KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Interpretation and prioritization of genomic single-nucleotide variation

Alejandro Sifrim

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering Science (PhD)

December 2016

Interpretation and prioritization of genomic single-nucleotide variation

Alejandro SIFRIM

Examination committee:
Prof. dr. ir. D. Vandermeulen, chair
Prof. dr. ir. Y. Moreau, supervisor
Prof. dr. ir. J. Vermeesch, co-supervisor
Prof. dr. ir. J. Aerts, co-supervisor
Prof. dr. ir. S. Van Huffel
Prof. dr. K. Devriendt
Prof. dr. L. De Raedt

Prof. dr. P. Robinson
(Jackson Laboratory, Maine, USA)
Prof. dr. T. Lenaerts
(Université Libre de Bruxelles, Brussels)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering Science (PhD)

December 2016

© 2016 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Alejandro Sifrim, Kasteelpark Arenberg 10, postbus 2440, B-3001 Heverlee (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, 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, electronic or any other means without written permission from the publisher.

Abstract

The field of human genetics has evolved at a dramatically fast pace over the past few decades. Breakthroughs in sequencing technologies and large-scale global initiatives such as the Human Genome Project and the Thousand Genomes Project have advanced our understanding of the human genome. The field of clinical genomics tries to leverage such resources in order to determine how variation in the genome influences human health, ranging from common diseases with complex genetic architectures to rare disorders caused by a single severe allele. As more human genomes are sequenced and other large-scale omics data sources (i.e. transcriptomics, proteomics, interactomics, phenomics) become available, we will be able to study more complex aspects of human biology. This will however require novel data analytical techniques capable of processing and integrating these large quantities of information.

This thesis describes the development of computational analysis tools in order to facilitate the interpretation of single-nucleotide variants discovered by whole-exome sequencing datasets in the context of rare genomic disorders. After a general introduction into human genetics and the methods used therein, it presents a case study of a successful application of whole-exome sequencing and a common data analysis workflow for the identification of the genetic cause of a rare genomic disorder (e.g. Nicolaides Baraitser Syndrome). Based on this workflow we describe a software tool called *Annotate-it* which allows clinical geneticists to easily perform similar analyses, aggregates useful information and handles the data management issues coupled with these type of studies. Although these approaches have been shown to be powerful in several exome sequencing studies, they are not always feasible due to their design requirements and rely heavily on the geneticist's expert interpretation of the aggregated information. To remedy this we developed a novel variant prioritization method using a machine learning approach in order to automatically prioritize single-nucleotide variants by integrating information about the disease at hand with evolutionary and biochemical data for the variant. In an extensive benchmark we show that such an approach substantially outperforms classical phenotype-

agnostic prediction methods. In cases where the disease is described by a series of individual phenotypes we elaborate on different aggregation techniques for variant prioritization. Finally we discuss the benefits and limitations of the presented approaches and suggest possible technical, methodological and conceptual improvements to the methods discussed in this thesis.

In conclusion, this thesis presents computational methodology to assist the clinical geneticist in interpreting single-nucleotide variation found in the human genome in order to identify the underlying disease-causing variants.

Beknopte samenvatting

Het domein van de menselijke genetica is in de afgelopen decennia aan een dramatisch tempo geëvolueerd. Doorbraken in sequeneringstechnologieën en grootschalige globale initiatieven zoals het Human Genome Project of het Thousand Genomes Project hebben onze kennis over het humaan genoom bevorderd. Het veld van de klinische genomica tracht deze middelen te benutten voor het bepalen hoe genomische variatie de menselijke gezondheid beïvloedt, gaande van vaak voorkomende ziektes met een complexe genetische architectuur tot zeldzame aandoeningen veroorzaakt door een enkel ernstig ziekte-allel. Naargelang er meer humane genomen gesequeneerd worden en er andere grootschalige "omics" databronnen (b.v. transcriptoom, proteeom, interactoom, fenoom) beschikbaar worden, zal het mogelijk worden om meer complexe aspecten van de humane biologie te bestuderen. Dit zal echter de ontwikkeling van nieuwe data-analytische technieken vereisen voor de verwerking en integratie van zulke grote hoeveelheden informatie.

Deze thesis beschrijft de ontwikkeling van computationele analysetools voor de interpretatie van enkel-nucleotide varianten, ontdekt door gehele-exoom sequenering, in de context van zeldzame genomische aandoeningen. Na een algemene inleiding in de humane genetica en de daarin gebruikte methoden, leggen we een succesvolle toepassing van geheel-exoom sequenering voor, aan de hand van een frequent gebruikte data analyse workflow, voor de identificatie van de genetische oorzaak van een zeldzame genomische aandoening (het Nicolaides-Baraitser Syndroom). Gebaseerd op deze workflow beschrijven we een software tool, *Annotate-it* genaamd, dat klinische genetici in staat stelt om eenvoudig gelijkaardige analyses uit te voeren, nuttige informatie samen te brengen en databeheer in exoom-sequeneringstudies te vereenvoudigen. Hoewel deze analysestrategieën krachtig zijn gebleken in meerdere exoom-sequeneringsstudies kunnen ze soms niet worden toegepast door de vereisten vqn het experimenteel design en zijn ze sterk afhankelijk van de expertise van de geneticus voor het interpreteren van de verzamelde informatie. Om deze problemen te verhelpen beschrijven we hier een nieuwe methodiek voor het

automatisch prioritizeren van enkel-nucleotide varianten, gebruikmakend van machine learning technieken. Dit wordt bekomen door het combineren van informatie omtrent het relevante ziektebeeld en evolutionaire en biochemische informatie over de genomische variant. We tonen op een uitgebreide manier aan dat deze aanpak substantiëel beter presteert dan klassieke phenotype-agnostische predictiemethoden. In gevallen waar de ziekte beschreven wordt door meerdere invididuele phenotypes, breiden we onze methodiek uit door het hanteren van verschillende aggregatietechnieken voor variant-prioritizatie. We besluiten met een discussie over de voordelen en limitaties van de besproken technieken, en suggereren potentiële technische, methodologische en conceptuele verbeteringen.

Concluderend, deze thesis beschrijft computationele methodologieën voor het assisteren van de klinische geneticus in de interpretatie van enkel-nucleotide variatie in het menselijk genoom, opdat ziekte-veroorzakende mutaties geïdentificeerd kunnen worden.

Contents

Abstract i

Contents v

List of Figures ix

List of Tables xi

1 Introduction 5

1.1 The human genome 5

1.2 Variation in the human genome 7

1.2.1 Variation consequence 8

1.2.2 Genetic diversity 11

1.2.3 Genetic disease 11

2 Material & Methods 15

2.1 Detecting genomic variation 15

2.1.1 Massively parallel sequencing 15

2.1.2 Genome alignment & variant calling 17

2.2 Interpreting genomic variation 18

2.2.1 Gene annotation 19

2.2.2 Variant annotation 20
2.2.3 Phenotype Annotation 21
2.2.4 Experimental design 22
2.2.5 Gene Prioritization 23
2.3 Thesis overview 24

3 Case study: Nicolaides Baraitser Syndrome 27
3.1 Summary 28
3.2 Case study: Nicolaides Baraitser Syndrome 28

4 SNV data management, annotation and interpretation 43
4.1 Summary 44
4.2 SNV data management, annotation and interpretation 44

5 Phenotype-specific variant prioritization 67
5.1 Summary 68
5.2 Phenotype-specific variant prioritization 68

6 Comparison of aggregation methods for multiphenotype variant prioritization 103
6.1 Summary 104
6.2 Comparison of aggregation methods for multiphenotype variant prioritization 104

