinases	ferment
	pyruvate	thiamin
	pyruvate decarboxylase	decarboxylas

Table 6.5: TXTGate profiling of clusters *a*, *b*, *c*, and *d* from Chaussabel and Sher [42] (GO vocabulary). Corresponding terms in Chaussabel and Sher [42] and TXTGate are in bold. TXTGate's profiles are comparably informative.

Gene symbol	Cluster terms in Blaschke *et al.* [29]		Terms from TXTGate	
Cluster a				
LPL	Lipoprotein		**lipoprotein**	scaveng_receptor
CD36L1	Density		lipas	high_densiti_lipoprotein
LDLR	Cholesterol		ldl	low_densiti_lipoprotein_receptor
	Lipid		ldl_receptor	low_densiti_lipoprotein
	Adipose		**cholesterol**	hdl
Cluster b				
UPA	Invasive	Collagenase	**metalloproteinas**	
PLAUR	Invasion	Collagen	**matrix**	
SERPIN	Metastasis	Matrix	metalloendopeptidas	
MMP1	UPAR	MMP	**collagenas**	
MMP10	UPA	Metalloproteinase	extracellular_matrix	
MMP14	Plasminogen	Molecule-1	alpha	
SPARC	Urokinase-type	Adhesion	**upar**	
	Urokinase	Vascular	**plasminogen_activ**	
	Plasmin	Endothelial	interstiti	
	Activator		**invasion**	
Cluster c				
AMPD3	Adenosine		purinerg	
ADA	A2A		**adenosin**	
ADORA2A	A1		deaminas	
ADORA3	Antagonist		p2	
P2RX	Agonist		p2x	
P2RX1	NM		p1	
P2RX7			**agonist**	
			receptor	
			adenosin_receptor	
			ada	
Cluster d				
IP10	Interferon		tumor_necrosi_factor	
MIP1A	IFN-alpha		cytokin	
MIP1B	IFN		induc	
IL8	Interferon-gamma		**interferon**	
STAT4	IFN-gamma		inflammatori	
IL12B	Inducible		antigen	
TNFRSF9			lymphocyt_activ	
TNFSF9			stimul	
SLAM			chemokin	
TNFRSF5			monocyt	
CD83				

Table 6.6: Comparison of the terms in cluster *e* found by Chaussabel and Sher [42] with those found by TXTGate (OMIM vocabulary). The diversity of the diseases the member genes are related to makes the relevant terms display high variance, rather than high mean. The terms that were also found by Chaussabel and Sher [42] after manual investigation are marked in bold.

Gene symbol	Terms in Chaussabel and Sher [42]	Terms from TXTGate
Cluster e		
CKB	Population	deaminas
AMPD3	Frequency	**lipoprotein_lipas**
ADA	Allele	creatin
ADORA2A	Unrelated	lipoprotein
ADORA3	Families	**krabb**
P2RX	Recessive	**epidermolysi_bullosa**
P2RX1	Autosomal	**alagil**
P2RX7	Disorder	bear
GEM	Severe	leukodystrophi
ARHH	Patient	receptor
LPL	Deficiency	down
CD36L1		**corneal_dystrophi**
LDLR		**deaf**
BF		hdl
GALC		nucleosid
LAMB3		retinoblastoma
GJB2		junction
TGFBI		adhesion
JAG1		congenit_heart_defect
DSCR1		**hear_loss**

6.2.3 Textual information through the eyes of different vocabularies

Another major feature of TXTGate is its ability to present textual information (most importantly MEDLINE abstracts) from different perspectives. This is implemented by offering indices built on GO-, OMIM-, MeSH-, eVOC-, and gene nomenclature-based domain vocabularies respectively. Each configuration is meant to expose a different view of the literature. TXTGate mirrors the dual approach adopted by the external databases it links to, which separate keyword and gene-symbol queries. This, in part, motivated the strategy to construct both term- and gene-centric vocabularies.

To compare the term-based vocabularies, a group of genes involved in colon and colorectal cancer extracted from the OMIM Morbid Map database was profiled. Table 6.7 shows the top 10 terms for each of the retrieved profiles. As can be seen, there is little difference between the MeSH and OMIM profiles, whose terms are mainly medical and disease-related ('colorect_cancer', 'colon_cancer', 'colorect_neoplasm', 'hereditari'), whereas the scope of the GO profile is focused more on metabolic functions of genes ('mismatch_repair', 'dna_repair', 'tumor_suppressor', 'kinas') and the eVOC profile contains terms more related to cell type and development ('growth', 'cell', 'carcinoma', 'metabol', 'fibroblast'). TXTGate's link-out feature allows a more profound analysis of the retrieved terms. Top-ranking terms can be sent to PubMed to retrieve relevant publications. Because all MEDLINE entries are tagged with MeSH keywords, using terms from the MeSH vocabulary assures a successful query. When using the GO-derived vocabulary, terms can be mapped back directly to the GO tree with AmiGO[23] to investigate the term's neighborhood. Other databases available for querying include Entrez Gene and OMIM.

Table 6.7: Various perspectives on textual information in TXTGate. Here it is shown how term-centric vocabularies based on GO, OMIM, MeSH and eVOC profile a group of genes involved in colon and colorectal cancer.

GO	**OMIM**	**MeSH**	**eVOC**
mismatch_repair	colorect	colorect_neoplasm	colorect
tumor	colorect_cancer	mismatch	tumour
dna_repair	tumor	cancer	malign_tumour
mismatch	kinas	colorect	colon
pair	colon	mutat	growth
tumor_suppressor	hereditari	repair	cell
apc	cancer	dna_repair	carcinoma
kinas	colon_cancer	colon	metabol
somat	associ	neoplasm_protein	fibroblast
ra	on	tumor	chain

The same colon cancer case was used to test the ability of the human gene symbol vocabulary in screening for co-linkage of genes. Two different index tables were constructed - one with and one without alternative gene symbols; the former was constructed by mapping all synonymous symbols to the primary gene symbol. The first table has the disadvantage of not being able to disambiguate alternative gene

[23] AmiGO Gene Ontology browser (`http://www.godatabase.org`)

symbols that are mapped to different primary gene symbols; the second does not take synonyms into account, as only true occurrences of a symbol were counted. As a consequence, frequently used symbols are ranked highly, although they are not the official gene symbols. Examples of this are p21 and dra, whose primary symbols are CDKN1A and SLC26A3, respectively. The top 25 gene symbols using the first index table are given in Table 6.8. Most of the retrieved gene names are also in the query list. TXTGate's link-out feature was used to investigate the role of the genes that were not in the input list by sending them as a query to Entrez Gene and GeneCards. This way it was possible to determine their function and their relation to colon and colorectal cancer, as can be seen in Table 6.8.

Table 6.8: Co-linkage analysis of genes with gene-centric vocabularies. This table shows the top-25 co-linked gene symbols in the pool of abstracts of the colon and colorectal cancer case. Genes that were not in the query list are indicated in bold.

Gene name	**Description**
hnpcc	Hereditary nonpolyposis colon cancer
apc	Adenomatous polyposis coli protein
p53	Cellular tumor antigen P53 (tumor suppressor P53)
mlh1	DNA mismatch repair protein MLH1 (mutL protein homolog 1)
muts	E. coli mismatch repair gene mutS
p21	Cyclin-dependent kinase inhibitor 1A
msh2	DNA mismatch repair protein MSH2 (mutS protein homolog 2)
bax	BAX protein, cytoplasmic isoform delta
wnt	Wingless-type MMTV integration site family members
pms2	DNA mismatch repair protein PMS2
src	Proto-oncogene tyrosine protein kinase SRC
dcc	Tumor suppressor protein DCC precursor (colorectal cancer suppressor)
mcc	Colorectal mutant cancer protein MCC
braf	Proto-oncogene serine/threonine protein kinase B-RAF
fgfr3	Fibroblast growth factor receptor 3 precursor
hcc	Hepatocellular carcinoma
dra	Chloride anion exchanger DRA
axin2	AXIS inhibition protein 2
pms1	DNA mismatch repair protein PMS1
abl	Abelson murine leukemia viral oncogene homolog 1
bub1	Mitotic checkpoint serine/threonine protein kinase BUB1
ptp	Protein tyrosine phosphatase family
bcl10	B cell lymphoma/leukemia 10
ptp_pest	Protein tyrosine phosphatase family with C-terminal PEST-motif
prlts	PDGF-receptor beta-like tumor suppressor

6.3 Application of TXTGate to a real-life research problem

In the framework of an ongoing collaboration with a medical research group, the TXTGate system was deployed to tackle a current research issue [119, 263]. A group of 350 genes that were up-regulated in a mouse model for human benign tumors of the salivary glands was analyzed and the results were evaluated in a biological context. A medical researcher constructed a summary of pathological and genetic observations, reflecting relevant literature and expert knowledge. From this, a list of important terms was derived. This list was cross-referenced with textual profiles retrieved from TXTGate using different domain vocabularies. As pathology and developmental issues were the focus of the summary in this case, the eVOC domain vocabulary proved most appropriate, as can be seen from the occurrence of terms such as 'fibroblast', 'embryo', 'tumor', 'teratoma' and so on (see Table 6.9).

Table 6.9: Textual profile of a gene group from a mouse model for human benign tumors of the salivary glands. This table shows the 25 top-ranking terms (for both mean and variance) of the textual profile of a group of 350 genes that were up-regulated in a mouse model for human benign tumors of the salivary glands processed with the eVOC domain vocabulary.

Terms sorted by mean	**Terms sorted by variance**
organ	organ
intern	intern
normal	growth
red	development
male	fibroblast
femal	tumour
visual	red
capillari	nucleu
system	normal
optic	embryo
retina	tera
viral	depend
bacteri	stem_cell
adult	kidnei
chain	epithelium
cell	visual
growth	multipl
tissu	skin
development	muscl_cell
metabol	system
embryo	capillari
fibroblast	mammari
tumour	type_ii
depend	bacteri
genet	male

It can be concluded that the choice of domain vocabulary depends on the experimental context and focus of the investigation. This supports the strategic choice of offering different domain vocabularies.

As a measure of textual coherence C_G, the median distance in vocabulary space from the profile of the group G to the individual profiles g_i of all genes in that group was calculated:

$$C_G = \text{med}_i \left\{ \text{d}(g_i, \hat{g}) \right\}_{j=1..350} \tag{6.8}$$

with

$$\hat{g} = avg\{g_i\}_{i=1..350} \tag{6.9}$$

As background, 5,000 random gene clusters were generated, of both the same size and random sizes (see Figure 6.2), and their coherence was calculated as in Equation 6.8. Two background distributions were derived in order to model the information content for random clusters. This allows the calculation of a p-value for a cluster of genes, expressing the probability that the observed textual coherence occurs by chance. The cluster profile of the 350 up-regulated mouse genes was significant against both the background for random cluster size (p-value 1.8×10^{-3}) and for cluster size 350 (p-value $\leq 10^{-8}$).

6.4 Case 2: Integration of microarray and textual data through Bayesian networks

In this section, we investigate the integration of textual data with microarray expression data. This subject has been profoundly studied by Glenisson [84] (dissertation). In the current context, we look specifically at how text and expression array data can be used to improve the prognosis prediction of breast, lung and ovarian cancer patients.

Microarray data are notoriously noisy, such that models predicting clinically relevant outcomes often contain many false positive genes. Integration of other data sources can alleviate this problem and enhance gene selection and model building. Probabilistic models provide a natural solution to integrate information by using the prior over model space. We investigated if the use of text information from PUBMED abstracts in the structure prior of a Bayesian network could improve the prediction of the prognosis in cancer. Our results show that prediction of the outcome with the text prior was significantly better compared to not using a prior, both on a well known microarray data set and on three independent microarray data sets.

6.5 Introduction

Integration of data sources has become very important in bio-informatics. This is evident from the numerous publications involving multiple data sources to discover new biological knowledge [15, 130, 23]. This is due to the rise in publicly available databases and also the number of databases has increased significantly [76]. Much knowledge still is contained in publications in unstructured form as opposed to being deposited in public databases where they can be amenable to use in algorithms. Therefore we attempted to mine this vast resource and transform it to the gene domain such that

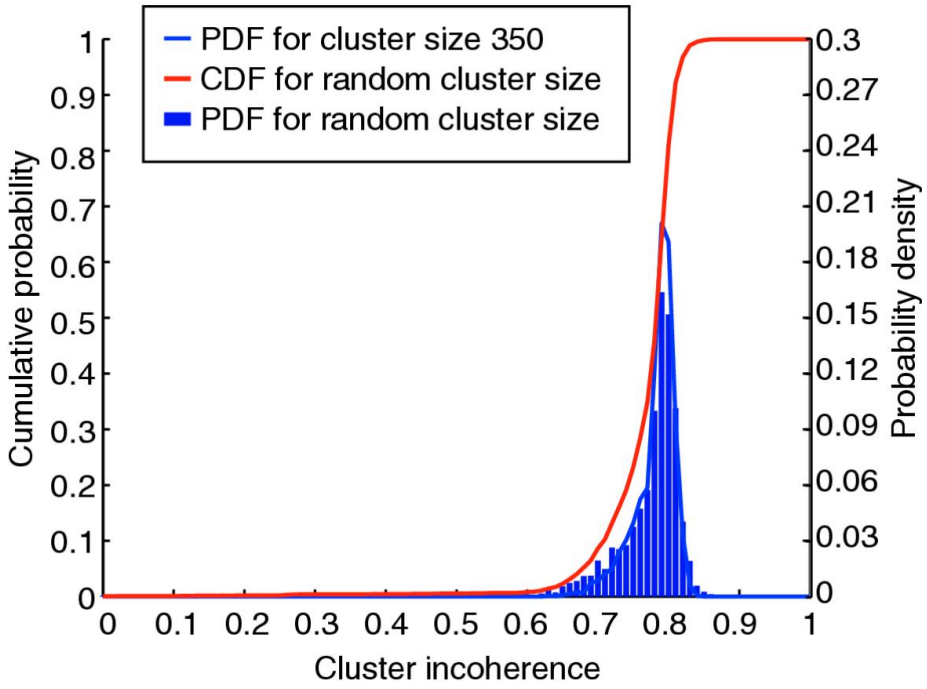


Figure 6.2: Background distributions for cluster incoherence. Cluster incoherence is defined as the median distance in vector space between the mean cluster profile and all individual gene profiles. Probability density functions (pdf) are shown for random clusters of size 350 (blue curve) and random clusters of random size (blue bars). For randomly sized clusters, the cumulative distribution function (cdf) is also shown (red curve).

it can be used in combination with gene expression data. Microarray data are known for their low signal-to-noise ratio and microarray data sets often suffer from a small sample size. This causes that genes are often differently expressed between clinically relevant outcomes purely by chance. Integration of prior knowledge can improve model building in general and gene selection in particular.

