Mapping Biomedical Concepts onto the Human Genome by Mining Literature on Chromosomal Aberrations

Steven Van Vooren[1], Bernard Thienpont[2], Björn Menten[3],
Frank Speleman[3], Bart De Moor[1], Joris Vermeesch[2], Yves Moreau[1]

February 21, 2006

Abstract

Biomedical literature provides a rich but unstructured source of associations between chromosomal regions and biomedical concepts. By mining PubMed abstracts, we annotate the human genome at the level of cytogenetic bands. Our method creates a set of chromosomal aberration maps that associate cytogenetic bands to biomedical concepts from a variety of controlled vocabularies, including disease, dysmorphology, anatomy, development, and Gene Ontology. The association between a band (e.g., 4p16.3) and a concept (e.g., microcephaly) is assessed by the statistical overrepresentation of this concept in the abstracts relating to this band. Our 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 http://www.esat.kuleuven.be/abandapart/.

1 Introduction

Many developmental disorders are caused by chromosomal imbalances and translocations. There are multiple examples where the identification of such aberrations has led (through genotype-phenotype association and positional cloning) to the identification of genes involved in human development and pathology.

To speed up the process of gene discovery, some attempts have been made to associate genomic rearrangements (such as subchromosomal 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 [1] and duplications [2] 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 [17], the Human Cytogenetics Database and ECARUCA [16], Decipher [12], the Chromosome Anomaly Collection [14], the Mitelman Database of Chromosome Aberrations in Cancer [10], the Mendelian Cytogenetics Network DataBase [9], Orphanet [13], ... These efforts differ in setup but aim at aggregating chromosomal aberration information and charting phenotypes and case reports. Some catalogs are available only in print or for a licence fee, other databases merely require registration. Others 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

[1]Department of Electrotechnical Engineering, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium
[2]Center for Human Genetics, Leuven University Hospital, Herestraat 49, B-3000 Leuven, Belgium
[3]Center for Medical Genetics, Ghent University Hospital MRB 2nd floor, De Pintelaan 185, B-9000 Ghent, Belgium

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.

We have developed a method to automatically create chromosomal aberration maps from PubMed 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 being 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.

2 Materials and Methods

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

To discover overrepresented association between concepts and cytobands, we must first locate cytogenetic band identifiers and concepts from the vocabularies (and their synonyms) in the PubMed corpus. We have extended Lucene [4], a high-performance text indexing engine written in Java, to parse all PubMed abstracts and extract cytogenetic bands, ranges of bands, and biomedical concepts that are present in our different structured vocabularies.

2.1 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 [18]. 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 (Human Chromosome Aberration Database [5]), a web-based text-mining tool supporting analysis of human breakpoint data, using the nomenclature for translocations.

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 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 and as an option for the user, an association to a certain abstract can be 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 PubMed 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. Using PubMed (all licensed MEDLINE updates prior to September 6th, 2005), we identified 47,984 abstracts mentioning at least one cytogenetic band or range of bands.

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 noise documents (i.e., documents dealing with

nonhuman species in our corpus). We considered a set of 30000 documents that mention one or more cytogenetic bands. We 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 [11]. 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 1.

rank	phrase	rank	phrase
12522	human	108	rabbit
3054	mouse	108	saccharomyces cerevisiae
957	rat	105	xenopus
523	rodent	102	cat
424	hamster	97	escherichia coli
410	drosophila	95	drosophila melanogaster
165	bovine	92	primates
141	chicken	87	pig
136	caenorhabditis elegans	82	human papillomavirus
132	porcine	81	wolf

Table 1: The most frequently occurring species in a set of 30000 cytogenetic PubMed abstracts mentioning cytogenetic bands

Note that the results from Table 1 do not imply that 12522 documents discuss human cases and 3054 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% of the test document set. As the rat chromosome nomenclature closely follows the human cytogenetic nomenclature [8], abstracts dealing with rat band patterns are a potential source of contamination—but 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.

2.2 Vocabularies

Geneticists, pediatricians or MDs 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 2.

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

2.3 Statistical overrepresentation

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

name	function	example	size
GO.B	Biological Processes	cell growth, signal transduction	941
GO.C	Cellular Components	proteasome, nucleus	388
GO.M	Molecular Functions	ATPase activity	675
GO.E	Gene Ontology	all of the above	1954
LDDB	London Dysmorphology & Neurology Database	microcephaly	1286
LDDB.S	LDDB with synonyms	microcephaly or small head	796
OMIM	Genetic Disorders	attention deficit hyperactivity disorder	1642
CBIL	Human Anatomy	heart muscle	294
OHDA	Embryo Development	early stage, fetus	370
TDMS.s	Systems, Tissues and Sites	cardiovascular system	386
TDMS.l	Microscopic Lesions	disseminated intravascular coagulation	194

Table 2: 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

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 overrepresentation, we assume a hypergeometric distribution as a model.