7 Conclusion 119
7.1 Conceptual improvements 122
7.1.1 Annotate-it 122
7.1.2 Variant Prioritization 123

8 Postscript 129

Bibliography **133**

Curriculum Vitae **153**

List of publications **155**

List of Figures

1.1 Overview of classes of genic variation 9

1.2 Relationship of the effect size of an allele and its frequency in the population 13

3.1 Photographs of four individuals with NBS in whom *SMARCA2* mutations were identified. 33

3.2 Exons, protein domains, conserved motifs and mutation spectrum for *SMARCA2*. 37

4.1 Overview of Annotate-it 45

4.2 Screenshot of interactive graphical filter 53

4.3 Randomization scheme used for the semi-synthetic Miller data set 56

4.4 Rank statistics for DHODH based on 1,000 randomizations . . 58

5.1 Overview diagram for eXtasy variant prioritization 71

5.2 Receiver operating characteristic (ROC) curves comparing eXtasy and classical deleteriousness prediction scores 73

6.1 Schematic representation of the benchmarking scheme 109

6.2 Empirical Cumulative Distribution Function (ECDF) of ranks comparing different aggregation schemes 112

6.3 Schematic representation of the benchmarking scheme 114

List of Tables

3.1 Summary of clinical findings in individuals with NBS examined in this study 32

3.2 Annotation of *SMARCA2* mutations identified in individuals with NBS 39

4.1 Comparison of features of different tools for next-generation sequencing annotation and interpretation 49

4.2 Statistics of initial exomes used for randomizations: 8 previously acquired exomes were used to generate 8000 synthetic exomes by randomization. The initial exomes were sequenced by either Illumina GA2 or Illumina HiSeq 2000, then aligned with BWA and variants were called using the GATK pipeline. 55

4.3 Top 10 candidate genes in Schinzel-Giedion case study: The data was analyzed using Annotate-it. Genes were considered a hit if they contained either nonsynonymous, nonsense or essential splice site variants which were not present in dbSNP or found in the 1000 Genomes Project. Genes were then sorted based on number of affected samples in which the gene was a candidate and conservation in placental mammals. Additionaly the total number of variants present across all samples and the number of variants unique to only one sample was given. 60

6.1 Summary statistics for the overall comparison of different aggregation methods 112

6.2 Summary statistics for the comparison of different aggregation schemes by fraction of informative phenotypes 115

Acknowledgements

> Words do not express thoughts very well. they always become a little different immediately after they are expressed, a little distorted, a little foolish. And yet it also pleases me and seems right that what is of value and wisdom to one man seems nonsense to another.
>
> Hermann Hesse

Ever since I was a prepubescent young child I've always wanted to become a scientist. Although I do not remember the exact reason for this ambition, I can say with certainty that it was likely based on a naive understanding of what it means to dedicate oneself to the scientific profession. As years went on, I madly pursued this vision of discovering new knowledge, grooming a critical mindset, and perhaps solving puzzles which could potentially better our understanding of life in this universe. It however took me nearly a decade of practicing science to learn that being a scientist is not just about the science, but also about the people with whom we share the journey. Be it the people that we learn from, the people that we teach, the people with whom we exchange ideas, the people which it affects, or the friends and family that support us. As one does not get many occasions to explicitly thank these fellow travellers (or at least not as much as they deserve) I will use this opportunity to take a few words to express my deepest gratitude. As many of you know this defense (and the writing of this acknowledgment) comes much delayed, therefore my sincerest apologies if you have temporarily slipped beyond the crevaces of my mind.

First and foremost I would like to thank my supervisor, Prof. Yves Moreau. Thank you for the opportunity to work in this exciting multidisciplinary field of computational biology and the freedom to sometimes pursue my ideas (even if they weren't always that great), I've learned a great deal from my time in your group. I would also like to thank my co-supervisors: Prof. Joris Vermeesch (without your course in cytogenetics I would have never landed in this fantastic field) and Prof. Jan Aerts (for the stimulating chats and the inspiration to

pursue a postdoc at the Sanger Institute). I would also like to thank Prof. Koen Devriendt who has always been an inspiration and a role model of what research in clinical genetics should be. I'd also like to thank the members of my examination committee: Prof. Dirk Vandermeulen, Prof. Sabine Van Huffel, Prof. Luc De Raedt; and my external jury members: Prof. Tom Lenaerts and Prof. Peter Robinson, for your time and dedication to evaluate and correct this PhD dissertation.

This work is obviously no one-man job, but the result of an extensive collaboration between colleagues. That is a phrase often used in PhD acknowledgements, but the reality is that this work would not only have been impossible, but more importantly, it would have been a lot less fun without such wonderful colleagues. Therefore, thank you to my former colleagues at ESAT: Amin, Arnaud, Babis, Georgios, Griet, Jiqiu, Leo, Lieven, Marc, Minta, Peter, Pooya, Raf, Raf 2.0, Ryo, Tim, Sarah,Yousef. I'd also like to thank my collaborators at the hospital: Beata, Francesca, Jeroen B. , Jeroen V.H., Mala, Paul. I'd also like to thank my more-recently-former colleagues at the Sanger Institute who always anticipated this day would eventually come: Caroline, Dan, Di, Elena, Gabi, Giuseppe, Jawahar, Jeff, Jenny, Jeremy, Joanna, Juliet, Keren, Marc, Matt, Nadia, Patrick, Raheleh, Ro, Sarah, TJ, Tom. Also a sincere thank you to the patients and their families who contributed the data to make this research possible. Also thanks to my students: Tugce and Mahmoud, I'm always very proud to see the progress during their current PhDs. Also thanks to Kristof, my master's thesis supervisor, who has probably taught me the most and pushed me to pursue a career in computational biology. I'd especially like to thank two individuals who have made life as a PhD student significantly (in the non-statistical sense) more interesting: thank you Dusan, aka *the Bear from Serbia*, for forcing me to experience the boundaries of hypothermia several times a day while discussing the finer nuances of the collective Serbian psychosis and/or ensemble classifiers. Also thank you Nico, aka Becky, for being my partner-in-crime from the very beginning of this adventure, keep firing them lasers like the ninja warrior we all know you really are.

I'm also fortunate enough to have good friends and family who supported me, even while I approached new levels of annoyingness, and who made life as a PhD student seem like a picnic (I don't like walking) in the park. Thank you Driss, Elke, Sem, (Lord) Stuart, Kevin, Dieter, Jonas, Rik, Nick, Arwen, Lara, Maarten, Annette, Liesbet. A special shout-out goes to my former and current cats (who are the best consolation when a manuscript gets rejected): Felix, Agatha, Couscous and Nommie Chomsky. Thank you to my "in-laws": Caroline, Johan, Josiane, Frans; I don't always understand Tiens but you're great. Thank you to my family in Argentina and other parts of the world. I may not have much family close by, but the ones that I do have are magnificent. Thank you to

my elder brother Pablo and sister-in-law Anika, you guys are phenomenal, your fortitude and character are awe-inspiring. Loren, there are no words to thank you enough (only cats and food) for getting me going when times were rough, or for the happiness you provided when things were joyful. You're the best! Finally, I'd like to dedicate this work to both my parents: Isabel Fernandez and Daniel Sifrim. For all their quirks, they are the best parents one could wish for and have shaped me into the person I am today.