In this section, we present an approach to integrate information from literature abstracts into probabilistic models of gene expression data. Integration of different data sources into a single framework potentially leads to more reliable models and at the same time it can reduce over-fitting [130]. Probabilistic models provide a natural solution to this problem since information can be incorporated in the prior distribution over the model space. This prior is then combined with other data to form a posterior distribution over the model space which is a balance between the information incorporated in the prior and the data.

Specifically, we investigated how the use of text information as a prior of a Bayesian network can improve the prediction of prognosis in cancer when modeling expression data. Bayesian networks provide a straightforward way to integrate information in the prior distribution over the possible structures of its network. By mining abstracts we can easily represent genes as term vectors and create a gene-by-gene similarity matrix. After appropriate scaling, such a matrix can be used as a structure prior to build Bayesian networks. In this manner text information and gene expression data can be combined in a single framework. Our approach builds further on methods for integrating prior information with Bayesian networks for other types of data [8, 79] where the authors have shown that structure prior information improves model selection especially when few data is available.

In this study we investigated whether a Bayesian network model with a text prior can be used to predict the prognosis in cancer. Bayesian networks and their combination with prior information have already been studied by others [23, 74, 167, 110]. However, to the authors' knowledge, none have investigated the influence of priors in a classification setting or, more specifically, when predicting the outcome or phenotypic group of cancer patients.

First, we will show how the prior performs on a well known breast cancer data set and examine the effect of the prior in more detail. Subsequently, we will validate our approach on three other data sets studying breast, lung and ovarian cancer.

6.6 Bayesian networks

A Bayesian network is a probabilistic model that consists of two parts: a directed acyclic graph – which is called the structure of the model – and local probability models [174].

Bayesian networks are models that combine probability theory and graph theory, and hence provide a natural tool for dealing with uncertainty and complexity, two problems that occur throughout applied mathematics and engineering. Bayesian networks are modular: complex models are built by combining simpler parts. Graph and probability theory together provide a consistent system to interface models to data [118].

The dependency structure specifies how the variables (in this case, gene expression levels) are related to each other by drawing directed edges between the variables without creating directed cycles. In our case, each variable x_i models the expression of a particular gene. Such a variable or gene is dependent on a possibly empty set

of other variables. These variables are called the parents. Biologically, these parents could, *e.g.*, be putative regulators for these genes. Formally, the probability model is expressed as:

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i | Pa(x_i)) \tag{6.10}$$

where $Pa(x_i)$ are the parents of x_i and n is the total number of variables. Usually the number of parents for each variable is small and therefore a Bayesian network is a sparse way of writing down a joint probability distribution. The second part of this model, the local probability models, specifies how the variables or gene expressions depend on their parents. We used discrete-valued Bayesian networks, which means these local probability models can be represented with Conditional Probability Tables (CPTs). Such a table specifies the probability that a variable takes a certain value given the value or state of its parents.

6.6.1 Model building

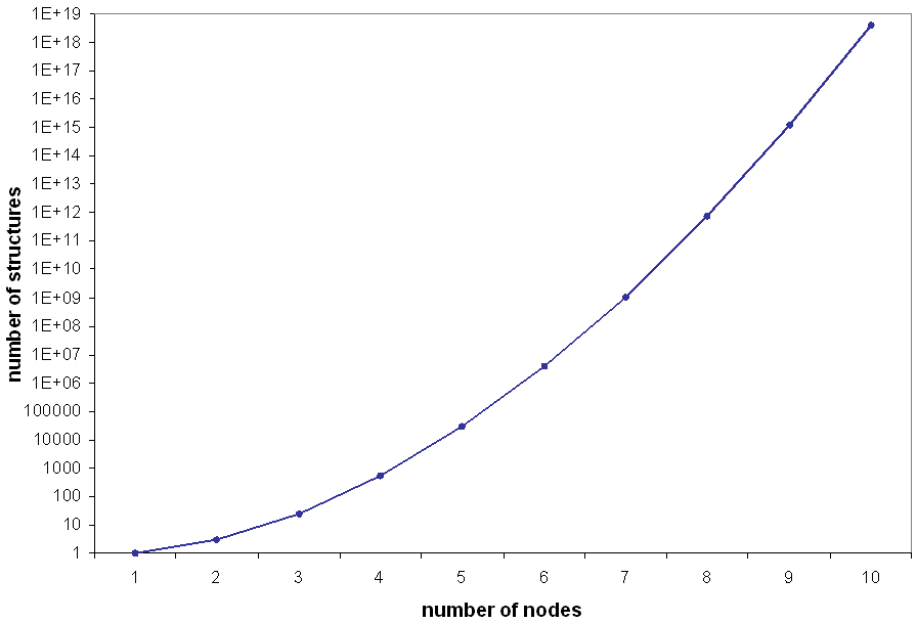
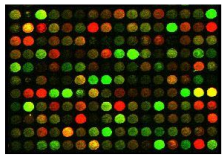


Figure 6.3: The number of possible structures increases super-exponentially with the number of variables in a Bayesian network.

We already mentioned that a discrete valued Bayesian network consists of two parts: the structure and the local probability models. Consequently, there are two steps to be performed during model building: structure learning and learning the parameters of the CPTs. First, the structure is learned through a search strategy. Since the number of possible structures increases super-exponentially with the number

of variables (see Figure 6.3), we used the well-known greedy search algorithm K2 [47] in combination with the Bayesian Dirichlet (BD) scoring metric [47, 80, 98]:

$$p(S|D) \propto p(S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \right], \tag{6.11}$$



$$p(S|D) \propto \overbrace{\prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \right]} \; \overbrace{p(S)}$$

Figure 6.4: Bayesian Dirichlet scoring metric

with N_{ijk} the number of cases in the data set D having variable x_i in state k associated with the j-th instantiation of its parents in current structure S. Γ corresponds to the gamma distribution. Next, N_{ij} is calculated by summing over all states of a variable: $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. In our case the state of a variable refers to the expression of the corresponding gene where each variable can have one of three states: over-expressed, under-expressed or no expression. Next, N'_{ijk} and N'_{ij} have similar meanings as N_{ijk} and N_{ij} but refer to prior knowledge for the parameters. When no knowledge is available they are estimated using the following approach [98]:

$N'_{ijk} = \frac{N}{r_i q_i}$ with N the equivalent sample size, r_i the number of states of variable x_i and q_i the number of instantiations of the parents of variable x_i. K2 uses a prior ordering of the variables to restrict the number of structures that can be built. The order of the variables reflects the causal relationship between the variables, this means that regulators should come before their targets in the ordering. Because the prior ordering of the variables is not known in advance we repeat the model building process for a set of randomly drawn variable orderings and choose the model with the highest posterior BD score. The next step consists of estimating the parameters of the local probability models of each variable in the structure with the highest BD score. This amounts to filling in a CPT for every variable and every possible value of its parents using the data. This Bayesian network can then be used to predict future data.

6.6.2 Structure prior

Previously, two approaches have been used to define informative prior distributions over Bayesian network structures [8]. First there are penalization methods that start from a prior structure and score structures based on the difference with the prior

structure [98]. Secondly there are pairwise methods which define the prior probability of a Bayesian network structure by combining individual edge scores between variables. This method assumes that being a parent of some node is independent of any other parental relation. We have chosen the second approach to model the structure prior where the prior probability of a structure is decomposed as:

$$p(S) = \prod_{i=1}^{n} p(Pa(x_i) \rightarrow (x_i)) \tag{6.12}$$

The probability of a local structure (i.e. $p(Pa(x_i) \rightarrow x_i)$) is then calculated by multiplying the probability that there is an edge between the parents of x_i and, the probability there is no edge between the other variables and x_i:

$$p(Pa(x_i) \rightarrow x_i) = \prod_{p \in Pa_{(x_i)}} p(p \rightarrow x_i) \prod_{y \notin Pa_{(x_i)}} p(y \nrightarrow x_i) \tag{6.13}$$

where $\nrightarrow$ means no edge between y and x_i. These individual edge probabilities can be represented in a matrix. In the Text Prior Section, we will be able to derive a matrix S from the literature where the elements represent the connectedness or similarity between the genes. Rather than using these values immediately as edge probabilities, we will introduce an extra parameter, ν, which is the mean density and which controls the density of the networks that will be generated from the distribution. We will transform all the matrix elements in the prior with an exponent ζ such that the average of the mean number of parents per local substructure is according to the given mean density [8].

Finding the exponent ζ that gives rise to the correct mean number of parents can be done with any single variable optimization algorithm. With this mean number of parents, we can control the complexity of the networks that will be learned.

6.6.3 Inference

After learning the model, we can use it to predict unseen data. This means that we can use a Bayesian network model to predict the value of a variable given the values of the other variables. We used the probability propagation in the *tree of cliques* algorithm [98] to predict the state of the class variable (i.e. the prognosis in cancer). This inference algorithm was then used to evaluate the effect of using a text prior in combination with the expression data described below. To accomplish this, we used a randomization approach where we randomly distributed the data in 70% used to build a model and 30% to estimate the Area Under the ROC curve (AUC). This process was repeated 100 times to have a robust estimate of the generalization performance of the two approaches: with text prior and without text prior. Then these 100 AUCs were averaged and reported. Next, a model was built using the complete data set for both methods and we investigated the possible differences between the Markov blanket variables (*i.e.*, the set of genes which are sufficient to predict the outcome). The average AUC with and without prior are compared by calculating the p-value with a two-sided Wilcoxon rank sum test. The p-values are considered statistically significant if smaller than 0.05.

6.7 Prior data

6.7.1 Gene prior

Since microarray data usually reference thousands of genes, it is infeasible to manually construct a structure prior as described earlier. Therefore, prior construction involves methods based on an automatic elicitation of relationships between genes. In this work, we propose the use of priors that consist of gene-by-gene similarity matrices based on biomedical literature mining. To accomplish this, genes are represented in the Vector Space Model. In the VSM model, each position of a gene vector corresponds to a term or phrase in a controlled vocabulary. In our case, we have constructed a cancer specific vocabulary which was extracted from the National Cancer Institute Thesaurus. Using a fixed vocabulary has several advantages. Firstly, simply using all terms that occur in the corpus of literature linked to the genes involved in the microarray experiment at hand, will result in vectors of considerable size, which means genes are represented in a high dimensional space. As this 'curse of dimensionality' is detrimental to the strength of a metric, the use of only a relatively small set of concepts will improve the quality of calculated gene-to-gene distances. Further reduction of the dimensionality is accomplished by performing stemming, which will allow different terms that in essence convey a same meaning (coughing, coughs, coughed) to be treated as a single concept (cough). Secondly, the use of phrases reduces noise in the data set, as genes will only be compared to each other from a highly domain specific view. Thirdly, a structured vocabulary will enable the use of multi-word phrases as opposed to just single terms, without having to resort to co-occurrence statistics on the corpus to detect them. Fourthly, there is no need to filter out articles and stop words, as only highly specific cancer related terms are considered.

The gene vectors themselves are constructed as follows. For each gene, manually curated literature references are extracted from Entrez Gene. All PUBMED abstracts linked to these genes are then indexed using the aforementioned vocabulary. As a result, all PUBMED abstracts are represented in a high dimensional vector space using IDF (Inverse Document Frequency) weights for non-zero vector positions. The resulting vectors (which represent abstracts, not genes) are normalized to bring them on the union hyper sphere in the vector space, which facilitates cosine similarity calculation. Gene vectors are then constructed by averaging the vectors of all the abstracts associated to that gene by Entrez Gene. Finally, the cosine measure is used to obtain gene-to-gene distances between 0 and 1. These gene-to-gene distances can then be represented as a symmetric matrix S which forms the structure prior for the Bayesian network modeling.

6.7.2 Class variable prior

We have already defined the way the prior is determined between the genes. Since we are developing models which predict the prognosis in cancer, the need exists for an additional variable in the model, namely the outcome class of the patients. This variable describes to which group each sample belongs, for example, good prognosis and poor prognosis. Hence, we need to define the prior relation between the class variable and the genes. To accomplish this, we used terms in the vocabulary which are related to the prediction of the prognosis of cancer, such as outcome, prognosis and metastasis. Next, we counted the number of associations each gene had with prognosis related terms and increased the gene-to-outcome similarity for every additional term

the gene was associated with. Genes which had no association with either term were given a prior probability of 0.5. This information was added to the gene prior creating a structure prior for all the variables studied (i.e. genes and patient outcome). This structure prior is then, after scaling according to the mean density, used in Bayesian network learning.

6.8 Data

To test our approach we used publicly available microarray data on breast cancer [251] (Veer data). This data set consists of 46 patients that belonged to the poor prognosis group and 51 patients that belonged to the good prognosis group. DNA microarray analysis was used to determine the mRNA expression levels of approximately 25000 genes for each patient. Every tumor sample was hybridized against a reference pool made by pooling equal amounts of RNA from each patient. The ratio of the sample and the reference was used as a measure for the expression of the genes and they constitute the microarray data set. This data set was already background corrected, normalized and log-transformed. Preprocessing was done similarly as in [251]. This resulted in 232 genes that were correlated with the patient outcome which were used in our models.

To validate our results we used three publicly available data sets from Bild et al. [27] studying breast, lung and ovarian cancer (Bild data). These data sets contained data on 171 breast cancer patients, 147 ovarian cancer patients and 91 lung cancer patients. The three groups of tumors were analyzed on different Affymetrix chips; the breast tumors were hybridized on Hu95Av2 arrays, the ovarian tumors on Hu133A arrays and the lung tumors on Human U133 2.0 plus arrays. The data were already pre-processed using RMA. For all cancer sites survival data was available and patients were split up in two groups according to the following thresholds: 53 months for breast cancer, 62 months for ovarian cancer and 36 months for lung cancer. The thresholds were chosen to make sure both classes contained approximately the same number of samples. Genes were selected similarly as in the Veer data set by selecting the top 100 genes after ranking them by their correlation with patient survival data.