Let A be the total number of abstracts. We want to qualify the strength of the link between band b and concept c. Let O_{bc} be the observed number of papers that are associated to band b and mention concept c or one of its synonyms. Let B be the number of documents associated to band b and C the number of documents associated to concept C or its synonyms. The p-value is then given by the hypergeometric cumulative distribution function

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

$$= 1 - \sum_{i=0}^{O_{bc}-1} \frac{\binom{B}{i}\binom{A-B}{C-i}}{\binom{A}{C}} \quad (2)$$

$$= \sum_{i=O_{bc}}^{\min(B,C)} \frac{\binom{B}{i}\binom{A-B}{C-i}}{\binom{A}{C}} \quad (3)$$

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 mentioning concept c when drawing B band-related documents without replacement from a corpus of A abstracts.

2.4 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 overrepresented vocabulary concepts found in the literature. The user indicates wihich controlled vocabulary is to be used, according to his/her 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 overrepresented chromosomal regions. If the concept is not found, the application will suggest alternatives with similar spelling. Overrepresented bands are listed together with their p-values, and the raw counts that were used to calculate each p-value. The highly overrepresented 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 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.* [1, 2].

3.1 Heart disease

We now 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 docu-

ments associated to it. The five most relevant hits are shown in Table 3.

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: 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 [19]. Cardiac defects are strongly penetrant in those patients. The third best association, linking *heart* to 20p12, is corroborated by literature on the Alagille syndrome [7], 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 [6]. 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 4.

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 the central nervous system to be involved. 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 microdeletion within 7q11.2 including the elastin gene, leading to disorganized preelastic and mature elastic fibers [15]. Through this brief discussion we have illustrated how different domain vocabularies each provide a specific view towards a genotype-phenotype association.

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 4: 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.

3.2 NIH data set — Genes and Disease

The online NIH book *Genes and Disease* [3] discusses a set of genes and the diseases that they are proven 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.

For Chromosome 1, results are shown in Table 5. 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	1q42.1	0.24E-2

Table 5: 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 provided as supplementary material.

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 interface at http://www.esat.kuleuven.be/abandapart.

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 an relatively high number of cancer related concepts.

Secondly, misses can also be explained by the fact that there exists no literature in the PubMed 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.

3.3 Congenital malformations

To further validate our methodology, we evaluate its agreement with a chromosome map of autosomal deletions composed by Brewer *et al.* [1, 2]. In this work, clinical and cytogenetic information from the Human Cytogenetics Database was used to associate different congenital malformations to nonmosaic 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.

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 6: 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.

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, 1 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.

4 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.

This tool will provide 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 will support genetic counseling and an educated followup of clinical cases.

It will aid 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 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.

Acknowledgements

Our research is supported by grants from several funding agencies and sources: Research Council KUL: GOA-Mefisto-666, GOA-Ambiorics, IDO (IOTA Oncology, Genetic networks); Flemish Government: - FWO: PhD/postdoc grants, projects G.0115.01 (microarrays/oncology), G.0240.99 (multilinear algebra), G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0241.04 ((Functional Genomics), research communities (ICCoS, ANMMM, MLDM); - IWT: PhD Grants, STWW-Genprom (gene promotor prediction), GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors); Belgian Federal Government: Belgian Federal Science Policy Office: IUAP V-22 (Dynamical Systems and Control: Computation, Identification, Modelling, 2002-2006); EU: RTD: FP5 CAGE (Compendium of Arabidopsis Gene Expression); ERNSI: European Research Network on System Identification; FP6 NoE Biopattern; NoE E-tumours.

References

[1] C Brewer, S Holloway, P Zawalnyski, A Schinzel, and D FitzPatrick. A chromosomal deletion map of human malformations. *Am J Hum Genet*, 63(4):1153–9, Oct 1998.

[2] 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):1702–8, Jun 1999.

[3] Genes and NIH Disease. http://www.ncbi.nlm.nih.gov/books/.

[4] Erik Hatcher and Otis Gospodnetić. *Lucene in Action*. Manning Publications co., 2004.

[5] 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.

[6] Rika Kosaki, Kenjiro Kosaki, Kazushige Matsushima, Norimasa Mitsui, Naomichi Matsumoto, and Hirofumi 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. Case Reports.

[7] 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(1):56–60, May 1999. Case Reports.

[8] G. Levan. Nomenclature on G-bands in rat chromosomes. *Hereditas*, 77:37–52, 1974.

[9] MCNdb. http://www.mcndb.org/index.jsp.

[10] F. Mitelman, B. Johansson, and F Mertens. *Mitelman Database of Chromosome Aberrations in Cancer*. http://cgap.nci.nih.gov/Chromosomes/Mitelman, 2005.

[11] NCBI. http://www.ncbi.nlm.nih.gov/Taxonomy/.

[12] DECIPHER: DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources. http://decipher.sanger.ac.uk.

[13] OrphaNET. http://www.orpha.net/.

[14] Anomaly Register. http://www.som.soton.ac.uk/research/geneticsdiv/.

[15] 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.

[16] A. Schinzel. The ecaruca project. http://www.ecaruca.net/.

[17] A. Schinzel. *Catalogue of unbalanced chromosome aberration in man*. de Gruyter, 2001.

[18] L.G. Shaffer and N. Tommerup. *ISCN 2005*. Karger, 2005.

[19] Yakut T, Kilic SS, Cil E, Yapici E, and Egeli U. FISH investigation of 22q11.2 deletion in patients with immunodeficiency and/or cardiac abnormalities. *Pediatr Surg Int*, pages 1–4, Feb 2006.