Chapter 1

Introduction

> In 50 years - or 20 years, or 200 years - our current epistemic horizon (the Big Bang, roughly) may look as parochial as the horizon Newton had to settle for in his day, but no doubt there will still be good questions whose answers elude us.
>
> Daniel Dennett

During the past decade, high-throughput technologies have caused a true revolution in the fields of molecular biology and clinical genetics. The focus of research has shifted from the analysis of single, or a limited group of genes to a full genomic view. Processing this kind of heterogeneous data to identify genes causal for specific phenotypes is far from trivial, especially when dealing with disorders involving multiple genes. The sheer amount and the low signal-to-noise ratio of the data demand automatic and statistical approaches to find genes that are relevant to the biological process or pathology of interest. In this chapter we offer a brief introduction into to the human genome, genomic variation and genetic disease. In the next chapter we describe sequencing technologies and analysis software and how these can be used investigate the genome and to discover disease-causing mutations, there we also provide a brief introduction into some of the methods used in the following the chapters.

1.1 The human genome

Biology, from the simplest to the most complex organisms, is dictated by an imperfect transfer or *inheritance* of information from one generation to the next.

It is through this process first described by Charles Darwin in his *On the Origin of Species* (1859) that organisms evolve to adapt to their changing environments. The unit if heredity, first defined by Gregor Mendel (1965), is called the *gene* and can exist in different versions or *alleles*. Genes are organized into different chromosomes (Morgan T.H., 1910). The molecular basis of genes is a molecule, first described by Friedrich Miescher (1869), called *deoxyribonucleic acid* (DNA) [1]. Famously discovered by Watson and Crick [2], the structure of DNA consists of a right-handed double-helix. This helix is a dimer composed of nucleotide subunits. These nucleotides are linked through phosphate groups and consist of a deoxyribose sugar and one of 4 nucleotide bases: adenine (A), guanine (G), cytosine (C), thymine (T). The nucleotide bases form complementary pairs ($A \leftrightarrow T$, $C \leftrightarrow G$) in the double-helix through hydrogen bonding, making DNA replication an elegant and relatively simple process of complementary matching of a template DNA strand [3]. The DNA only codes the genetic information but does not serve a structural or functional role in living organisms, this is mainly achieved by a class of biomolecules called *proteins*. Proteins consist of polypeptide chains of 20 different amino acids, whose structure and chemical properties allow it to perform an endless variety of functions. The sequence of amino acids is coded in the DNA by a degenerate coding scheme where each triplet of nucleotides (called a *codon*) represents a certain amino acid. This sequence of codons is initially transcribed into *ribonucleic acid* (RNA) by a protein called RNA polymerase. This RNA can then in its turn be translated into protein by an RNA/protein hybrid cellular machinery called the ribosome. This process of transfer of information through transcription and translation is called the *Central Dogma* of biology.

A complete set of genetic information (including all genes) is called a *genome*. Every human cell, with the exception of red blood cells, platelets, male and female gametes, contains two copies of the genome (making humans *diploid*, in contrast to *haploid* organisms which contain only one copy of their genome). The diploid genome is organized into two sets of 22 autosomal chromosomes and a set of sex chromosomes (called X and Y chromosomes). Males carry one X and one Y chromosome, whereas females carry two X chromosomes. In a process called cell division, or *mitosis*, a parent cell duplicates its diploid genomes and segregate a full diploid copy into each of two daughter cells. In organisms capable of sexual reproduction, a process called *meiosis* allows the creation of haploid gamete cells through random selection of one of the two sets of chromosomes (male gamete cells carrying either X or Y chromosomes while female gametes carrying a single X chromosome). Haploid cells from two different individuals can then fuse to form a (diploid) zygote with a mixture of two different genomes.

The haploid human genome in its entirety consists of approximately 3 billion

nucleotides. Only roughly 1% of the human genome codes the 20000-25000 protein coding genes. Although the remaining 99% was believed to be of lesser importance, recent large-scale studies have shown that a majority of the genome might play some biochemical role, be it in the dynamic regulation of genes are in the chromosomal architecture [4]. Any position or *locus* in the genome can occur in many different alternate forms or *alleles* in the modern Mendelian sense (although most of the genome is conserved across individuals of a single species or closely related species, as we will see in Section 1.2). An individual carrying two different alleles of a locus is named *heterozygous* for that locus, while an individual only carrying one allele of a locus is named *homozygous*. The exact allelic composition of a locus is called a *genotype*. The first method for efficiently determining genotypes was proposed by Frederic Sanger through the method of dideoxy-sequencing, or popularly known as Sanger sequencing [5]. Sanger sequencing allows the determination of the exact nucleotide sequence and genotype of a short stretch of DNA. A combination of this sequencing method with rapid breakthroughs in selective DNA amplification [6], vector cloning [7] and sequencing alignment/assembly algorithms paved the way to sequencing the first full human genome sequence by the Human Genome Project and Celera culminating in the early 2000s [8, 9].

1.2 Variation in the human genome

Genetic diversity is the driving force of evolutionary adaptation to selective pressures. Genetic diversity is achieved by two major processes: *mutation* and *recombination*. A mutation is a change in the DNA, effectively being a source for new alleles. Mutations can arise either spontaneously as a random occurrence in the DNA replication process or by exposure to mutagens such as radiation or chemicals in the environment. The mutation rate in living organisms is kept extremely low through an elaborate cellular proofreading mechanism, as a high mutation rate would lead to a rapid loss of beneficial alleles. In humans the mutation rate is estimated to be approximately 2.5×10^{-8} mutations per nucleotide, or 175 mutations per generation of a diploid human genome [10]. A second level of variation is achieved by the cellular processes which generate new groupings of alleles. This is, for example, very elegantly achieved in sexual reproduction through meiosis, where chromosomes from each parent containing each a particular set of alleles are recombined into cells with new pairs of chromosomes. More complex recombination processes take place in living cells, where chromosomes are structurally altered through large-scale insertion, deletion, inversion and translocation events. Although these processes are extremely interesting, and an important source of genetic variation (and play a major role in genetic disease), in this thesis we focus primarily

on single nucleotide mutations. In this section we provide an overview of the different aspects of mutations. First (subsection 1.2.1) we describe the possible consequences that a mutation can have depending on its position in the genome and the nature of the change. Then, we look at the relation between allele and phenotype, more specifically on how certain alleles can cause genetic disease (subsection 1.2.3. A final important remark when talking about variation in the following sections is that variation is obtained by comparing two genomes, often times one of the genomes being the *human reference genome*. It is therefore important to remember that any variation could potentially be a mutation in the reference genome while being wild-type in our genome of interest. This problem is mitigated to a certain extent by the reference genome being constructed as a consensus of several individuals, yet only population frequency information is truly reliable when asking the question if a variant is the wild-type or mutated allele.

1.2.1 Variation consequence

Not every variant is born equal. Its functional effect, if any, depends on its position and nature of the change. In this thesis we focus on small variation of a single to a few nucleotides. A single nucleotide variant (SNV) also known as a *point mutation* can be classified into two main types *base substitutions* and *base insertions* or *deletions* (*indels*). Base substitutions are often further divided into two classes: *transitions* and *transversions*. Transitions are changes for which the replacement base belongs to the same chemical category (purine or pyrimidine) as the wild-type ($A \leftrightarrow G$, $C \leftrightarrow T$). Transversions are then simply the opposite of transitions, meaning an interchange from a purine to a pyrimidine or vice versa.