6.8.1 Discretization

We have chosen discrete valued Bayesian networks; therefore the microarray data has to be discretized. We specifically tried to minimize the loss of relationships between the variables by applying the algorithm of Hartemink [95]. The gene expression values were discretized in three categories or bins: baseline, over-expression and under-expression. This was done using a multivariate discretization method which minimizes the loss of mutual information between the gene expression measurements [95]. First a simple discretization method with a large number of bins is used as a starting point (*e.g.* interval discretization where the complete range of values is divided in a number of equally large bins). Then, the multivariate algorithm starts and for each variable it joins the neighboring bins together which have the smallest decrease in mutual information. This is iterated until each variable has three bins. The resulting discretized data set is used as input into the Bayesian network learning algorithms.

Table 6.10: Results of 100 randomizations of the Veer data set with the Text prior and without prior. The mean AUCs are reported together with the p-value.

Mean Density	Text prior mean AUC	Uniform prior mean AUC	P-value
1	0.80	0.75	0.000396
2	0.80	0.75	0.000002
3	0.79	0.75	0.005770
4	0.79	0.74	0.000006

6.9 Implementation

The software implementation is based on a combination of C++, java, MatLab and perl. The Bayesian network algorithms were implemented in C++. Java Lucene was used for indexing PubMed. MatLab scripts were used for discretization and to construct the structure priors. Perl was used to glue the different steps in the workflow together. A typical analysis took between 6 and 25 minutes depending on the data set size. All analyses were run on AMD dual core opteron 2.4 GHz with between 4 and 16 GB RAM memory.

6.10 Results and discussion

6.10.1 Veer data

First, we assessed the performance of the text prior regarding prediction of outcome on the Veer data set. We performed 100 randomizations of the data set without a prior and 100 randomizations with the text prior (as described in the Model building and testing Section in Materials and methods). We repeated the analysis for different values of the mean density to asses if this parameter had an influence on the results. Table 6.10 shows the mean AUC for both methods and for increasing mean density. The most important conclusion that can be drawn from Table 6.10 is that using the text prior significantly enhances the prediction of the outcome (P-value $\prec 0.05$). The text prior guides model search and favors genes which have a prior record related to prognosis. This knowledge improves gene selection and most likely wards off genes which are differentially expressed by chance. Additionally, Table 6.10 shows that the mean density has no influence on the result in the tested range. The mean density controls the complexity of the network; therefore, large values should be avoided since the danger of over-fitting increases. Note that the results for the mean AUC without prior are essentially the same as our previously obtained result [80].

Next, we used the complete data set and we built one model with text prior and one model without the text prior, to evaluate the set of genes which are sufficient to predict the outcome (i.e. the genes in the Markov blanket of the outcome). We call the former, the TXTmodel and the latter UNImodel. Table 2 shows the gene names that appear in both models. The average text score (i.e. the probability the gene is related to patient outcome according to literature) of the genes in the TXTmodel is 0.85 compared to only 0.58 for the UNImodel. The text prior thus has its expected effect and includes genes which have a prior tendency to be associated with the prognosis of cancer. There are only 10 genes in the TXTmodel compared to 15 genes in the UNImodel

Table 6.11: Genes sufficient to predict the outcome variable for the TXTmodel and the UNImodel.

TXTmodel:	MYLIP,TP53,ACADS,VEGF,ADM,NEO1,IHPK2,CA9,MMP9,BIRC5
UNImodel:	PEX12,LOC643007,WISP1,SERF1A,QSER1,ARL17P1,LGP2,IHPK2, TSPYL5,FBCO31,LAGE3,IGFBP5,AYTL2,TP53,PIB5PA

Table 6.12: Results of 100 randomizations of the three Bild data sets with the Text prior and without prior. The mean AUCs are reported together with the p-value.

Mean Density	Text prior mean AUC	Uniform prior mean AUC	P-value
Breast	0.79	0.75	0.00020
Lung	0.69	0.63	0.00002
Ovarian	0.76	0.74	0.02540

which indicates that TXTmodel needs fewer genes. Moreover, the TXTmodel has many genes which have been implicated in breast cancer or cancer in general such as TP53, VEGF, MMP9 [171], BIRC5, ADM [168], and CA9. Next, ACADS, NEO1 and IHPK2 have a weaker link to cancer outcomes whereas MYLIP has no association. In the UNImodel, as expected, far less genes are present which have a strong link with cancer outcomes which likely increases the probability of false positives. Only WISP1, FBXO31, IGFBP5 and TP53 have a relation with breast cancer outcome. The other genes have mostly unknown function or are not related.

Finally two genes appear in both sets: TP53 and IHPK2. TP53 is perhaps the best known gene to be involved in cancer. Therefore, it is bound to appear in the TXTmodel and it is no surprise that it is also present in the UNImodel. IHPK2 however has a weak prior relation with prognosis in cancer; therefore, this gene proves that genes with a low text prior still can be selected in the TXTmodel. Additionally, genes which appear in both models can be considered more reliable.

6.10.2 Bild data

Finally we validated our approach on three independent data sets on breast, ovarian and lung cancer [27] to assess if the results on the Veer data set can be confirmed. Based on the results presented in Table 6.10 we chose a mean density of 1 for these data sets. Again 100 randomizations of the data set with and without the text prior were performed. Table 6.12 shows the average AUC for the three Bild data sets and confirms that the text prior significantly improves the prediction of the prognosis on independent data sets and for other cancer sites.

6.11 Conclusions

6.11.1 Microarray and text data for gene cluster profiling

A framework was described for advanced textual profiling of groups of genes. TXTGate is implemented as a web application designed to efficiently process queries of up to 200 genes, although this is not a strict limit. It can be assumed that the application scales well enough to be of use in, for example, microarray cluster validation.

Supported by the work of Stephens *et al.* [238] and more recently that of Chiang and Yu [44], it has been the aim to complement the limitations of a single, more general text index by offering different views. Nevertheless, some vocabularies could still be optimized to improve the information content of the profiles. For example, at the time of evaluation, some general or non-informative terms were still scoring high because of the stemming and phrase-detection methods used (for example, 'ii', 'protein', 'alpha').

Finally, although the citations in Entrez Gene and SGD constitute good sources for retrieving relevant gene-related MEDLINE abstracts, weighting the information according to the context and eliminating poorly informative or contaminating annotations (such as sequence-related articles) still need to be taken into account in future incarnations of the software. Document-classification strategies as in Leonard *et al.* [135] or Raychaudhuri *et al.* [185] can be adopted to this end.

As with GO annotations, transfer of literature references according to homology can be used to characterize poorly annotated genes [9, 184]. At this stage, the application allows for the study of homologs within all organisms contained in Entrez Gene, provided the user inputs the corresponding Entrez Gene identifiers. This type of operation will be increasingly supported with future additions of literature indices from other organisms and databases.

In conclusion, TXTGate's approach to summarizing database annotations and literature via specific vocabularies, along with its options to perform further analysis via clustering or query building, make it a flexible gateway to explore text-based information comprehensively.

6.11.2 Microarray and text data for Bayesian Networks

In these sections we have shown a method to integrate information from literature abstracts with gene expression data using Bayesian network models. This prior information was integrated in the prior distribution over the possible Bayesian network structures after scaling. The results of the randomization analysis in Table 6.10 and 6.12 have shown that for both the Veer data set and the three Bild data sets the text prior improves the prediction of the prognosis of cancer patients significantly.

A possible limitation of our approach is the discretization of the data. It is inevitable that some information is lost in the process of discretization. We have chosen discrete valued Bayesian networks because the space of arbitrary continuous distributions is large. A solution could be to restrict ourselves to the use of Gaussian Bayesian networks but this class of models assumes linear interactions between the variables which, in our opinion, would restrict too much the type of relations among genes that are modeled. Moreover, by using the algorithm of Hartemink we are performing a multivariate discretization, keeping the relationships between the variables as much as possible intact.

Secondly, by using text information, which is often described as highly biased, one could risk focussing too much on the hot genes disregarding novel important genes. However, in our case the emphasis is not so much on biomarker discovery and more on developing models which can accurately predict the prognosis of disease. There are already many genes known to be involved in different types of cancer based on individual studies or because they are member of a cancer profile. Finding the minimal set of genes which is able to predict the prognosis of disease, however, is still an open problem. Our Bayesian network framework attempts to address this issue by tackling the disadvantages of cancer microarray data sets (low signal-to-noise ratio,

high dimensional, small sample size, ...) by using information from the literature as a guide.

Finally, the presented framework is complimentary to our previously published method to integrate clinical and microarray data with Bayesian networks [80]. Thus creating a Bayesian network framework which enables modeling of various data sources (i.e. clinical, microarray and text) to improve decision support of outcome (i.e phenotypic group) prediction in cancer or other genetic diseases. Moreover, our definition of the structure prior makes no assumptions about the nature of prior information. Therefore other sources of information can be combined with the text prior (*e.g.*, known protein-DNA interactions from TRANSFAC, known pathways from KEGG or motif information). Thus, creating a white box framework that visualizes how decisions are made by a model. This is in contrast to for example a kernel framework where model parameters are not readily interpretable.

Chapter 7

Conclusions and perspectives

"A person is not an originating agent; he is a locus, a point at which many genetic and environmental conditions come together in a joint effect."

BF Skinner, 1974

7.1 Accomplishments and conclusions

In this dissertation, we have started with a discussion on management and processing of raw data coming out of Array CGH assays. We have discussed how quality criteria were set up during the early stage of the introduction of array technology into clinical practice. Guidelines on how to analyze and interpret array data prove essential to allow the technique to move from research into everyday routine. Next, we have discussed analysis tools that aim at supporting the raw data analysis and initial interpretation. We have seen how databases augmented with tools and custom algorithms support both research as well as diagnostic processes in genetic screening and discussed an alternative experiment design, and the statistical models that come with its implementation. We have concluded Chapter 2 with a discussion on cross-experiment data mining, its pitfalls, and its benefits in charting variation in the human genome.

In Chapter 3, we have forayed into functional interpretation of genetic information resulting from Array CGH assays, aiming at leveraging patient information (such as phenotype and clinical traits) in order to interpret copy number variation, as well as mining public data sources such as biomedical literature. This functional analysis aims at identifying candidate causal regions that play a role in congenital anomaly or developmental delay of the patient, as well as at supporting the study of aetiology of new syndromes characterized by submicroscopic chromosomal imbalances. We have initially focused on capturing and using knowledge embedded in literature sources, and went on discussing text as a data source in the broader context of data fusion as a tool for candidate gene prioritization.

As an extension on functional interpretation, Chapter 4 focused on the clinical side of data interpretation, which is directed at diagnosis decision support. We have highlighted the collaborative aspects of biomedical research and diagnostics, and the

need for aggregation of case information in database efforts. This chapter showed how the tools and approaches discussed in the previous chapter, together with consortium and database initiatives, can form a comprehensive platform that can address specific challenges in both genetic diagnosis as well as research in constitutional cytogenetics.

Databases that aggregate rare patient case details and intelligent tools that leverage this information for research and diagnostics form structured repositories of knowledge annotated to the genome. In order to facilitate new discoveries and clinical interpretation of assay data, a consensus on how to distribute this information as well as technical means to do so are needed. This is why in Chapter 5 we have introduced genome annotation, allowing continuously updated and amassed knowledge can be made available in a uniform, distributed, and standardized way to the research community. We have discussed the Distributed Annotation Server protocol and how it serves its purpose in context of the databases and tools discussed earlier. We have also discussed a case study for genome annotation with information on cis-regulatory elements mined from publicly available biomedical literature and curated data from a consortium effort.

Finally, in Chapter 6, we have concluded with a deeper discussion on the text mining aspects of the different approaches discussed earlier. As a first specific case, we have looked into fusing biomedical text profiles on genes with assay data from expression assays in order to help identify groups of genes that are jointly regulated and functionally related. Secondly, we have looked into leveraging biomedical text as a prior data source for Bayesian networks applied to outcome prediction in oncology and to elucidating gene regulation networks.

7.2 Future research and perspectives

In applied medical genetic research, the creation of algorithms and statistics as an outcome of scientific research needs to focus on use in a high throughput context, on real data, in an actual research or diagnostic environment. This means that software packages that result from such efforts need to be stable and robust, but also flexible and extensible. This, in turn, means the work is never finished. Throughout the dissertation, we have highlighted concrete possibilities for extension of the tools discussed, such as the future work on Clone Wars (Section 2.4.3), extending LOOP for high density arrays and integrating it in the Bench platform, finding smart ways of dealing with the constant change in genome assemblies, whether relating to coordinate positions, metadata and annotations, creating clone and reporter specific models to be able to better mine and interpret array CGH log ratio data, ... However, in this discussion, rather than to dwell on the technical progress that follows evolving bio-informatics tools for medical genetics, we will focus on a number of general perspectives to take into account.

7.2.1 Future proof platforms

The adoption of Array CGH in biomedical research and clinical practice is but a single step in a continuing evolution of genetic screening. Charting the genome of an individual such as a patient is done at an ever increasing resolution, the assay itself is performed in less and less time, and the cost per base pair is dropping year after year, in a way that reminds us of Moore's law [164]. Array CGH as a platform increases in resolution quarter after quarter, as commercial companies drive the technology

forward. The use of Array CGH is undoubtedly but a phase in the ongoing rush towards the thousand dollar genome, a concept pointing to the genome of a single individual being sequenced in a short time and at a reasonable cost. Already, so called next generation sequencing platforms are on the market and being used, mostly in a research context. While still expensive, sequencing runs are becoming cheaper and the thousand dollar genome is getting closer. Full affordable genome sequencing will open up a vast new field of personalized medicine, and will revolutionize clinical diagnostics and genetic research. While bio-informaticians and statisticians used to eagerly await biological data to produce and validate their research, they are now threatened by a deluge of data we will hardly be able to cope with. The most important challenge in bio-informatics applied to clinical genetics will be to deal with these many genomes, and the depth at which we will be able to look at them. This change of paradigm will impact the way diagnostics and research will be organized in the future (a deeper integration of bio-informatics teams and biomedical or clinical groups will need to take place, and in hospitals, wet lab technicians will be complemented by a larger amount of IT technicians than currently present), and it will change the way we write software (high performance, grid or cloud computing will become core strategies) and the way we mine data. Next generation sequencing data is about to wash ashore and it is time we prepare for small and big questions it brings with it: how do we prepare or extend existing platforms for next generation sequencing data? How will NGS impact the clinical interpretation of sequence data, when in a clinical context, even now, parts of the data are simply ignored for lack of unambiguous ways of clinical interpretation?

Related to this (r)evolution, the field of bio-informatics will need to accommodate the continued convergence of several fields involved in genetic screening: that of constitutional cytogenetics, where bands and now smaller regions in copy gain or loss are linked to developmental delay and congenital anomaly, the world of SNP assays and regions in loss of heterozygosity, information from specific targeted assays that characterize small regions of the genome in a highly detailed way (for example, BRCA and CFTR gene tests), the results from next generation sequencing, results from methylation assays that are growing in popularity and use, and although they have been around for a while, data from expression assays, and results form the huge scale whole genome association studies, that merge information on thousands of individuals. Bringing these different data sources together and jointly interpreting them as different kinds of genetic aberrations or diversions from a reference genome and the interpretation of these data in a functional genomics and population specific context will bring an interesting array of challenges, as will the determining of a reference genome itself, as pointed out by the Sanger Case Control Consortium efforts and the Thousand Genomes Project.

7.2.2 Collecting and representing the genome and phenome

Genome data is soon going to be ubiquitously present in research consortium repositories, public databases, patient health records, and so forth. In order to make this wealth of data amenable to mining, genome and phenome data will have to be stored according to consensus standards. Standards will have to be set not just to represent the genetic and phenotypic information itself, but also to establish at a later stage how it was obtained, how its quality can be ascertained, and what related meta-data is available on the individual and the population it stems from. The importance of the

origin of individuals will increase as local genetic differences will have to be charted in order to be able to predict disease susceptibility in a context of population background: we already see differences in loci of benign copy number variation with different origins. Local genetic diversity will be mappable thanks to large scale genome screening or even sequencing initiatives.

Data quality, and especially correct and accurate phenotyping of individuals is a key element in setting up such initiatives. Also, uniformly describing genome variation versus reference genomes and formally representing case control information are key to make this data not only universally available for research purposes but also amenable to data mining. Both NCBI and EBI are currently working on data representation efforts and ontologies are being developed to start solving this problem. In this work we have highlighted a number of problems related to the use of classification systems such as ontologies to annotate phenotype data. The phenotype of a patient is dynamic: it evolves over time (as the organism develops normally or as a disease progresses), phenotypes are not always visible (penetrance) and can be qualified as more or less severe. Only recently, an ontology initiative has been released into the public that addresses these issues. The HPO (Human Phenotype Ontology) captures over 9000 human phenotype traits and allows describing modifications on annotations such as onset and penetrance. Linked strongly into OMIM, provided in established computer readable formats, supported with several tools and mapping files, and updated frequently, HPO holds the promise to become a major phenotype ontology and is already being adopted by the community..

Apart from sound data representation, there is a clear need for aggregation of large volumes of uniformly formatted genotype and phenotype information in order to advance the elucidation of gene function in human development and genetic disease. Several large scale initiatives have started and are currently being set up to address the challenge of mapping the presence of copy number variation in the population to understand genome plasticity and to better understand the implications of genome variation for disease susceptibility, complex diseases and personal traits. Understanding the genetic basis of many disorders will only be possible through large scale association studies based on increasing volumes of patient and reference individual data.

Moreover, the boundaries between benign and pathogenic variation are becoming blurred, since certain imbalances are associated with developmental disorders in only a proportion of carriers. Whereas the higher resolution arrays were thought to have the prospect of increasing the diagnostic yield even further, they currently lead to the identification of a large amount of copy number variants of which the potential role in developmental disorders is currently unclear. These findings are the cause of confusion to both clinicians and patients and currently not only hamper but even threaten clinical genetics. The reason for the confusion is the absence of adequately characterized genomic maps of copy number variation. Considering that these copy number variants vary across the population, and some variants may be rare, reference genome-wide CNV maps need to be built from a large number of individuals derived from different populations. The challenge is to relate genome-wide CNV and SNP data with complex and frequent disorders such as congenital malformations and mental retardation.

The goal of such initiatives is clearly not just the advance of science and medicine in itself, but equally to advance diagnosis and prognosis, improve treatment, and aid the families involved.

7.2.3 Clinical validation

Drawing conclusions in silico for diagnosis support is acceptable practice for research purposes but when real patients are involved, clinical validation is key and FDA approval lurks around the corner for clinical and commercial applications. Black box algorithms such as advanced classification systems or complex and untransparent data fusion approaches will on the one hand pass regulatory approval processes with great difficulty, and on the other hand, tend to scare off practitioners because they 'cannot see inside the software', cannot understand why an algorithm makes a certain conclusion, and hence refuse to bear responsibility for the consequences of accepting a computer-supported or guided diagnosis. This problem reaches further than the mere need for sound clinical validation studies on information systems approaches on medical or genetic data. However, software and data mining can make the task of a clinical practitioner easier, can help avoid mistakes and bias, can guide the process of diagnosis, and can certainly remove the daunting task of analyzing very large amounts of data manually.

7.2.4 Towards personal genomics

Drug pipelines are drying up and the development of new drugs increases in cost. Strengthened by the advent of future generation genetic screening assays, the field of clinical genetics is going to extend towards and at least partially merge into pharmacogenomics. The question of interpretation of high resolution array CGH data and the ethical dilemma faced in patients is just a prelude to the daunting tasks ahead when full genome sequencing will be feasible.

7.2.5 The future of genome annotation

Population ageing will put medical insurance and reimbursement schemes under pressure as time goes on, and different ways of optimizing and making health care more efficient will be sought. A trend towards deeper informatization and health care systems integration is emerging and is an important aspect in reducing health care cost. This will in turn lead to a greater pressure towards standardization of health care systems: databases and health information systems will have to communicate better, and creating interoperable systems will have to become easier. Different standards have arisen to this goal and a convergence towards a few of these standards (such as the IHE initiative) seems ongoing, but a lot of work lies ahead: issues like informed consent, data ownership, data protection and security, public policy on privacy, ethical and legal aspects, the social aspects of genetic information as an input for prime calculation by insurance companies, open vs. closed consortia, but also technical aspects such as data point duplicates, unifying identifiers, interoperable communication standards, dealing with distributed and frequently updated data, consistent, qualitative and standardized knowledge representation, dealing with very large volumes of data over a networked infrastructure, revoking distributed information, data access (who gets to see what in what circumstances), and so on. Patient records are becoming global in member states of the European union as well as outside, the Google health initiative is growing fast, and companies like 23andMe have clearly seen a market in consumer genomics. However, the scientific, ethical and legal frameworks behind these initiatives and evolutions are still developing. A crucial aspect will be to inform and empower people that are confronted with genetic screening, and ensure that the

individual remains in control over his or her genetic data.

Bibliography

[1] E.A. Adie, R.R. Adams, K.L. Evans, D.J. Porteous, and B.S. Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6:55, 2005.

[2] E.A. Adie, R.R. Adams, K.L. Evans, D.J. Porteous, and B.S. Pickard. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22:773–774, Mar 2006.

[3] B. Adryan and S.A. Teichmann. FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster. *Bioinformatics*, 22:1532–1533, Jun 2006.

[4] S. Aerts, M. Haeussler, S. van Vooren, O.L. Griffith, P. Hulpiau, S.J. Jones, S.B. Montgomery, and C.M. Bergman. Text-mining assisted regulatory annotation. *Genome Biol.*, 9:R31, 2008.

[5] S Aerts, D Lambrechts, S Maity, P Van Loo, B Coessens, F De Smet, L-C Tranchevent, B De Moor, P Marynen, B Hassan, P Carmeliet, and Y Moreau. Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24(5):537–544, May 2006.

[6] F Al-Shahrour, R Diaz-Uriarte, and J Dopazo. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, Mar 2004.

[7] D.G. Albertson, B. Ylstra, R. Segraves, C. Collins, S.H. Dairkee, D. Kowbel, W.L. Kuo, J.W. Gray, and D. Pinkel. Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat. Genet.*, 25:144–146, Jun 2000.

[8] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. De Moor. Using literature and data to learn bayesian networks as clinical models of ovarian tumours. *Artif Intell Med*, 30:257–81, 2004.

[9] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25:25–29, May 2000.

[10] M. Ashburner and C.M. Bergman. Drosophila melanogaster: a case study of a model genomic sequence and its consequences. *Genome Res.*, 15:1661–1667, Dec 2005.

[11] L. Baala, S. Briault, H.C. Etchevers, F. Laumonnier, A. Natiq, J. Amiel, N. Boddaert, C. Picard, A. Sbiti, A. Asermouh, T. Attié-Bitach, F. Encha-Razavi, A. Munnich, A. Sefiani, and S. Lyonnet. Homozygous silencing of T-box transcription factor EOMES leads to microcephaly with polymicrogyria and corpus callosum agenesis. *Nat. Genet.*, 39:454–456, Apr 2007.

[12] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999.

[13] I. Balikova, K. Martens, C. Melotte, M. Amyere, S. Van Vooren, Y. Moreau, D. Vetrie, H. Fiegler, N.P. Carter, T. Liehr, M. Vikkula, G. Matthijs, J.P. Fryns, I. Casteels,

K. Devriendt, and J.R. Vermeesch. Autosomal-dominant microtia linked to five tandem copies of a copy-number-variable region at chromosome 4p16. *Am. J. Hum. Genet.*, 82:181–187, Jan 2008.

[14] C.A. Ball and A. Brazma. MGED standards: work in progress. *OMICS*, 10:138–144, 2006.

[15] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, 21(11):1337–1342, November 2003.

[16] Michael Baraitser and Robin Winter. *London Neurogenetics Database (LNDB) and Dysmorphology Photo Library*. Oxford University Press, 1999.

[17] B. Bardoni, M. Castets, M.E. Huot, A. Schenck, S. Adinolfi, F. Corbin, A. Pastore, E.W. Khandjian, and J.L. Mandel. 82-FIP, a novel FMRP (fragile X mental retardation protein) interacting protein, shows a cell cycle-dependent intracellular localization. *Hum. Mol. Genet.*, 12:1689–1698, Jul 2003.

[18] M. Baudis and M.L. Cleary. Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics*, 17:1228–1229, Dec 2001.

[19] B.A. Bejjani and L.G. Shaffer. Clinical utility of contemporary molecular cytogenetics. *Annu Rev Genomics Hum Genet*, 9:71–86, 2008.

[20] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*, 57:289–300, 1995.

[21] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler. GenBank. *Nucleic Acids Res*, Dec 2007.

[22] C.M. Bergman, J.W. Carlson, and S.E. Celniker. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics*, 21:1747–1749, Apr 2005.

[23] A. Bernard and A. J. Hartemink. Informative structure priors joint learning of dynamic regulatory networks from multiple types of data. *PSB*, 10:459–70, 2005.

[24] M. W. Berry, editor. *Computational information retrieval*. Society for Industrial and Applied Mathematics, 2001.

[25] M. W. Berry, editor. *Survey of Text Mining*. Springer-Verlag New York, Inc., 2003.

[26] M. W. Berry and M. Browne. *Understanding search engines: mathematical modeling and text retrieval*. Society for Industrial and Applied Mathematics, 1999.

[27] Andrea H. Bild, Guang Yao, Jeffrey T. Chang, Quanli Wang, Anil Potti, Dawn Chasse, Mary-Beth Joshi, David Harpole, Johnathan M. Lancaster, Andrew Berchuck, John A. Olson, Jeffrey R. Marks, Holly K. Dressman, Mike West, and Joseph R. Nevins. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, November 2005.

[28] E. Blanco, D. Farré, M.M. Albà, X. Messeguer, and R. Guigó. ABS: a database of Annotated regulatory Binding Sites from orthologous promoters. *Nucleic Acids Res.*, 34:D63–67, Jan 2006.

[29] C. Blaschke, J.C. Oliveros, and A. Valencia. Mining functional information associated with expression arrays. *Funct. Integr. Genomics*, 1:256–268, Mar 2001.

[30] J. Bogaerts, F. Cardoso, M. Buyse, S. Braga, S. Loi, J.A. Harrison, J. Bines, S. Mook, N. Decker, P. Ravdin, P. Therasse, E. Rutgers, L.J. van 't Veer, and M. Piccart. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. *Nat Clin Pract Oncol*, 3:540–551, Oct 2006.

[31] DS Borgaonkar. *Chromosomal variation in man; a catalogue of chromosomal variants and anomalies*. Wiley, 1997.

[32] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, 29:365–371, Dec 2001.

[33] J. Breckpot, Y. Takiyama, B. Thienpont, S. Van Vooren, J.R. Vermeesch, E. Ortibus, and K. Devriendt. A novel genomic disorder: a deletion of the SACS gene leading to Spastic Ataxia of Charlevoix-Saguenay. *Eur. J. Hum. Genet.*, Apr 2008.

[34] C. Brewer, S. Holloway, P. Zawalnyski, A. Schinzel, and D. FitzPatrick. A chromosomal deletion map of human malformations. *Am. J. Hum. Genet.*, 63(4):1153–1159, Oct 1998.

[35] C. Brewer, S. Holloway, P. Zawalnyski, A. Schinzel, and D. FitzPatrick. A chromosomal duplication map of malformations: regions of suspected haplo- and triplolethality–and tolerance of segmental aneuploidy–in humans. *Am J Hum Genet*, 64(6):1702–1708, Jun 1999.

[36] C.E. Bruder, C. Hirvela, I. Tapia-Paez, I. Fransson, R. Segraves, G. Hamilton, X.X. Zhang, D.G. Evans, A.J. Wallace, M.E. Baser, J. Zucman-Rossi, M. Hergersberg, E. Boltshauser, L. Papi, G.A. Rouleau, G. Poptodorov, A. Jordanova, H. Rask-Andersen, L. Kluwe, V. Mautner, M. Sainio, G. Hung, T. Mathiesen, C. Moeller, S.M. Pulst, H. Harder, A. Heiberg, M. Honda, M. Niimura, S. Sahlén, E. Blennow, D.G. Albertson, D. Pinkel, and J.P. Dumanski. High resolution deletion analysis of constitutional DNA from neurofibromatosis type 2 (NF2) patients using microarray-CGH. *Hum. Mol. Genet.*, 10:271–282, Feb 2001.

[37] Richard F. Burton. *Biology by Numbers: An Encouragement to Quantitative Thinking*. Cambridge University Press, 1998.

[38] M. Buyse, S. Loi, L. van't Veer, G. Viale, M. Delorenzi, A.M. Glas, M.S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. Harris, J. Bogaerts, P. Therasse, A. Floore, M. Amakrane, F. Piette, E. Rutgers, C. Sotiriou, F. Cardoso, and M.J. Piccart. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl. Cancer Inst.*, 98:1183–1192, Sep 2006.