Variation in the coding regions of the genome

The functional consequence of variants is determined by their position and genotypes (Figure 1.1). First we consider the possible consequences that happen when variants arise in genic regions of the genome. Genes consist of 2 types of sequence: *introns* and *exons*. Introns are stretches of nucleotides which are excised by the cellular splicing machinery when converting pre-mRNA to mature mRNA, and as such do not code for amino acids in the protein. In contrast, exons make up the sequence of the mature mRNA and make up the protein-coding sequence and *5'* and *3' untranslated regions* (UTRs) of the gene. Introns and UTRs are generally under lower evolutionary constraints than protein-coding sequences and are thus believed to be more tolerant to variation. These regions however can contain regulatory motifs (such as transcription factor or miRNA

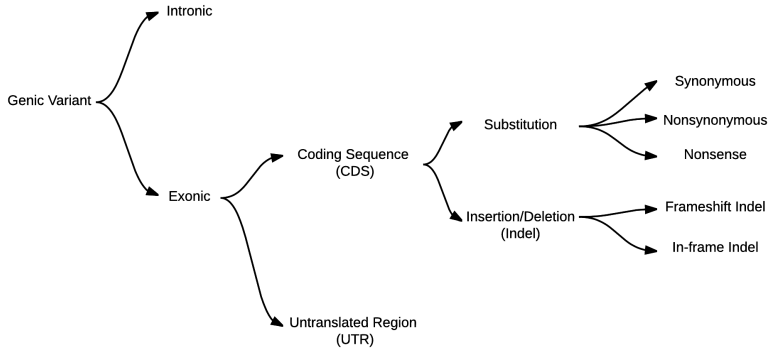


Figure 1.1: Overview of classes of genic variation: Genic variants are first categorized by their presence within an exon. Variants within the amino-acid coding sequence (CDS) are further classified according to their effect on the protein sequence: no change (e.g. synonymous variants), amino-acid substitutions (e.g. nonsynonymous and in-frame indels) or truncation (e.g. nonsense and frameshift indels).

binding sites) which can affect gene expression or alternative splicing (leading to other protein isoforms), and as such should never be fully disregarded when considering functionally relevant variation. Due to our limited understanding of these phenomena, functional interpretation of intronic variation is currently limited to splice donor and acceptor sites as changes in these sequences have the ability to dramatically alter the protein product.

The consequence of exomic variation is determined by its effect on the degenerated code which relates codons to amino acids. Three types of single nucleotide substitutions exist in the protein-coding sequence: *synonymous*, *nonsynonymous* (or missense) and nonsense variants. Synonymous variants are changes in the DNA which result in a codon coding the same amino acid as the wild-type codon. Because these variants don't change the underlying protein sequence, they are believed to be the least impactful of all protein-coding substitutions. This is however under some debate as it has been shown that specific codon selection might be important for protein translation kinetics and regulatory motifs [11, 12, 13, 14]. Nonsynonymous or missense variants change the codon so that the coded amino acid is different as that from the wild-type protein. The functional severity of such a change varies greatly from case to case, depending on the region of the protein which is affected (e.g. the

crucial active site of an enzyme) and the nature of the change. If a crucial amino acid is changed into an amino acid with similar chemical characteristics (*conservative substitution*), its effects on protein function might be less severe compared to a change to a radically different (e.g. hydrophilic ↔ hydrophobic) amino acid (*nonconservative substitution*). The most drastic effect is when a mutation alters the signal that tells the ribosome to stop translating the mRNA, the *stop codon*. This type of mutations is called a nonsense variant and either generates a stop codon where there was none before (*stop gain*) or removes an existing stop codon replacing it with a codon coding for an amino acid (*stop loss*). These mutations lead respectively to a truncated protein or a protein with an appended sequence of previously untranslated amino acids. The closer a nonsense mutation event occurs to the 3' end of the *open reading frame* (ORF) the more likely the protein might still posses some biological functionality.

Finally, indels can also be divided into two categories: *frameshift* and *non-frameshift*, depending on if the number of bases inserted/deleted forms a multiple of three or not. If the number of bases inserted/deleted is no multiple of three then this results in a shifted reading frame, changing all codons and thus amino acids of the protein downstream of the indel until a stop codon is encountered. Hence, these frameshift mutations have a dramatic disruptive impact on protein function. If the reading frame is not shifted ($3n$ bases deleted/inserted) then these mutations behave similarly to missense variants, depending on the protein region and chemical nature of the change.

Variation in the non-coding regions of the genome

The non-coding genome has long been considered of relative little functional importance compared to the protein-coding regions of the genome. However, recent studies such as the Encyclopedia of DNA Elements (ENCODE) Project suggest that a major fraction ($\tilde{8}0\%$) of the non-coding genome might play some biochemical function [4]. Non-coding RNAs (RNA molecules which are not translated into proteins), from small micro RNAs (miRNAs) to long non-coding RNAs (lncRNAs) have been shown to play an important role in the regulation of protein-coding genes [15]. Additionally, the importance of transcription factor binding sites has long been known to be crucial in gene regulation. These sites can not only be found close to the regulated gene but sometimes quite distant (several hundreds of thousands of nucleotides up/downstream). Nevertheless, understanding the complexities (e.g. tissue specificity) and interplay of these regulatory mechanisms is extremely difficult, let alone understanding the impact of variation on these mechanisms. Because of this, and the sheer amount of non-coding variation, most studies discard these variants as it's unfeasible to comprehensively interpret them at this time.

1.2.2 Genetic diversity

Not all types of mutations described in 1.2.1 occur equally frequently in a population of individuals. The frequency of an allele is dependent on several factors such as migration, drift, population size but most importantly natural selection. The more deleterious an allele is for the fitness of its carrier, the more negative selective pressure it will experience and the lower its population frequency will be. Whole-genome sequencing population studies [16, 17, 18] illustrate how these different mutation types exhibit allele frequencies in line with their estimated deleteriousness and evolutionary conservation. Good estimates of allele frequency are crucial in the identification of disease-causing variation as one does not expect deleterious alleles to be commonly present in the healthy population. Additionally, these studies show that the human genome seems robust to some degree of highly deleterious loss-of-function (LOF) mutations as any individual will carry ± 100 of these rare mutations, effectively knocking out over 1000 different genes in the sequenced population [19].

It is important to note here the difference between deleteriousness of a genomic variant and its putative pathogenicity (i.e. its contribution to human disease). A deleterious variant will perturb gene function at the molecular level for that particular allele of the gene; be it by changing its expression level, altering the protein function or abolishing the gene product altogether. A deleterious variant in a particular allele of a gene may lead to disease (i.e. be pathogenic) if it manages to perturb the biological system and it's underlying pathways sufficiently to cause a phenotype at either the cellular or organismal level. For example, in genes known to be responsible for recessive disease, one single deleterious allele is not enough to cause the disease phenotype, two deleterious alleles are needed (one in each homologous copy of the gene) to sufficiently alter biological function and only then will these deleterious variants be considered pathogenic. In more complex diseases, many different deleterious variants in many dif ferent genes and their interplay with environmental factors might concurrently play a role in provoking the necessary perturbations to cause the disease phenotype.

1.2.3 Genetic disease

In 1.2.1 we described different ways in which mutations can affect biological function, primarily by altering protein-coding genes. Depending on the biological processes in which such a mutated gene is involved, this alteration at the molecular level might directly lead to a macroscopic manifestation or phenotype. When this phenotype is experienced as illness, it is referred to as a genetic disorder. There are many different types of disorders depending on the etiology

of the disease. Some genetic disorders can be caused by a single mutation in a single gene with a large effect size (*monogenic disorders*) whereas other disorders might be caused by many mutations in many different genes with moderate to small effect sizes(*complex genetic disease*).

The frequency at which disease-causing mutations are found in the population is directly correlated with their effect size or penetrance (i.e. the probability of having the disease given the carrier status of the allele). Variants leading to a severe reductions of fitness will be under strong negative natural selection and therefore appear to be rare in the general population (with the exception of variants which benefit some germ cells over others [20] or have other adaptive benefits [21]). Inversely, variants with low to moderate effects will be under less evolutionary constraint and will therefore occur more frequently in the general population. In the case of inherited germ line variants, a direct correlation between the effect size of the allele and its frequency in the population, as more severe reductions in fitness will lead to higher levels of negative natural selection (Figure 1.2) [22]. Some variants may also appear spontaneously by *de novo* mutation events either in early (pre- or early post-zygotic) cell divisions in the germ line or as somatic mutational events later in life (e.g. cancer). In this thesis we will mainly focus on rare monogenic disorders which are predominantly caused by very rare, often *de novo*, variants with in most cases complete or near-complete penetrance.