[39] R Calogero, G Iazzetti, S Motta, G Pedrazzi, S Rago, E Rossi, and R Turra. Medmole: mining literature to extract biological knowledge by microarray data. *in proc Virtual Conf Genomics Bioinformatics*, 2:9–14, 2002.

[40] N.P. Carter, H. Fiegler, and J. Piper. Comparative analysis of comparative genomic hybridization microarray technologies: report of a workshop sponsored by the Wellcome Trust. *Cytometry*, 49:43–48, Oct 2002.

[41] C I Castillo-Davis and D L Hartl. GeneMerge–post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19(7):891–892, May 2003.

[42] D. Chaussabel and A. Sher. Mining microarray expression data by literature profiling. *Genome Biol.*, 3:RESEARCH0055, Sep 2002.

[43] B. Chi, R.J. DeLeeuw, B.P. Coe, C. MacAulay, and W.L. Lam. SeeGH–a software tool for visualization of whole genome array comparative genomic hybridization data. *BMC Bioinformatics*, 5:13, Feb 2004.

[44] J.H. Chiang and H.C. Yu. MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*, 19:1417–1422, Jul 2003.

[45] European Chromosome 16 Tuberous Sclerosis Consortium. Identification and characterization of the tuberous sclerosis gene on chromosome 16. *Cell*, 75:1305–1315, Dec 1993.

[46] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945, Oct 2004.

[47] G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9, 1992.

[48] J. Crim, R. McDonald, and F. Pereira. Automatically annotating documents with normalized gene lists. *BMC Bioinformatics*, 6 Suppl 1:S13, 2005.

[49] M.A. Crosby, J.L. Goodman, V.B. Strelets, P. Zhang, and W.M. Gelbart. FlyBase: genomes by the dozen. *Nucleic Acids Res.*, 35:D486–491, Jan 2007.

[50] de Ravel, T, Devriendt, K, Fryns, J. P., and Vermeesch, J. R. What's new in karyotyping? The move towards array comparative genomic hybridisation (CGH). *Eur J Pediatr*, 2007.

[51] B.B. de Vries, R. Pfundt, M. Leisink, D.A. Koolen, L.E. Vissers, I.M. Janssen, S. Reijmersdal, W.M. Nillesen, E.H. Huys, N. Leeuw, D. Smeets, E.A. Sistermans, T. Feuth, C.M. van Ravenswaaij-Arts, A.G. van Kessel, E.F. Schoenmakers, H.G. Brunner, and J.A. Veltman. Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.*, 77:606–616, Oct 2005.

[52] R W Doerge. Bioinformatics and computational biology solutions using R and bioconductor edited by Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. *Biometrics*, 62(4):1270–1271, Dec 2006.

[53] M.J. Donlin. Using the Generic Genome Browser (GBrowse). *Curr Protoc Bioinformatics*, Chapter 9:Unit 9.9, Mar 2007.

[54] R.D. Dowell, R.M. Jokerst, A. Day, S.R. Eddy, and L. Stein. The distributed annotation system. *BMC Bioinformatics*, 2:7, 2001.

[55] E.E. Eichler, D.A. Nickerson, D. Altshuler, A.M. Bowcock, L.D. Brooks, N.P. Carter, D.M. Church, A. Felsenfeld, M. Guyer, C. Lee, J.R. Lupski, J.C. Mullikin, J.K. Pritchard, J. Sebat, S.T. Sherry, D. Smith, D. Valle, and R.H. Waterston. Completing the map of human genetic variation. *Nature*, 447:161–165, May 2007.

[56] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95:14863–14868, Dec 1998.

[57] C.C. Elbers, N.C. Onland-Moret, L. Franke, A.G. Niehoff, Y.T. van der Schouw, and C. Wijmenga. A strategy to search for common obesity and type 2 diabetes genes. *Trends Endocrinol. Metab.*, 18:19–26, 2007.

[58] C.G. Elsik, K.C. Worley, L. Zhang, N.V. Milshina, H. Jiang, J.T. Reese, K.L. Childs, A. Venkatraman, C.M. Dickens, G.M. Weinstock, and R.A. Gibbs. Community annotation: procedures, protocols, and supporting tools. *Genome Res.*, 16:1329–1333, Nov 2006.

[59] J. Fantes, N.K. Ragge, S.A. Lynch, N.I. McGill, J.R. Collin, P.N. Howard-Peebles, C. Hayward, A.J. Vivian, K. Williamson, V. van Heyningen, and D.R. FitzPatrick. Mutations in SOX2 cause anophthalmia. *Nat. Genet.*, 33:461–463, Apr 2003.

[60] I. Feenstra, H.G. Brunner, and C.M. van Ravenswaaij. Cytogenetic genotype-phenotype studies: improving genotyping, phenotyping and data storage. *Cytogenet. Genome Res.*, 115:231–239, 2006.

[61] I Feenstra, J Fang, D A Koolen, A Siezen, C Evans, R M Winter, M M Lees, M Riegel, B B A de Vries, C M A Van Ravenswaaij, and A Schinzel. European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA); an online database for rare chromosome abnormalities. *Eur J Med Genet*, 49(4), 2006.

[62] L. Feuk, A.R. Carson, and S.W. Scherer. Structural variation in the human genome. *Nat. Rev. Genet.*, 7:85–97, Feb 2006.

[63] Heike Fiegler, Philippa Carr, Eleanor J Douglas, Deborah C Burford, Sarah Hunt, Carol E Scott, James Smith, David Vetrie, Patricia Gorman, Ian P M Tomlinson, and Nigel P Carter. DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer*, 36(4):361–374, Apr 2003.

[64] R.D. Finn, J.W. Stalker, D.K. Jackson, E. Kulesha, J. Clements, and R. Pettett. ProServer: a simple, extensible Perl DAS server. *Bioinformatics*, 23:1568–1570, Jun 2007.

[65] H. V. Firth, S. M. Richards, A. P. Bevan, S. Clayton, M. Corpas, D. Rajan, S. Van Vooren, Y. Moreau, R. M. Pettett, and N. P. Carter. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.*, 84:524–533, Apr 2009.

[66] D.R. FitzPatrick, I.M. Carr, L. McLaren, J.P. Leek, P. Wightman, K. Williamson, P. Gautier, N. McGill, C. Hayward, H. Firth, A.F. Markham, J.A. Fantes, and D.T. Bonthron. Identification of SATB2 as the cleft palate gene on 2q32-q33. *Hum. Mol. Genet.*, 12:2491–2501, Oct 2003.

[67] J. Flint, A.O. Wilkie, V.J. Buckle, R.M. Winter, A.J. Holland, and H.E. McDermid. The detection of subtelomeric chromosomal rearrangements in idiopathic mental retardation. *Nat. Genet.*, 9:132–140, Feb 1995.

[68] L. Franke, H. van Bakel, B. Diosdado, M. van Belzen, M. Wapenaar, and C. Wijmenga. TEAM: a tool for the integration of expression, and linkage and association maps. *Eur. J. Hum. Genet.*, 12:633–638, Aug 2004.

[69] L. Franke, H. van Bakel, L. Fokkens, E.D. de Jong, M. Egmont-Petersen, and C. Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, 78:1011–1025, Jun 2006.

[70] J.L. Freeman, G.H. Perry, L. Feuk, R. Redon, S.A. McCarroll, D.M. Altshuler, H. Aburatani, K.W. Jones, C. Tyler-Smith, M.E. Hurles, N.P. Carter, S.W. Scherer, and C. Lee. Copy number variation: new insights in genome diversity. *Genome Res.*, 16:949–961, Aug 2006.

[71] N. Freimer and C. Sabatti. The human phenome project. *Nat. Genet.*, 34:15–21, May 2003.

[72] J. Freudenberg and P. Propping. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18 Suppl 2:S110–115, 2002.

[73] J.M. Friedman, A. Baross, A.D. Delaney, A. Ally, L. Arbour, L. Armstrong, J. Asano, D.K. Bailey, S. Barber, P. Birch, M. Brown-John, M. Cao, S. Chan, D.L. Charest, N. Farnoud, N. Fernandes, S. Flibotte, A. Go, W.T. Gibson, R.A. Holt, S.J. Jones, G.C. Kennedy, M. Krzywinski, S. Langlois, H.I. Li, B.C. McGillivray, T. Nayar, T.J. Pugh, E. Rajcan-Separovic, J.E. Schein, A. Schnerch, A. Siddiqui, M.I. Van Allen, G. Wilson, S.L. Yong, F. Zahir, P. Eydoux, and M.A. Marra. Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation. *Am. J. Hum. Genet.*, 79:500–513, Sep 2006.

[74] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *J Comput Biol*, 7:601–20, 2000.

[75] S.M. Gallo, L. Li, Z. Hu, and M.S. Halfon. REDfly: a Regulatory Element Database for Drosophila. *Bioinformatics*, 22:381–383, Feb 2006.

[76] Michael Y. Galperin. The Molecular Biology Database Collection: 2006 update. *Nucl. Acids Res.*, 34(suppl 1):D3–5, 2006.

[77] K.J. Gaulton, K.L. Mohlke, and T.J. Vision. A computational system to select candidate genes for complex human traits. *Bioinformatics*, 23:1132–1140, May 2007.

[78] R.A. George, J.Y. Liu, L.L. Feng, R.J. Bryson-Richardson, D. Fatkin, and M.A. Wouters. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res.*, 34:e130, 2006.

[79] O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Van Huffel, Y. Moreau, D. Timmerman, B. De Moor, and G. Condous. Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. *Human reproduction*, 21, 2006.

[80] O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Proceedings of the 14th Annual ISMB, Bioinformatics special issue*, 2006.

[81] O. Gevaert, S. Van Vooren, and B. De Moor. A framework for elucidating regulatory networks based on prior information and expression data. *Ann. N. Y. Acad. Sci.*, 1115:240–248, Dec 2007.

[82] O. Gevaert, S. Van Vooren, and B. de Moor. Integration of microarray and textual data improves the prognosis prediction of breast, lung and ovarian cancer patients. *Pac Symp Biocomput*, pages 279–290, 2008.

[83] D. Ghosh. Object-oriented transcription factors database (ooTFD). *Nucleic Acids Res.*, 28:308–310, Jan 2000.

[84] P. Glenisson. *Integrating scientific literature with large scale gene expression analysis, PhD thesis, Faculty of Engineering*. Katholieke Universiteit Leuven, 2004.

[85] P Glenisson, P Antal, J Mathys, Y Moreau, and B De Moor. Evaluation of the vector space representation in text-based gene clustering. *in proc of Pac Symp Biocomput*, pages 391–402, 2003.

[86] P. Glenisson, B. Coessens, S. Van Vooren, J. Mathys, Y. Moreau, and B. De Moor. TXTGate: profiling gene groups with text-based information. *Genome Biol.*, 5:R43, 2004.

[87] K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, and A.L. Barabsi. The human disease network. *Proc. Natl. Acad. Sci. U.S.A.*, 104:8685–8690, May 2007.

[88] J. Greshock. DNA copy number data analysis using the CGHAnalyzer software suite. *Methods Mol. Biol.*, 396:255–266, 2007.

[89] O.L. Griffith, S.B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M.C. Sleumer, M. Bilenky, M. Haeussler, M. Griffith, S.M. Gallo, B. Giardine, B. Hooghe, P. Van Loo, E. Blanco, A. Ticoll, S. Lithwick, E. Portales-Casamar, I.J. Donaldson, G. Robertson, C. Wadelius, P. De Bleser, D. Vlieghe, M.S. Halfon, W. Wasserman, R. Hardison, C.M. Bergman, and S.J. Jones. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res*, Nov 2007.

[90] P. Groth, N. Pavlova, I. Kalev, S. Tonov, G. Georgiev, H.D. Pohlenz, and B. Weiss. PhenomicDB: a new cross-species genotype/phenotype resource. *Nucleic Acids Res.*, 35:D696–699, Jan 2007.

[91] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. Cambridge, MA: MIT Press, 2001.

[92] J.L. Hand, V.V. Michels, M.J. Marinello, R.P. Ketterling, and S.M. Jalal. Inherited interstitial deletion of chromosomes 5p and 16q without apparent phenotypic effect: further confirmation. *Prenat. Diagn.*, 20:144–148, Feb 2000.

[93] D. Hanisch, K. Fundel, H.T. Mevissen, R. Zimmer, and J. Fluck. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 Suppl 1:S14, 2005.

[94] C.T. Harbison, D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, J.B. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K. Pokholok, M. Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, and R.A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, Sep 2004.

[95] A. J. Hartemink. *Principled computational methods for the validation and discovery of genetic regulatory networks*. PhD thesis, MIT, 2001.

[96] E Hatcher and O Gospodnetić. *Lucene in Action*. Manning Publications Co., 2004.

[97] M.A. Hauser, Y.J. Li, S. Takeuchi, R. Walters, M. Noureddine, M. Maready, T. Darden, C. Hulette, E. Martin, E. Hauser, H. Xu, D. Schmechel, J.E. Stenger, F. Dietrich, and J. Vance. Genomic convergence: identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage. *Hum. Mol. Genet.*, 12:671–677, Mar 2003.

[98] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[99] S.J. Ho Sui, J.R. Mortimer, D.J. Arenillas, J. Brumm, C.J. Walsh, B.P. Kennedy, and W.W. Wasserman. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res.*, 33:3154–3164, 2005.

[100] R Hoffmann, J Dopazo, JC Cigudosa, and A Valencia. HCAD, closing the gap between breakpoints and genes. *Nucleic Acids Res*, 33(Database Issue):D511–3, Jan 2005.

[101] R. Hoffmann and A. Valencia. A gene network for navigating the literature. *Nat. Genet.*, 36:664, Jul 2004.

[102] R Hoffmann and A Valencia. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21 Suppl 2:ii252–ii258, Sep 2005.

[103] D. Hristovski, B. Peterlin, J.A. Mitchell, and S.M. Humphrey. Using literature-based discovery to identify disease candidate genes. *Int J Med Inform*, 74:289–298, Mar 2005.

[104] Y. Hu, L.M. Hines, H. Weng, D. Zuo, M. Rivera, A. Richardson, and J. LaBaer. Analysis of genomic and proteomic data using advanced literature mining. *J. Proteome Res.*, 2:405–412, 2003.

[105] T. Hubbard. Biological information: making it accessible and integrated (and trying to make sense of it). *Bioinformatics*, 18 Suppl 2:S140, 2002.