Rare monogenic disorders, also known as Mendelian disorders, are conceptually easy to understand as they are caused by highly penetrant mutations in a single gene. They can however follow different possible inheritance patterns. If a single heterozygous allele causes the disease phenotype then that disease is considered to have a *dominant* inheritance pattern (e.g. Huntington's disease, caused by a single heterozygous mutation in the *HTT* gene). If the disease is caused by a single homozygous allele then the disease is said to follow a *recessive* inheritance pattern (e.g. Cystic Fibrosis, caused by homozygous mutations in the *CFTR* gene). When the disease is caused by two heterogeneous recessive alleles at a particular locus, both in a heterozygous state, then the disease is said to follow a *compound heterozygous* inheritance pattern. Recessive disorders can often also be caused by compound heterozygosity as both alleles are inactivated, but sometimes show reduced penetrance in cases with compound heterozygosity depending on the degree of gene inactivation by the individual alleles (e.g. Phenylketonuria [23], Tay-Sachs Disease [24]). Finally, inheritance patterns vary if the causative locus lies on one of the sex chromosomes, resulting in differences in disease occurrence in males and females as males are virtually hemizygous for all loci on their X chromosome (e.g. Hemophilia A, Duchenne muscular dystrophy).

Rare genetic disorders are effectively catalogued by OMIM (Online Mendelian

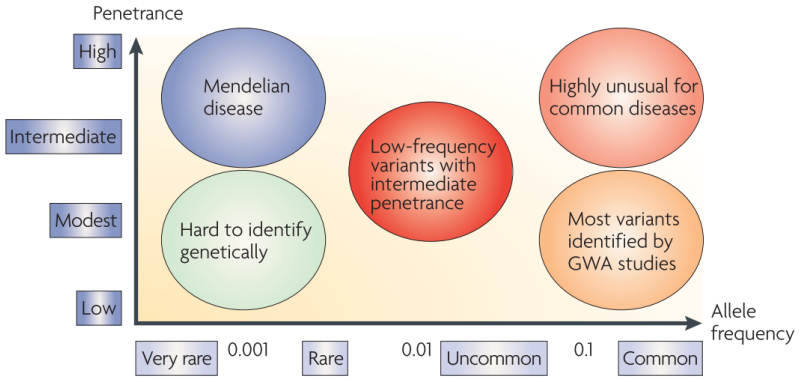


Figure 1.2: Diagram of the relationship between the penetrance of an allele and its frequency in the population (adapted from MacCarthy et al., 2008). Highly penetrant variants are usually found at very low frequencies due to strong negative selection, these variants can cause Mendelian monogenic disease. Combinations of common variants with small effect sizes can jointly cause complex diseases, these associations are usually found to classical genome-wide assocation (GWA) studies in large sample sizes.

Inheritance of Man), a manually curated database of Mendelian phenotypes and their reported molecular causes. As of January 2014, OMIM reports 3674 Mendelian phenotypes with known molecular basis and 1765 phenotypes with unknown molecular basis to this date. The World Health Organisation estimates the amount of rare diseases in the population to be in the area of 5000-8000 individual diseases. Although these disorders are rare (by definition less than 1 in 2000), the European Organisation of Rare Diseases (EURORDIS, `http://www.eurordis.org`) estimated 3-4% of births in the European Union to suffer from a rare genetic disorder, of which only 80% have a successful molecular diagnosis. Although few therapeutic outcomes exist for many of these disorders, the families of the 1% of births without possible diagnosis usually suffer great psychological stress and embark on expensive "diagnostic odysseys" often lasting decades, to a great cost for healthcare organizations.

The study of Mendelian disorders does not only have a substantial societal impact, it also provides fundamental scientific insight into human genetic disease. Most importantly, they provide a window into the function of the affected gene and on how this affects the phenotype, overall enriching the available genomic annotation. Additionally, they provide novel mechanistic insights into pathogenicity of mutations (e.g. the discovery of trinucleotide repeat expansions

in neurodegenerative disorders [25]). Also, in cases where the disorder shows some degree of phenotypic variability they provide grounds for the study of modifier genes or alleles (e.g. the role of intronic short repeats in cystic fibrosis [26]). As such some Mendelian disorders may help us in understand *oligogenic* inheritance which in turn could provide links to common *polygenic* diseases caused by various subtle mutations in many interacting genes.

Some investigators argue that animal gene knock-out models for all genes in the genome (so called *reverse genetics*) will provide a much more systematic approach into the study of human disease compared to classical *forward genetics* [27, 28]. However, these animal model approaches have several shortcomings: not all human genes have a corresponding gene (*orthologue*) in the animal model, some phenotypes might be difficult to assay in the animal model (e.g. cognitive behaviour) and naturally occurring disease mutations might have a more nuanced effects than large knock-out deletions in the animal model. These reasons accentuate the need for human studies in a clinical setting, although recent developments in human pluripotent stem cell technologies might allow us to cross the gap between animal and human disease models in the near future [29].

Chapter 2

Material & Methods

> He who seeks for methods without having a definite problem in mind seeks in the most part in vain.
>
> David Hilbert

2.1 Detecting genomic variation

To determine the genomic variation present in a given sample complex protocols of experimental and data analysis techniques have to be performed. In this section we first describe current sequencing technologies which are used to determine the DNA sequence of the sample. These technologies don't produce a single contiguous read of the whole chromosome but rather hundreds of thousands to millions of short overlapping reads. We then proceed to describe the algorithms needed to determine genetic variants from these short reads.

2.1.1 Massively parallel sequencing

The first generation of DNA sequencing biotechnology was based on the Sanger method (also known as capillary sequencing) of dideoxy-termination sequencing reactions. This method in combination with recombinant DNA vector techniques was successfully used in order to sequence the first human genome. Although the Sanger method is still considered by many the gold standard, it is low-throughput and labor and cost-intensive. In the last decade advancements in

sequencing chemistry, imaging technologies and computational methods have led to what was coined "Next-Generation Sequencing" (NGS) or Massively Parallel Sequencing (MPS). Most of these sequencing technologies consist broadly of four major steps: template preparation, sequencing, signal detection and data analysis. Different combinations of different protocols for each step determine the resulting differences between sequencing platforms (e.g. Illumina/Solexa, Roche 454, SOLiD, Helicos BioSciences, PacBio). The resulting differences in read lengths, error rates, cost, run time, throughput determine their applicability in certain experimental scenarios.

After DNA extraction from the sample (e.g. blood, saliva or tissue), the DNA needs to be prepared for the following sequencing reaction in a so-called template preparation step. First the DNA is randomly broken into fragments of sizes of the desired length. Most imaging systems are not capable of detecting fluorescent events at the single-molecule level, therefore representative and non-biased amplification of these fragments is required through methods such as emulsion PCR [30] and solid-phase amplification [31]. These methods allow highly parallelized PCR reactions in which a single DNA fragment goes through clonal amplification. These clonal amplification products are immobilized be it in the forms of beads in PicoTiterPlate wells [32] or on a glass slide in solid-phase amplification through covalently attached forward and reverse primers. These amplification steps do come at a price, as DNA polymerase can introduce errors in the template and sequence-specific amplification bias can occur leading to under or over-representation of certain fragments.

Clonal amplification and immobilization result in a population of identical DNA fragments spatially clustered together. When these fragments undergo a sequencing reaction cycle which produces a specific fluorescent event, the observed signal is the consensus of all the reactions across the identical templates for that cycle. It is of crucial importance that this reaction is efficient, because if any lag occurs this leads to *dephasing* of the signal, dramatically increasing the signal-to-noise ratio in subsequent reaction cycles. Several different sequencing reactions have been used such as cyclic reversible termination [33] (e.g. Illumina/Solexa), sequencing by ligation [34] (e.g. SOLiD) and pyrosequencing [35] (e.g. Roche/454). For an extensive review of the different reactions we refer to Metzker *et al.* [36]. Every fluorescent event in the incorporation reactions is captured by advanced charge-coupled device (CCD) cameras. After image processing one can assign the probability of observing a given base (a process called *basecalling*) for each clonal cluster and determine the most likely sequence of bases. This can be done for hundreds of thousands to millions of clusters simultaneously generating large amounts of sequence reads in parallel.

Despite the dramatic cost reduction per base-pair compared to traditional Sanger

sequencing, MPS still comes at a considerable cost (although costs are steadily declining). In order alleviate this, sequencing efforts can be focussed on specific regions of the genome, such as all exonic (also known as whole-exome sequencing). This can be achieved by targetted capturing of said regions using massively multiplex PCR with specific primers for the regions one wants to capture. This can be done in solid-phase hybridization using microarray technology [37] (e.g. NimbleGen arrays), in solution hybridization [38] (e.g. Agilent SureSelect) or by microfluidics technologies [39] (e.g. RainDance Technologies). Although these capturing technologies allow us to selectively sequence desired regions, differential capturing efficiency and amplification can cause biases in library representation and coverage compared to whole-genome sequencing. A particular known issue is the difficulty to capture GC-rich regions (often found in the first exon of genes).

2.1.2 Genome alignment & variant calling

Once reads have been generated, computational analysis can be performed depending on the research question at hand. Reads can be used to either determine the sequence of a previously unsequenced species, also known as *de novo assembly*, or in a resequencing study (e.g. sequencing humans) to determine variation in the sequenced individual by *genomic alignment*. Although genome assembly is a fascinating and methodologically complicated problem, in this thesis we focus on human resequencing studies and thus refer the reader to Flicek and Birney *et al.* [40] for a review of assembly methods. In genome alignment one tries match short reads (usually around 100 bp) to the already assembled reference genome (e.g. for humans 3 Gb of sequence). A multitude of public and commercial aligner tools exist, emphasizing different aspects of accuracy, speed and memory footprints. These characteristics are often time dependent on the intrinsic properties of the sequencing platform used (e.g. read length, error rate, quality degradation per cycle). For a comparison of different alignment strategies and implementations we refer to Li *et al.* [41]. Commonly used aligners are SSAHA2 [42](for long reads),BWA [43] and Bowtie [44] (for short reads). The output of most aligners is a Sequence Alignment/Map (SAM) file [45], a tab-delimited file in which each line represents a single read and contains information regarding the mapped chromosomal position, potential mismatches with the reference, quality metrics for the base calls and alignment (e.g. uniqueness of alignment). This SAM file can be then compressed to a binary version called the Binary Alignment/Map (BAM) file.

When all the sequenced reads have been matched to a specific genomic location (be it either on the reverse or forward strand), differences with the reference genome are identified in a process called *variant calling*. Here the information at

a single locus, across many reads, is considered to determine if the sample varies from the reference for that locus. If the locus contains a variant, the genotype is determined by assessing the number of reads containing the observed variant. As the reads are sampled from homologous pairs of chromosomes (with the exception of hemizygous sex chromosomes or aneuploidy events), a heterozygous variant will only be present in roughly halve of the reads, while a homozygous variant will be observed in all reads (with the exception of sequencing errors). The higher the number of reads available at a genomic position (also known as the *coverage* for that position), the more confidence one can have in the call. Given the error-rate of MPS platforms and the binomial variance, 30X coverage usually gives sufficient confidence for heterozygous calls. As with short-reads aligners, a multitude of freely available variant calling tools exist [46], using different modelling approaches to accurately call variants. The most popular variant callers are the Genome Analysis Toolkit (GATK) [47] and SAMtools [43]. Because of alignment and platform-specific issues, indels are much more difficult to correctly call than SNVs. Therefore specific methods have been developed in order to tackle indels seperately such as Indel [48] (performing local de novo reassembly) and Indel [49]. For a comparison of indel callers we refer to Neumann *et al.* [50]. Variant callers usually output variants (be it indels or Sn Vs) in Variant Calling Format (VCF4) files [51].

2.2 Interpreting genomic variation

VCF4 files can contain tens of thousands (for whole-exome sequencing studies) to millions of variants (for whole-genome studies) per individual. Based on the genomic position alone it is impossible to determine which variants are likely to be important in the studied phenotype. In this section we discuss different ways by which VCF4 files can be interpreted in order to identify disease-causing variation. This process usually starts by annotating variants with current knowledge about genes overlapping the given locus and the consequence of the observed variants for these genes (as discussed in 1.2.1). Additionally trained predictive models can be used to score variants based on their likely deleteriousness or on the chance that gene function is maintained given a single functional copy of the gene. Furthermore, we can look at the occurence of the observed variant healthy populations. Once variants have been selected for certain degrees of functional impact and rarity, we can look at how these variants segregate in familial pedigree (either to find de novo variants or familial segregation in the case of several affected individuals) or in populations of cases compared to controls. If other genes are known to be involved for the particular phenotype under study, gene prioritization methods can be used in order to narrow down the list to candidate genes due to similarity with known

genes. Finally databases of model organism knock-outs or knock-downs can be consulted to find genes which experimentally produce similar phenotypes.

2.2.1 Gene annotation

Genes are genomic regions containing transcripts that have either been experimentally observed or for whom an open reading-frame has been predicted. Due to the massive amount of genetic research, the current state of knowledge about genes is constantly in flux. Projects such as Ensembl [52], RefSeq [53] and the UCSC Genome Browser [54] aim to manage this deluge of information by curating sets of transcripts and genes. Additionally projects such as ENCODE [4] provide a wealth of information concerning non-coding regions of the genome. This publicly available information can be used to annotate variants in a VCF file using one of the many available annotation tools such as ANNOVAR [55], the Variant Effect Predictor [56], SNPeff [57]. These software not only check for overlap of transcript definitions with the genomic positions in the VCF file but also compute the consequence of the nucleotide changes for the transcripts in question. It is important to note that the choice of transcript set and annotation tool can have a dramatic impact on the resulting annotation. McCarthy *et al.* have shown that annotating a set of variants with either Ensembl or RefSeq transcripts using a single annotation tool only resulted in 44% concordance in annotation. They also show that annotating variants with the same transcript set but with two different tools (VEP and ANNOVAR) resulted in only 67% mainly due to the way in which a consequence is defined and chosen when multiple transcripts are available (e.g. longest transcript, worst consequence over transcripts) for a single position [58].

Once gene annotations have been added, additional knowledge about these genes can be appended. Such as for example Gene Ontology terms describing implicated cellular components, molecular function and biological processes [59]. Known clinical associations can be annotated using Human Phenotype Ontology terms [60] and the Online Mendelian Inheritance in Man gene mappings [61]. Also several databases such as KEGG [62], REACTOME [63] and the manually curated INTACT [64] can be used to annotate the biochemical pathways in which the gene participates. When a monoallelic dominant inheritance model is expected, one can annotate genes with haploinsufficiency predictions based on general properties of the gene (length, expression level, protein-protein interaction network centrality, occurrence of LOF variants or deletions in the healthy population) [19, 65].