[106] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. The Ensembl genome database project. *Nucleic Acids Res.*, 30:38–41, Jan 2002.

[107] T.J. Hubbard, B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S.C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X.M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Res.*, 35:D610–617, Jan 2007.

[108] A. Hunter, M.H. Kaufman, A. McKay, R. Baldock, M.W. Simmen, and J.B. Bard. An ontology of human developmental anatomy. *J. Anat.*, 203:347–355, Oct 2003.

[109] A.J. Iafrate, L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, and C. Lee. Detection of large-scale variation in the human genome. *Nat. Genet.*, 36:949–951, Sep 2004.

[110] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1, 2002.

[111] A S Ishkanian, C A Malloff, S K Watson, R J DeLeeuw, B Chi, B P Coe, A Snijders, D G Albertson, D Pinkel, M A Marra, V Ling, C MacAulay, and W L Lam. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*, 36(3):299–303, 2004.

[112] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[113] A.M. Jenkinson, M. Albrecht, E. Birney, H. Blankenburg, T. Down, R.D. Finn, H. Hermjakob, T.J. Hubbard, R.C. Jimenez, P. Jones, A. Khri, E. Kulesha, J.R. Macas, G.A. Reeves, and A. Prlic. Integrating biological data - the Distributed Annotation System. *BMC Bioinformatics*, 9 Suppl 8:S3, Jul 2008.

[114] T.K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, 28:21–28, May 2001.

[115] G. Jimenez-Sanchez, B. Childs, and D. Valle. Human disease genes. *Nature*, 409:853–855, Feb 2001.

[116] P. Jones, N. Vinod, T. Down, A. Hackmann, A. Kahari, E. Kretschmann, A. Quinn, D. Wieser, H. Hermjakob, and R. Apweiler. Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics*, 21:3198–3199, Jul 2005.

[117] K. Jong, E. Marchiori, G. Meijer, A.V. Vaart, and B. Ylstra. Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, 20:3636–3637, Dec 2004.

[118] M. I. Jordan. *Learning in Graphical Models*. MIT Press, 1998.

[119] K. Kas, M.L. Voz, E. Rijer, A.K. Astrm, E. Meyen, G. Stenman, and W.J. Van de Ven. Promoter swapping between the genes for a novel zinc finger protein and beta-catenin in pleiomorphic adenomas with t(3;8)(p21;q12) translocations. *Nat. Genet.*, 15:170–174, Feb 1997.

[120] J. Kelso, J. Visagie, G. Theiler, A. Christoffels, S. Bardien, D. Smedley, D. Otgaar, G. Greyling, C.V. Jongeneel, M.I. McCarthy, T. Hide, and W. Hide. eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.*, 13(6A):1222–1230, Jun 2003.

[121] P Khatri and S Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, Sep 2005.

[122] E. Klopocki, H. Schulze, G. Strauss, C.E. Ott, J. Hall, F. Trotier, S. Fleischhauer, L. Greenhalgh, R.A. Newbury-Ecob, L.M. Neumann, R. Habenicht, R. Knig, E. Seemanova, A. Megarbane, H.H. Ropers, R. Ullmann, D. Horn, and S. Mundlos. Complex inheritance pattern resembling autosomal recessive inheritance involving a microdeletion in thrombocytopenia-absent radius syndrome. *Am. J. Hum. Genet.*, 80:232–240, Feb 2007.

[123] S.J. Knight, R. Regan, A. Nicod, S.W. Horsley, L. Kearney, T. Homfray, R.M. Winter, P. Bolton, and J. Flint. Subtle chromosomal rearrangements in children with unexplained mental retardation. *Lancet*, 354:1676–1681, Nov 1999.

[124] D.A. Koolen, L.E. Vissers, R. Pfundt, N. de Leeuw, S.J. Knight, R. Regan, R.F. Kooy, E. Reyniers, C. Romano, M. Fichera, A. Schinzel, A. Baumer, B.M. Anderlid, J. Schoumans, N.V. Knoers, A.G. van Kessel, E.A. Sistermans, J.A. Veltman, H.G. Brunner, and B.B. de Vries. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat. Genet.*, 38:999–1001, Sep 2006.

[125] J O Korbel, T Doerks, L J Jensen, C Perez-Iratxeta, S Kaczanowski, S D Hooper, M A Andrade, and P Bork. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol*, 3(5):e134, May 2005.

[126] J.O. Korbel, A.E. Urban, J.P. Affourtit, B. Godwin, F. Grubert, J.F. Simons, P.M. Kim, D. Palejev, N.J. Carriero, L. Du, B.E. Taillon, Z. Chen, A. Tanzer, A.C. Saunders, J. Chi, F. Yang, N.P. Carter, M.E. Hurles, S.M. Weissman, T.T. Harkins, M.B. Gerstein, M. Egholm, and M. Snyder. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318:420–426, Oct 2007.

[127] R. Kosaki, K. Kosaki, K. Matsushima, N. Mitsui, N. Matsumoto, and H. Ohashi. Refining chromosomal region critical for Down syndrome-related heart defects with a case of cryptic 21q22.2 duplication. *Congenit Anom (Kyoto)*, 45(2):62–64, Jun 2005.

[128] I.D. Krantz, R. Smith, R.P. Colliton, H. Tinkel, E.H. Zackai, D.A. Piccoli, E. Goldmuntz, and N.B. Spinner. Jagged1 mutations in patients ascertained with isolated congenital heart defects. *Am. J. Med. Genet.*, 84:56–60, May 1999.

[129] K. Lage, E.O. Karlberg, Z.M. Strling, P.I. Olason, A.G. Pedersen, O. Rigina, A.M. Hinsby, Z. Tmer, F. Pociot, N. Tommerup, Y. Moreau, and S. Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, 25:309–316, Mar 2007.

[130] G R G Lanckriet, T De Bie, N Cristianini, M I Jordan, and W S Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[131] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, P. de Jong, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y.J. Chen, and J. Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, Feb 2001.

[132] C Le Caignec, C Spits, K Sermon, M De Rycke, B Thienpont, S Debrock, C Staessen, Y Moreau, J-P Fryns, A Van Steirteghem, I Liebaers, and J R Vermeesch. Single-cell chromosomal imbalances detection by array CGH. *Nucleic Acids Res*, 34(9):e68, 2006.

[133] Jason S M Lee, Gurpreet Katari, and Ravi Sachidanandam. GObar: a gene ontology based analysis and visualization tool for gene sets. *BMC Bioinformatics*, 6:189, 2005.

[134] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J.B. Tagne, T.L. Volkert, E. Fraenkel, D.K. Gifford, and R.A. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298:799–804, Oct 2002.

[135] J.E. Leonard, J.B. Colombe, and J.L. Levy. Finding relevant references to genes and proteins in Medline using a Bayesian approach. *Bioinformatics*, 18:1515–1522, Nov 2002.

[136] G. Levan. Nomenclature for G-bands in rat chromosomes. *Hereditas*, 77:37–52, 1974.

[137] H. Liu, Z.Z. Hu, M. Torii, C. Wu, and C. Friedman. Quantitative assessment of dictionary-based protein named entity tagging. *J Am Med Inform Assoc*, 13:497–507, 2006.

[138] D.P. Locke, A.J. Sharp, S.A. McCarroll, S.D. McGrath, T.L. Newman, Z. Cheng, S. Schwartz, D.G. Albertson, D. Pinkel, D.M. Altshuler, and E.E. Eichler. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.*, 79:275–290, Aug 2006.

[139] Núria López-Bigas and Christos A Ouzounis. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.*, 32:3108–3114, 2004.

[140] X. Lu, C.A. Shaw, A. Patel, J. Li, M.L. Cooper, W.R. Wells, C.M. Sullivan, T. Sahoo, S.A. Yatsenko, C.A. Bacino, P. Stankiewicz, Z. Ou, A.C. Chinault, A.L. Beaudet, J.R. Lupski, S.W. Cheung, and P.A. Ward. Clinical implementation of chromosomal microarray analysis: summary of 2513 postnatal cases. *PLoS ONE*, 2:e327, 2007.

[141] R. Lucito, J. Healy, J. Alexander, A. Reiner, D. Esposito, M. Chi, L. Rodgers, A. Brady, J. Sebat, J. Troge, J.A. West, S. Rostan, K.C. Nguyen, S. Powers, K.Q. Ye, A. Olshen, E. Venkatraman, L. Norton, and M. Wigler. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.*, 13:2291–2305, Oct 2003.

[142] J.R. Lupski. Structural variation in the human genome. *N. Engl. J. Med.*, 356:1169–1171, Mar 2007.

[143] J.R. Lupski, R.M. de Oca-Luna, S. Slaugenhaupt, L. Pentao, V. Guzzetta, B.J. Trask, O. Saucedo-Cardenas, D.F. Barker, J.M. Killian, C.A. Garcia, A. Chakravarti, and P.I. Patel. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*, 66:219–232, Jul 1991.

[144] J.R. Lupski and P. Stankiewicz. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet.*, 1:e49, Dec 2005.

[145] Y. Lussier, T. Borlawsky, D. Rappaport, Y. Liu, and C. Friedman. PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. *Pac Symp Biocomput*, pages 64–75, 2006.

[146] Y A Lussier and Y Liu. Computational approaches to phenotyping: high-throughput phenomics. *Proc Am Thorac Soc*, 4(1):18–25, 2007.

[147] X. Ma, H. Lee, L. Wang, and F. Sun. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics*, 23:215–221, Jan 2007.

[148] N.M. Maas, S. Van Vooren, F. Hannes, G. Van Buggenhout, M. Mysliwiec, Y. Moreau, K. Fagan, A. Midro, O. Engiz, S. Balci, M.J. Parker, Y. Sznajer, K. Devriendt, J.P. Fryns, and J.R. Vermeesch. The t(4;8) is mediated by homologous recombination between olfactory receptor gene clusters, but other 4p16 translocations occur at random. *Genet. Couns.*, 18:357–365, 2007.

[149] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[150] K.K. Mantripragada, I. Tapia-Pez, E. Blennow, P. Nilsson, A. Wedell, and J.P. Dumanski. DNA copy-number analysis of the 22q11 deletion-syndrome region using array-CGH with genomic and PCR-based targets. *Int. J. Mol. Med.*, 13:273–279, Feb 2004.

[151] A.A. Margolin, J. Greshock, T.L. Naylor, Y. Mosse, J.M. Maris, G. Bignell, A.I. Saeed, J. Quackenbush, and B.L. Weber. CGHAnalyzer: a stand-alone software package for cancer genome analysis using array-based DNA copy number data. *Bioinformatics*, 21:3308–3311, Aug 2005.

[152] C.L. Martin, D.J. Waggoner, A. Wong, S. Uhrig, J.A. Roseberry, J.F. Hedrick, S.D. Pack, K. Russell, E. Zackai, W.B. Dobyns, and D.H. Ledbetter. "Molecular rulers" for calibrating phenotypic effects of telomere imbalance. *J. Med. Genet.*, 39:734–740, Oct 2002.

[153] D Martin, C Brun, E Remy, P Mouren, D Thieffry, and B Jacq. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol*, 5(12):R101, 2004. Comparative Study.

[154] M Masseroli, O Galati, and F Pinciroli. GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Res*, 33(Web Server issue):717–723, Jul 2005.

[155] D.R. Masys, J.B. Welsh, J. Lynn Fink, M. Gribskov, I. Klacansky, and J. Corbeil. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17:319–326, Apr 2001.

[156] V. Matys, O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34:D108–110, Jan 2006.

[157] R. McDonald and F. Pereira. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, 6 Suppl 1:S6, 2005.

[158] B. Menten, N. Maas, B. Thienpont, K. Buysse, J. Vandesompele, C. Melotte, T. de Ravel, S. Van Vooren, I. Balikova, L. Backx, S. Janssens, A. De Paepe, B. De Moor, Y. Moreau, P. Marynen, J.P. Fryns, G. Mortier, K. Devriendt, F. Speleman, and J.R. Vermeesch. Emerging patterns of cryptic chromosomal imbalance in patients with idiopathic mental retardation and multiple congenital anomalies: a new series of 140 patients and review of published reports. *J. Med. Genet.*, 43:625–633, Aug 2006.

[159] B. Menten, F. Pattyn, K. De Preter, P. Robbrecht, E. Michels, K. Buysse, G. Mortier, A. De Paepe, S. Van Vooren, J. Vermeesch, Y. Moreau, B. De Moor, S. Vermeulen, F. Speleman, and J. Vandesompele. arrayCGHbase: an analysis platform for comparative genomic hybridization microarrays. *BMC Bioinformatics*, 6:124, 2005.

[160] J.A. Mitchell, A.R. Aronson, J.G. Mork, L.C. Folk, S.M. Humphrey, and J.M. Ward. Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annu Symp Proc*, pages 460–464, 2003.

[161] T. Mitsumori, S. Fation, M. Murata, K. Doi, and H. Doi. Gene/protein name recognition based on support vector machine using dictionary as features. *BMC Bioinformatics*, 6 Suppl 1:S8, 2005.

[162] Suri Mohnish. Oxford Medical Databases: London Dysmorphology Database Version 3.0. *J Med Genet*, 39:782–783, 2002.

[163] S.B. Montgomery, O.L. Griffith, M.C. Sleumer, C.M. Bergman, M. Bilenky, E.D. Pleasance, Y. Prychyna, X. Zhang, and S.J. Jones. ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, 22:637–640, Mar 2006.

[164] Gordon E. Moore. Cramming more components onto integrated circuits. *Electronics Magazine*, 4, 1965.

[165] A.A. Morgan, B. Wellner, J.B. Colombe, R. Arens, M.E. Colosimo, and L. Hirschman. Evaluating the automatic mapping of human gene and protein mentions to unique identifiers. *Pac Symp Biocomput*, pages 281–291, 2007.

[166] C.L. Myers, M.J. Dunham, S.Y. Kung, and O.G. Troyanskaya. Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, 20:3533–3543, Dec 2004.

[167] N. Nariai, S. Kim, S. Imoto, and S. Miyano. Using Protein-Protein Interactions for Refining Gene Networks Estimated from Microarray Data by Bayesian Networks . *PSB*, 9:336–347, 2004.

[168] M. K. Oehler, D. C. Fischer, M. Orlowska-Volk, F. Herrle, D. G. Kieback, M. C. Rees, and R. Bicknell. Tissue and plasma expression of the angiogenic peptide adrenomedullin in breast cancer. *Br J Cancer*, 89(10):1927–1933, November 2003.

[169] M. Oti and H.G. Brunner. The modular nature of genetic diseases. *Clin. Genet.*, 71:1–11, Jan 2007.

[170] M. Oti, B. Snel, M.A. Huynen, and H.G. Brunner. Predicting disease genes using protein-protein interactions. *J. Med. Genet.*, 43:691–698, Aug 2006.

[171] J. L. Owen, V. Iragavarapu-Charyulu, and D. M. Lopez. T cell-derived matrix metalloproteinase-9 in breast cancer: friend or foe? *Breast Dis*, 20:145–153, 2004.

[172] Q.S. Padiath, K. Saigoh, R. Schiffmann, H. Asahara, T. Yamada, A. Koeppen, K. Hogan, L.J. Ptácek, and Y.H. Fu. Lamin B1 duplications cause autosomal dominant leukodystrophy. *Nat. Genet.*, 38:1114–1123, Oct 2006.

[173] E. Pasmant, A. de Saint-Trivier, I. Laurendeau, A. Dieux-Coeslier, B. Parfait, M. Vidaud, D. Vidaud, and I. Biche. Characterization of a 7.6-Mb germline deletion encompassing the NF1 locus and about a hundred genes in an NF1 contiguous gene syndrome patient. *Eur. J. Hum. Genet.*, Jul 2008.

[174] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Matteo, California, 1988.

[175] C. Perez-Iratxeta, P. Bork, and M.A. Andrade. Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, 31:316–319, Jul 2002.

[176] C. Perez-Iratxeta, M. Wjst, P. Bork, and M.A. Andrade. G2D: a tool for mining genes associated with disease. *BMC Genet.*, 6:45, 2005.

[177] D Pinkel and D G Albertson. Comparative genomic hybridization. *Annu Rev Genomics Hum Genet*, 6:331–54, 2005.

[178] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.L. Kuo, C. Chen, Y. Zhai, S.H. Dairkee, B.M. Ljung, J.W. Gray, and D.G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, 20:207–211, Oct 1998.

[179] J.R. Pollack, C.M. Perou, A.A. Alizadeh, M.B. Eisen, A. Pergamenschikov, C.F. Williams, S.S. Jeffrey, D. Botstein, and P.O. Brown. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, 23:41–46, Sep 1999.

[180] E. Portales-Casamar, S. Kirov, J. Lim, S. Lithwick, M.I. Swanson, A. Ticoll, J. Snoddy, and W.W. Wasserman. PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol*, 8:R207, Sep 2007.

[181] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[182] A. Prlić, T.A. Down, E. Kulesha, R.D. Finn, A. Kahari, and T.J. Hubbard. Integrating sequence and structural biology with DAS. *BMC Bioinformatics*, 8:333, 2007.

[183] K.D. Pruitt, K.S. Katz, H. Sicotte, and D.R. Maglott. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.*, 16:44–47, Jan 2000.

[184] S. Raychaudhuri, J.T. Chang, F. Imam, and R.B. Altman. The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res.*, 31:4553–4560, Aug 2003.

[185] S. Raychaudhuri, J.T. Chang, P.D. Sutphin, and R.B. Altman. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.*, 12:203–214, Jan 2002.

[186] S. Raychaudhuri, H. Schuetze, and R.B. Altman. Using text analysis to identify functionally coherent gene groups. *Genome Res.*, 12:1582–1590, Oct 2002.

[187] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14:656–664, 1998.

[188] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno. Text processing through Web services: calling Whatizit. *Bioinformatics*, 24:296–298, Jan 2008.

[189] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr. EBIMed–text crunching to gather facts for proteins from Medline. *Bioinformatics*, 23:e237–244, Jan 2007.

[190] R. Redon, S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, A.R. Carson, W. Chen, E.K. Cho, S. Dallaire, J.L. Freeman, J.R. Gonzlez, M. Gratacs, J. Huang, D. Kalaitzopoulos, D. Komura, J.R. MacDonald, C.R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M.J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D.F. Conrad, X. Estivill, C. Tyler-Smith, N.P. Carter, H. Aburatani, C. Lee, K.W. Jones, S.W. Scherer, and M.E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, 2006.

[191] J.S. Reece-Hoyes, B. Deplancke, J. Shingles, C.A. Grove, I.A. Hope, and A.J. Walhout. A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol.*, 6:R110, 2005.

[192] S.B. Rice, G. Nenadic, and B.J. Stapley. Mining protein function from text using term-based support vector machines. *BMC Bioinformatics*, 6 Suppl 1:S22, 2005.

[193] M. Riegel, A. Baumer, M. Jamar, K. Delbecque, C. Herens, A. Verloes, and A. Schinzel. Submicroscopic terminal deletions and duplications in retarded patients with unclassified malformation syndromes. *Hum. Genet.*, 109:286–294, Sep 2001.

[194] W.P. Robinson, J. Waslynka, F. Bernasconi, M. Wang, S. Clark, D. Kotzot, and A. Schinzel. Delineation of 7q11.2 deletions associated with Williams-Beuren syndrome and mapping of a repetitive sequence to within and to either side of the common deletion. *Genomics*, 34(1):17–23, May 1996.

[195] C. Rodrìguez-Penagos, H. Salgado, I. Martìnez-Flores, and J. Collado-Vides. Automatic reconstruction of a bacterial regulatory network using Natural Language Processing. *BMC Bioinformatics*, 8:293, 2007.

[196] C. Rosenberg, J. Knijnenburg, E. Bakker, A.M. Vianna-Morgante, W. Sloos, P.A. Otto, M. Kriek, K. Hansson, A.C. Krepischi-Santos, H. Fiegler, N.P. Carter, E.K. Bijlsma, A. van Haeringen, K. Szuhai, and H.J. Tanke. Array-CGH detection of micro rearrangements in mentally retarded individuals: clinical significance of imbalances present both in affected children and normal parents. *J. Med. Genet.*, 43:180–186, Feb 2006.

[197] M.J. Rosenberg, C. Killoran, L. Dziadzio, S. Chang, D.L. Stone, J. Meck, D. Aughton, L.M. Bird, J. Bodurtha, S.B. Cassidy, J.M. Graham, A. Grix, A.E. Guttmacher, L. Hudgins, C. Kozma, R.C. Michaelis, R. Pauli, K.F. Peters, K.N. Rosenbaum, C.J. Tifft, D. Wargowski, M.S. Williams, and L.G. Biesecker. Scanning for telomeric deletions and duplications and uniparental disomy using genetic markers in 120 children with malformations. *Hum. Genet.*, 109:311–318, Sep 2001.

[198] S Rossi, D Masotti, C Nardini, E Bonora, G Romeo, E Macii, L Benini, and S Volinia. Tom: a web-based integrated approach for identification of candidate disease genes. *Nucleic acids research*, 34(Web Server issue).

[199] A. Rovelet-Lecrux, D. Hannequin, G. Raux, N. Le Meur, A. Laquerrière, A. Vital, C. Dumanchin, S. Feuillette, A. Brice, M. Vercelletto, F. Dubas, T. Frebourg, and D. Campion. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat. Genet.*, 38:24–26, Jan 2006.

[200] Falcon S and Gentleman R. Using GOstats to test gene lists for GO term association. *Bioinformatics*, Nov 2006.

[201] L.H. Saal, C. Troein, J. Vallon-Christersson, S. Gruvberger, A. Borg, and C. Peterson. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol.*, 3:SOFTWARE0003, Jul 2002.

[202] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[203] D. Sanlaville, J.M. Lapierre, C. Turleau, A. Coquin, G. Borck, L. Colleaux, M. Vekemans, and S.P. Romana. Molecular karyotyping in human constitutional cytogenetics. *Eur J Med Genet*, 48:214–231, 2005.

[204] J. Saric, L.J. Jensen, R. Ouzounova, I. Rojas, and P. Bork. Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*, 22:645–650, Mar 2006.

[205] J. Saric, L.J. Jensen, and I. Rojas. Large-scale extraction of gene regulation for model organisms in an ontological context. *In Silico Biol. (Gedrukt)*, 5:21–32, 2005.

[206] A.J. Schaeffer, J. Chung, K. Heretis, A. Wong, D.H. Ledbetter, and C. Lese Martin. Comparative genomic hybridization-array analysis enhances the detection of aneuploidies and submicroscopic imbalances in spontaneous miscarriages. *Am. J. Hum. Genet.*, 74:1168–1174, Jun 2004.

[207] A. Schinzel. *Catalogue of unbalanced chromosome aberration in man.* de Gruyter, 2001.

[208] J. Schoumans, B.M. Anderlid, E. Blennow, B.T. Teh, and M. Nordenskjld. The performance of CGH array for the detection of cryptic constitutional chromosome imbalances. *J. Med. Genet.*, 41:198–202, Mar 2004.

[209] J. Schoumans, C. Ruivenkamp, E. Holmberg, M. Kyllerman, B.M. Anderlid, and M. Nordenskjld. Detection of chromosomal imbalances in children with idiopathic mental retardation by array based comparative genomic hybridisation (array-CGH). *J. Med. Genet.*, 42:699–705, Sep 2005.

[210] C. Schwaenen, M. Nessling, S. Wessendorf, T. Salvi, G. Wrobel, B. Radlwimmer, H.A. Kestler, C. Haslinger, S. Stilgenbauer, H. Dhner, M. Bentz, and P. Lichter. Automated array-based genomic profiling in chronic lymphocytic leukemia: development of a clinical tool and discovery of recurrent genomic alterations. *Proc. Natl. Acad. Sci. U.S.A.*, 101:1039–1044, Jan 2004.

[211] Cynthia L Sears. A dynamic partnership: celebrating our gut flora. *Anaerobe*, 11(5):247–251, Oct 2005.

[212] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y.H. Lee, J. Hicks, S.J. Spence, A.T. Lee, K. Puura, T. Lehtimki, D. Ledbetter, P.K. Gregersen, J. Bregman, J.S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M.C. King, D. Skuse, D.H. Geschwind, T.C. Gilliam, K. Ye, and M. Wigler. Strong association of de novo copy number mutations with autism. *Science*, 316:445–449, Apr 2007.

[213] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Manér, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T.C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler. Large-scale copy number polymorphism in the human genome. *Science*, 305:525–528, Jul 2004.

[214] R.F. Service. Gene sequencing. The race for the 1000 dollar genome. *Science*, 311:1544–1546, Mar 2006.

[215] L G Shaffer and B A Bejjani. Medical applications of array CGH and the transformation of clinical cytogenetics. *Cytogenet Genome Res*, 115(3-4):303–9, 2006.

[216] L.G. Shaffer and N. Tommerup. *ISCN 2005.* Karger, 2005.

[217] P K Shah, C Perez-Iratxeta, P Bork, and M A Andrade. Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, 4:20, May 2003. Evaluation Studies.

[218] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13:2498–2504, Nov 2003.

[219] A.J. Sharp, S. Hansen, R.R. Selzer, Z. Cheng, R. Regan, J.A. Hurst, H. Stewart, S.M. Price, E. Blair, R.C. Hennekam, C.A. Fitzpatrick, R. Segraves, T.A. Richmond, C. Guiver, D.G. Albertson, D. Pinkel, P.S. Eis, S. Schwartz, S.J. Knight, and E.E. Eichler. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.*, 38:1038–1042, Sep 2006.

[220] A.J. Sharp, D.P. Locke, S.D. McGrath, Z. Cheng, J.A. Bailey, R.U. Vallente, L.M. Pertz, R.A. Clark, S. Schwartz, R. Segraves, V.V. Oseroff, D.G. Albertson, D. Pinkel, and E.E. Eichler. Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.*, 77:78–88, Jul 2005.

[221] A.J. Sharp, R.R. Selzer, J.A. Veltman, S. Gimelli, G. Gimelli, P. Striano, A. Coppola, R. Regan, S.M. Price, N.V. Knoers, P.S. Eis, H.G. Brunner, R.C. Hennekam, S.J. Knight, B.B. de Vries, O. Zuffardi, and E.E. Eichler. Characterization of a recurrent 15q24 microdeletion syndrome. *Hum. Mol. Genet.*, 16:567–572, Mar 2007.

[222] H. Shatkay, S. Edwards, W.J. Wilbur, and M. Boguski. Genes, themes and microarrays: using information retrieval for large-scale gene analysis. *Proc Int Conf Intell Syst Mol Biol*, 8:317–328, 2000.

[223] C. Shaw-Smith, A.M. Pittman, L. Willatt, H. Martin, L. Rickman, S. Gribble, R. Curley, S. Cumming, C. Dunn, D. Kalaitzopoulos, K. Porter, E. Prigmore, A.C. Krepischi-Santos, M.C. Varela, C.P. Koiffmann, A.J. Lees, C. Rosenberg, H.V. Firth, R. de Silva, and N.P. Carter. Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat. Genet.*, 38:1032–1037, Sep 2006.

[224] C. Shaw-Smith, R. Redon, L. Rickman, M. Rio, L. Willatt, H. Fiegler, H. Firth, D. Sanlaville, R. Winter, L. Colleaux, M. Bobrow, and N.P. Carter. Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J. Med. Genet.*, 41:241–248, Apr 2004.

[225] L. Shi, L.H. Reid, W.D. Jones, R. Shippy, J.A. Warrington, S.C. Baker, P.J. Collins, F. de Longueville, E.S. Kawasaki, K.Y. Lee, Y. Luo, Y.A. Sun, J.C. Willey, R.A. Setterquist, G.M. Fischer, W. Tong, Y.P. Dragan, D.J. Dix, F.W. Frueh, F.M. Goodsaid, D. Herman, R.V. Jensen, C.D. Johnson, E.K. Lobenhofer, R.K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P.K. Wolber, L. Zhang, S. Amur, W. Bao, C.C. Barbacioru, A.B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X.M. Cao, T.A. Cebula, J.J. Chen, J. Cheng, T.M. Chu, E. Chudin, J. Corson, J.C. Corton, L.J. Croner, C. Davies, T.S. Davison, G. Delenstarr, X. Deng, D. Dorris, A.C. Eklund, X.H. Fan, H. Fang, S. Fulmer-Smentek, J.C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P.K. Haje, J. Han, T. Han, H.C. Harbottle, S.C. Harris, E. Hatchwell, C.A. Hauser, S. Hester, H. Hong, P. Hurban, S.A. Jackson, H. Ji, C.R. Knight, W.P. Kuo, J.E. LeClerc, S. Levy, Q.Z. Li, C. Liu, Y. Liu, M.J. Lombardi, Y. Ma, S.R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M.S. Orr, T.W. Osborn, A. Papallo, T.A. Patterson, R.G. Perkins, E.H. Peters, R. Peterson, K.L. Philips, P.S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B.A. Rosenzweig, R.R. Samaha, M. Schena, G.P. Schroth, S. Shchegrova, D.D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K.L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S.J. Walker, S.J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, 24:1151–1161, Sep 2006.