2.2.2 Variant annotation

Individual variants can also be annotated with more fine-scaled annotation. A typically useful annotation when searching for causal variation is the variant's frequency (also known as Minor Allele Frequency or MAF) in the general population. If the disorder under study is rare and is assumed to have a monoallelic inheritance mode then filtering out common polymorphisms (e.g. having a frequency greater than 1% in the population) can dramatically reduce the number of candidate variants. Setting hard thresholds on the MAF can be problematic as it runs the risk of discarding pathogenic alleles segregating in the population at low yet appreciable frequencies. This is especially the case when dealing with recessive disorders, where allelic carriers might not suffer from any severe phenotype [66]. Several large scale genotyping studies of healthy populations have been undertaken and provide publicly available databases of variants. Most notably are dbSNP [67](though often considered contaminated with pathogenic variants), the 1000Genomes Project [17] and the Exome Sequencing Project (ESP) [18]. Either a combination of these databases can be used in order to filter common variants or a choice can be made based on the ethnic proximity of the control population to the studied cases.

Other informative variant-level annotations are quantitative measures of evolutionary conservation. Functionally deleterious variants are often observed in sites which have undergone high degrees of purifying evolutionary selection, removing any mutation events that might occur from the germ-line. This degree of selection can be quantified using different strategies such as Genetic Evolutionary Rate Profiling (GERP), phyloP and phastCONS (see [68] for a comparison of different conservation measures). When considering evolutionary conservation, it is important to take into account the desired *phylogenetic scope*, that is the evolutionary time in which the specific function might have evolved (e.g. from yeast to mammals, from primates to humans). For example, it would not be informative to look at evolutionary conserved sites from yeast to humans when looking for genomic positions involved in cognitive development [69, 70].Additionally evolutionary measures will not be suitable in for the detection of causative variants resulting in gain-of-function in previously non-functional sequences or when the variant is located in rapidly evolving sequences [66].

More involved approaches can be used to predict the deleteriousness of protein-altering variants. One might consider the relative position of nonsense or frameshift variants, assuming that these variants would have more moderate effects if residing at the end of the protein-coding sequence [19]. The effect of nonsynymous variants however is much harder to predict, compared to truncating mutations. The amino-acid change here might be located in a

functionally non-relevant part of the protein or the biochemical nature of the change might not be radical enough to cause any significant perturbation to function. Approaches such as Polyphen2 [71], MutationTaster2 [72], MAPP [73], PANTHER [74], PMUT [75], SNAP [76], PhD-SNP [77] try to combine several properties of variant such as biochemical change (*hydrophobic* $\leftrightarrow$ *hydrophilic*), protein structure models and evolutionary conservation. They achieve this either by training a machine learning model on a training set of known deleterious variants or by simple heuristics. One or a combination [78] of these tools can then be used to filter for variants predicted to be deleterious. Although most of these tools show reasonable benchmark performance, overlap between tools is often poor [18] and even taking a majority vote approach yields roughly 15% of all variants being called deleterious (i.e. 1200 variants). These deleterious variants are likely rare, mildly deleterious variants which have emerged in recent generations due to an expansive population growth and only being subject to weak purifying selection [79]. Because these predictions don't discriminate this burden of mildly deleterious variants from the true phenotypically causative variant, further discriminatory power is needed.

2.2.3 Phenotype Annotation

In addition to annotating useful information at the gene or variant level it is often useful to annotate particular patients/samples with a thorough description of their respective phenotypic presentation. Traditionally the clinical geneticist provides a description of the disease presentation and important developmental milestones (e.g. age at first words, age at first walk) in natural text form. This however is makes systematic computational analysis difficult (e.g. computing similarities between patients [161, 195]) due to variability in the descriptions between different clinicians. Standardized conceptualization in the the form of structured ontologies can prove useful as this permits reproducible phenotypic annotations and allows to formally define relationships between related disease terms (e.g. "ventricular septal defect" is a "septal defect" which in its turn is an "abnormality of the heart"). Additionally cross-ontology relationships can be defined (e.g. gene-phenotype relationships) allowing easy information retrieval and pattern recognition. Several disease ontologies have been developed, most notably the Human Phenotype Ontology (HPO) [60] which comprises of over 11,000 terms describing phenotypic abnormalities and their semantic relationships. Other more domain-specific ontologies focus on specific areas of disease such as the European Paediatric Cardiac Coding (EPCC) ontology which specifically describes cardiac phenotypes.

2.2.4 Experimental design

Even after substantive filtering for rare variants with highly damaging consequences many candidate variants remain in a single exome. Additional evidence is needed in order to restrict the number of candidates to pursue in cost and labour prohibitive follow-up validation studies. One way of restricting the search space is to test for association of a given gene in a cohort of unrelated individuals with the same phenotype compared to a population of "healthy" individuals. We present such a study in Chapter 3. Statistical significance of association can be computed as a two-sample test (e.g. Fisher's exact test) for the number of cases containing a mutation, post filtering, in a given gene compared to the number of controls containing a mutation in the same gene. Due to the large number of genes to be tested, multiple testing correction (e.g. Bonferonni or False Discovery Rate) is advised. Special care is needed to assure that both the case and control cohort populations have comparable coverage across all the tested genes, are ethnically similar and have been processed using identical processing pipelines. These tests also assume strict monogenic causality, which might be confounded by clinically hard to distinguish phenotypes or disorders with oligenic causality (e.g Bardet-Biedl syndrome [80]). Oligogenic causality remains a difficult problem to tackle due to the combinatorially explosive number of interactions to be considered, leading to a large-scale multiple testing problem. Nevertheless, recent studies in large cohorts of cases can detect implicated functional processes by testing variants at groups of genes collapsed by functional annotations or by network proximity to detected *de novo* variation [81, 82].

When familial samples are available, information about the segregation of variants in affected and/or unaffected individuals can be an extremely powerful methods to aid in the identification of causative variation. If several individuals are affected within a single family, one can look through linkage analysis for a shared haplotype between all affected individuals which is not shared in the unaffected individuals. When a recessive inheritance model is expected, due to for example consanguinity, runs of homozygosity can point to the region of interest. See Bamshad *et al.* for several examples where these strategies have been applied [83]. In recent years, exome sequencing of parent-child trios has been an extremely effective method of identifying causative *de novo* mutations, specially in sporadic cases of disease. In this setup, variants which are present only in the child but have not been inherited from either parent are detected. As only very few of these mutational events happen from generation to generation, the chance of finding several of these variants in a single gene/pathway across several individuals with the same phenotype is extremely low by random chance [10]. This approach has been used to discover the cause of many Mendelian disorders (e.g. Schinzel-Giedion syndrome [84], Kabuki syndrome [85], Miller

syndrome [86] amongst others) and more genetically heterogenous diseases such as intellectual disability [87] and schizophrenia [81, 82]). For a more in depth review of the study of *de novo* mutations in human disease we refer to Veltman *et al.* [88].

2.2.5 Gene Prioritization

Previously mentioned strategies offer an unbiased analysis across the exome. However, due to limited resources (e.g. limited sequencing capacity, limited number of available case samples, no available matched control population, no familial samples), one might have to restrict the number of tests to a selection of genes in order to obtain sufficient statistical power to detect. Expert knowledge can be used to select a subset of potentially interesting genes by looking at the published literature or association to known involved genes through sequence similarity, biochemical pathways, coexpression, protein-interaction networks and other sources. Making an informed decision based on all these sources can be extremely time-consuming, is prone to observer bias and is becoming more unfeasible due to the vast amount of data available in the omics-era. Several computational strategies have been developed in order to automate this process by prioritizing the genes in the genome according to a "guilt-by-association" principle, where some measure of similarity to known disease-associated genes is computed across many different data sources [89, 90, 91].