[226] N. Sierro, T. Kusakabe, K.J. Park, R. Yamashita, K. Kinoshita, and K. Nakai. DBTGR: a database of tunicate promoters and their regulatory elements. *Nucleic Acids Res.*, 34:D552–555, Jan 2006.

[227] A. Slavotinek, M. Rosenberg, S. Knight, L. Gaunt, W. Fergusson, C. Killoran, J. Clayton-Smith, H. Kingston, R.H. Campbell, J. Flint, D. Donnai, and L. Biesecker. Screening for submicroscopic chromosome rearrangements in children with idiopathic mental retardation using microsatellite markers for the chromosome telomeres. *J. Med. Genet.*, 36:405–411, May 1999.

[228] N.G. Smith and A. Eyre-Walker. Human disease genes: patterns and predictions. *Gene*, 318:169–175, Oct 2003.

[229] G.K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

[230] A.M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A.K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J.P. Yue, J.W. Gray, A.N. Jain, D. Pinkel, and D.G. Albertson. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.*, 29:263–264, Nov 2001.

[231] L N Soldatova and R D King. Are the current ontologies in biology good ontologies? *Nat Biotechnol*, 23(9):1095–8, 2005.

[232] S. Solinas-Toldo, S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H. Dhner, T. Cremer, and P. Lichter. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, 20:399–407, Dec 1997.

[233] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297, Dec 1998.

[234] P. Stankiewicz and A.L. Beaudet. Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. *Curr. Opin. Genet. Dev.*, 17:182–192, Jun 2007.

[235] H. Starke, A. Nietzel, A. Weise, A. Heller, K. Mrasek, B. Belitz, C. Kelbova, M. Volleth, B. Albrecht, B. Mitulla, R. Trappe, I. Bartels, S. Adolph, A. Dufke, S. Singer, M. Stumm, R.D. Wegner, J. Seidel, A. Schmidt, A. Kuechler, I. Schreyer, U. Claussen, F. von Eggeling, and T. Liehr. Small supernumerary marker chromosomes (SMCs): genotype-phenotype correlation and classification. *Hum. Genet.*, 114:51–67, Dec 2003.

[236] L. Stein. Genome annotation: from sequence to biology. *Nat. Rev. Genet.*, 2:493–503, Jul 2001.

[237] L. Stein, P. Sternberg, R. Durbin, J. Thierry-Mieg, and J. Spieth. WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res.*, 29:82–86, Jan 2001.

[238] M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, and J. Mostafa. Detecting gene relations from Medline abstracts. *Pac Symp Biocomput*, pages 483–495, 2001.

[239] L. Tanabe, U. Scherf, L.H. Smith, J.K. Lee, L. Hunter, and J.N. Weinstein. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques*, 27:1210–1214, Dec 1999.

[240] L. Tanabe, N. Xie, L.H. Thom, W. Matten, and W.J. Wilbur. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6 Suppl 1:S3, 2005.

[241] P D Thomas, H Mi, and S Lewis. Ontology annotation: mapping genomic regions to biological function. *Curr Opin Chem Biol*, 11(1):4–11, 2007.

[242] N. Tiffin, E. Adie, F. Turner, H.G. Brunner, M.A. van Driel, M. Oti, N. Lopez-Bigas, C. Ouzounis, C. Perez-Iratxeta, M.A. Andrade-Navarro, A. Adeyemo, M.E. Patti, C.A. Semple, and W. Hide. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res.*, 34:3067–3081, 2006.

[243] N. Tiffin, J.F. Kelso, A.R. Powell, H. Pan, V.B. Bajic, and W.A. Hide. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.*, 33:1544–1552, 2005.

[244] M. Tompa, N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, W.J. Kent, V.J. Makeev, A.A. Mironov, W.S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 23:137–144, Jan 2005.

[245] L.C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens, B. De Moor, S. Aerts, and Y. Moreau. ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res.*, 36:W377–384, Jul 2008.

[246] F.S. Turner, D.R. Clutterbuck, and C.A. Semple. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, 4:R75, 2003.

[247] E. Tuzun, A.J. Sharp, J.A. Bailey, R. Kaul, V.A. Morrison, L.M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, M.V. Olson, and E.E. Eichler. Fine-scale structural variation of the human genome. *Nat. Genet.*, 37:727–732, Jul 2005.

[248] G. Van Buggenhout, C. Melotte, B. Dutta, G. Froyen, P. Van Hummelen, P. Marynen, G. Matthijs, T. de Ravel, K. Devriendt, J.P. Fryns, and J.R. Vermeesch. Mild Wolf-Hirschhorn syndrome: micro-array CGH analysis of atypical 4p16.3 deletions enables refinement of the genotype-phenotype map. *J. Med. Genet.*, 41:691–698, Sep 2004.

[249] M.A. van Driel, J. Bruggeman, G. Vriend, H.G. Brunner, and J.A. Leunissen. A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, 14:535–542, May 2006.

[250] M.A. van Driel, K. Cuelenaere, P.P. Kemmeren, J.A. Leunissen, H.G. Brunner, and G. Vriend. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res.*, 33:W758–761, Jul 2005.

[251] L. J. van 't Veer, H. Dai, M. J. van de Vijver, U. D. He, A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome in breast cancer. *Nature*, 415:530–536, 2002.

[252] S. Van Vooren, B. Coessens, B. De Moor, Y. Moreau, and J.R. Vermeesch. Array comparative genomic hybridization and computational genome annotation in constitutional cytogenetics: suggesting candidate genes for novel submicroscopic chromosomal imbalance syndromes. *Genet. Med.*, 9:642–649, Sep 2007.

[253] S. Van Vooren, B. Thienpont, B. Menten, F. Speleman, B. De Moor, J. Vermeesch, and Y. Moreau. Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Res.*, 35:2533–2543, 2007.

[254] J.A. Veltman. Genomic microarrays in clinical diagnosis. *Curr. Opin. Pediatr.*, 18:598–603, Dec 2006.

[255] W N. Venables and B D. Ripley. *Modern Applied Statistics with S. Fourth Edition.* Springer, New York, 2002.

[256] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, J.D. Gocayne, P. Amanatides, R.M. Ballew, D.H. Huson, J.R. Wortman, Q. Zhang, C.D. Kodira, X.H. Zheng, L. Chen, M. Skupski, G. Subramanian, P.D. Thomas, J. Zhang, G.L. Gabor Miklos, C. Nelson, S. Broder, A.G. Clark, J. Nadeau, V.A. McKusick, N. Zinder, A.J. Levine, R.J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A.E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T.J. Heiman, M.E. Higgins, R.R. Ji, Z. Ke, K.A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G.V. Merkulov, N. Milshina, H.M. Moore, A.K. Naik, V.A. Narayan, B. Neelam, D. Nusskern, D.B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M.L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru,

A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N.N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J.F. Abril, R. Guig, M.J. Campbell, K.V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291:1304–1351, Feb 2001.

[257] J.R. Vermeesch, H. Fiegler, N. de Leeuw, K. Szuhai, J. Schoumans, R. Ciccone, F. Speleman, A. Rauch, J. Clayton-Smith, C. Van Ravenswaaij, D. Sanlaville, P.C. Patsalis, H. Firth, K. Devriendt, and O. Zuffardi. Guidelines for molecular karyotyping in constitutional genetic diagnosis. *Eur. J. Hum. Genet.*, 15:1105–1114, Nov 2007.

[258] J.R. Vermeesch, C. Melotte, G. Froyen, S. Van Vooren, B. Dutta, N. Maas, S. Vermeulen, B. Menten, F. Speleman, B. De Moor, P. Van Hummelen, P. Marynen, J.P. Fryns, and K. Devriendt. Molecular karyotyping: array CGH quality criteria for constitutional genetic diagnosis. *J. Histochem. Cytochem.*, 53:413–422, Mar 2005.

[259] J.R. Vermeesch and A. Rauch. Reply to Hochstenbach et al. 'Molecular karyotyping'. *Eur. J. Hum. Genet.*, 14:1063–1064, Oct 2006.

[260] L.E. Vissers, B.B. de Vries, K. Osoegawa, I.M. Janssen, T. Feuth, C.O. Choy, H. Straatman, W. van der Vliet, E.H. Huys, A. van Rijk, D. Smeets, C.M. van Ravenswaaij-Arts, N.V. Knoers, I. van der Burgt, P.J. de Jong, H.G. Brunner, A.G. van Kessel, E.F. Schoenmakers, and J.A. Veltman. Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. *Am. J. Hum. Genet.*, 73:1261–1270, Dec 2003.

[261] L.E. Vissers, C.M. van Ravenswaaij, R. Admiraal, J.A. Hurst, B.B. de Vries, I.M. Janssen, W.A. van der Vliet, E.H. Huys, P.J. de Jong, B.C. Hamel, E.F. Schoenmakers, H.G. Brunner, J.A. Veltman, and A.G. van Kessel. Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat. Genet.*, 36:955–957, Sep 2004.

[262] L.E. Vissers, J.A. Veltman, A.G. van Kessel, and H.G. Brunner. Identification of disease genes by whole genome CGH arrays. *Hum. Mol. Genet.*, 14 Spec No. 2:R215–223, Oct 2005.

[263] M.L. Voz, J. Mathys, K. Hensen, H. Pendeville, I. Van Valckenborgh, C. Van Huffel, M. Chavez, B. Van Damme, B. De Moor, Y. Moreau, and W.J. Van de Ven. Microarray screening for target genes of the proto-oncogene PLAG1. *Oncogene*, 23:179–191, Jan 2004.

[264] J. Wang, L.A. Meza-Zepeda, S.H. Kresse, and O. Myklebost. M-CGH: analysing microarray-based CGH experiments. *BMC Bioinformatics*, 5:74, Jun 2004.

[265] W.W. Wasserman and J.W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, 278:167–181, Apr 1998.

[266] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L.Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D.J. Lipman, T.L. Madden, D.R. Maglott, V. Miller, J. Ostell, K.D. Pruitt, G.D. Schuler, M. Shumway, E. Sequeira, S.T. Sherry,

K. Sirotkin, A. Souvorov, G. Starchenko, R.L. Tatusov, T.A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 36:13–21, Jan 2008.

[267] R.D. Wolfinger, G. Gibson, E.D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R.S. Paules. Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, 8:625–637, 2001.

[268] J.D. Wren, W.H. Hildebrand, S. Chandrasekaran, and U. Melcher. Markov model recognition and classification of DNA/protein sequences within large text databases. *Bioinformatics*, 21(21):4046–4053, Nov 2005.

[269] G Wrobel, F Chalmel, and M Primig. goCluster integrates statistical analysis and functional interpretation of microarray expression data. *Bioinformatics*, 21(17):3575–3577, Sep 2005. Evaluation Studies.

[270] T. Yakut, S.S. Kilic, E. Cil, E. Yapici, and U. Egeli. FISH investigation of 22q11.2 deletion in patients with immunodeficiency and/or cardiac abnormalities. *Pediatr. Surg. Int.*, 22:380–383, Apr 2006.

[271] A Young, N Whitehouse, J Cho, and C Shaw. OntologyTraverser: an R package for GO analysis. *Bioinformatics*, 21(2):275–276, Jan 2005.

[272] S. Yu, S. Van Vooren, L.C. Tranchevent, B. De Moor, and Y. Moreau. Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. *Bioinformatics*, 24:i119–125, Aug 2008.

[273] F. Zahir, H.V. Firth, A. Baross, A.D. Delaney, P. Eydoux, W.T. Gibson, S. Langlois, H. Martin, L. Willatt, M.A. Marra, and J.M. Friedman. Novel deletions of 14q11.2 associated with developmental delay, cognitive impairment and similar minor anomalies in three children. *J. Med. Genet.*, 44:556–561, Sep 2007.

[274] B R Zeeberg, W Feng, G Wang, M D Wang, A T Fojo, M Sunshine, S Narasimhan, D W Kane, W C Reinhold, S Lababidi, K J Bussey, J Riss, J C Barrett, and J N Weinstein. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28, 2003.

[275] B Zhang, D Schmoyer, S Kirov, and J Snoddy. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, 5:16, Feb 2004.

[276] X Zhang, A Snijders, R Segraves, X Zhang, A Niebuhr, D Albertson, H Yang, J Gray, E Niebuhr, L Bolund, and D Pinkel. High-resolution mapping of genotype-phenotype relationships in cri du chat syndrome using array comparative genomic hybridization. *Am J Hum Genet*, 76(2):312–26, 2005.

[277] F. Zhao, Z. Xuan, L. Liu, and M.Q. Zhang. TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. *Nucleic Acids Res.*, 33:D103–107, Jan 2005.

[278] M. Zhu and S. Zhao. Candidate gene identification approach: progress and challenges. *Int. J. Biol. Sci.*, 3:420–427, 2007.

Biography

Steven Van Vooren was born on June 10th, 1977 in Gent (Belgium). He received a Master degree in Computer Science (Licentiaat Informatica optie informatie- en communicatietechnologie) at the Rijksuniversiteit Gent in 1999 (*cum laude*). From 1999 to 2003, he was a security engineer at BACOB Bank and became project manager after a merger with Dexia Bank. In 2003, he received the Master's degree in Artificial Intelligence (*magna cum laude*) and the Master's degree in Bioinformatics (*magna cum laude*) at the Katholieke Universiteit Leuven, Belgium. In Oktober 2003, he joined the ESAT-SCD group at the Department of Electrical Engineering (ESAT), K.U.Leuven to start working towards a PhD on data mining for molecular karyotyping in constitutional cytogenetics. During this PhD, he was a visiting worker at the Wellcome Trust Sanger Institute and the European Bioinformatics Institute in Hinxton, Cambridge, UK. Together with Dr. ir. Bert Coessens, Prof. dr. ir. Yves Moreau, Prof. dr. ir. Bart De Moor and Prof. dr. ir. Joris Vermeesch, he co-founded Cartagenia, a spin-off company at the K.U.Leuven that markets a software and database platform for management and analysis of genetic data and supports labs and hospitals applying screening assays such as Array CGH in a clinical diagnostic context.