Gene prioritization methods usually have two inputs: a set of candidate genes to be prioritized (e.g. genes in a linkage region, the whole genome) and a set of criteria used for prioritization. These criteria can be formulated in a set of biological function or disease terms (e.g. Gene Ontology), or directly into a set of 'seed' or 'training' genes that are linked to the disease of interest. The algorithm then prioritizes the set candidates based on putative links to the known-associated genes in the different data sources. For example a gene might be prioritized high if it interacts in a protein complex in the protein-interaction network with one or several of the seed genes. The methodological challenge of integrating many heterogeneous, and often noisy, data sources in a reliable way is great, and has been tackled in many different ways [92, 93]. In this thesis we use an in-house developed tool called Endeavour [94, 95] which integrates information concerning sequence similarity, co-expression, tissue-specific EST expression, CIS-regulatory motifs, several functional annotations (Gene Ontology, SwissProt), protein domains (InterPro) protein-protein interaction networks (STRING,HPRD, BioGRID, IntnetDB), text mining and biological pathways (KEGG). The method with which Endeavour computes scores to seed genes depends on the type of data for each source. For attribute-based data (Gene Ontology, EST expression, InterPro

and KEGG) it computes over-representation of terms using a binomial statistic, resulting in a multiple-testing corrected p-value. For vector-type data sources (literature, co-expression, regulatory motifs) it constructs an average vector profile for the seed genes and computes the Pearson correlation between the candidate gene and the average vector. Lastly, for some data sources (BLAST, PPI), Endeavour computes different similarity measures such as the e-value or network neighbour similarity scores. These similarity measures are transformed to ranks normalized by the number of genes being ranked. Different ranks across data sources are then merged into a single global rank using an Order Statistics approach by computing the probability that the ranks for a gene are generated under a null hypothesis that the ranks of the different sources come from a random process. The final result is then a ranked list of genes according to the their similarity to the list of seed genes. A top-ranking set of genes can then be used to design experiments targetting a limited number of genes, or to select candidate variants in top-ranking genes for further functional validation such as zebra fish knock-down modelling, mice knock-outs or yeast molecular functional assays.

2.3 Thesis overview

This thesis starts with a broad introduction of human genetics and clinical genomics in Chapter 1 and the methods used in Chapter 2.

Chapter 3 presents a typical case study in a rare Mendelian disorder called Nicolaides-Baraitser syndrome. Here an exome-sequencing strategy is shown in a group of unrelated individuals in order to successfully identify the cause of the disorder. After standard genome alignment, variant calling and variant filtering several potential candidate genes were identified with potentially interesting variants across several patients. These variants were validated in parental samples and found to be *de novo* mutational events and further validated in an additional cohort of families (which had not been exome-sequenced due to the costs of exome sequencing at the time) and in immunohistochemistry experiments.

Chapter 4 takes the lessons learned in Chapter 3 and proposes a web-tool called *Annotate-it* to facilitate the analysis of such studies. *Annotate-it* handles variant annotation, experiment (meta-)data management and offers an easy-to-use visual filtering interface. Where in Chapter 3 several bioinformaticians were needed in order to perform the data analysis, *Annotate-it* allows this process to be performed by non-technical users, such as clinical geneticists, allowing for quick and iterative hypothesis checking for different inheritance models. It

retrieves and visualizes all the information needed for the geneticist to interpret the wealth of genomic data at hand.

Where *Annotate-it* was one of the first tools to consider phenotypic information (through OMIM, GO, HPO, text-mining and Endeavour prioritization annotation) in order to interpret variants, this information was generally secondary to the multisample analysis methodology (either through familial or association analysis or a combination of both). In Chapter 5 we push this concept further by proposing a new tool called *eXtasy* which integrates gene prioritization, deleteriousness prediction scores and haploinsufficiency prediction using a Random Forest machine learning approach in order to propose the first phenotype-specific variant prioritization tool. *eXtasy* can be used even when only a single sample is available and/or when no familial data is available. We show that this phenotype-specific methodology substantially outperforms state-of-the-art prediction algorithms in discerning disease-causing variation from rare putatively mildly deleterious variation.

Most Mendelian disorders do not usually present themselves as a single phenotype but ares often a combination of several phenotypes. In Chapter 6 we expand *eXtasy* to tackle this problem by comparing different aggregation methods for multiphenotype variant-prioritization.

We end this thesis in Chapter 7 where we summarize our findings and propose future directions for follow-up research.

Chapter 3

Case study: Nicolaides Baraitser Syndrome

L'absurde, c'est la raison lucide qui constate ses limites

Albert Camus

3.1 Summary

Nicolaides-Baraitser syndrome (NBS) is an extremely rare Mendelian disorder characterized by sparse hair, distinctive facial morphology, distal-limb anomalies and intellectual disability. No causes of NBS were identified prior to the presented study. By exome-sequencing of ten unrelated individuals suffering from NBS we identified heteryzygous *de novo* mutations in the *SMARCA2* gene. The identified mutations are clustered in the helicase domain of the protein, indicating a potential gain-of-function molecular mechanism of disease. Shortly after publication, several studies found mutations in another partner of the SWI/SNF chromatin remodelling complex, ARID1B, causing Coffin-Syris syndrome. These findings point to the importance of the SWI/SNF complex in early development and its implication in rare genomic disorders.

3.2 Case study: Nicolaides Baraitser Syndrome

At the time the presented study was conducted, computational tools for genome alignment, variant calling and variant annotation with short-read sequencing were still in their infancy. Tool selection and parameter settings resulted in substantially differing variant calls. The PhD candidate compared several of the publicly available tools for alignment and variant calling. Additionally, early prototype versions of the tools presented in Chapter 4 were developed and used in the analysis of the NBS data, ultimately helping in the discovery of the causative gene. *Annotate-it* was used as a platform for disseminating the underlying data to the broad public.

This paper was published in *Nature Genetics*, 2012, Feb 26;44(4):445-9, S1. doi: 10.1038/ng.1105.

Heterozygous missense mutations in *SMARCA2* cause Nicolaides-Baraitser syndrome

Jeroen K J Van Houdt[1,39], Beata Anna Nowakowska[1,2,39], Sérgio B Sousa[3,4,39], Barbera D C van Schaik[5,39], Eve Seuntjens[6,7], Nelson Avonce[8,9], Alejandro Sifrim[10], Omar A Abdul-Rahman[11], Marie-José H van den Boogaard[12], Armand Bottani[13], Marco Castori[14], Valérie Cormier-Daire[15], Matthew A Deardorff[16], Isabel Filges[17],Alan Fryer[18], Jean-Pierre Fryns[1], Simone Gana[19], Livia Garavelli[20], Gabriele Gillessen-Kaesbach[21], Bryan D Hall[22], Denise Horn[23], Danny Huylebroeck[6,7], Jakub Klapecki[2], Malgorzata Krajewska-Walasek[24], Alma Kuechler[25], Matthew A Lines[26], Saskia Maas[27], Kay D MacDermot[28], Shane McKee[29], Alex Magee[29], Stella A de Man[30,31], Yves Moreau[10], Fanny Morice-Picard[32], Ewa Obersztyn[2], Jacek Pilch[33], Elizabeth Rosser[34], Nora Shannon[35], Irene Stolte-Dijkstra[36], Patrick Van Dijck[8,9], Catheline Vilain[37], Annick Vogels[1], Emma Wakeling[28], Dagmar Wieczorek[25], Louise Wilson[34], Orsetta Zuffardi[38], Antoine H C van Kampen[5], Koenraad Devriendt[1], Raoul Hennekam[27,40] & Joris Robert Vermeesch[1,40]

[1]Center for Human Genetics, Catholic University Leuven, University Hospital Gasthuisberg, Leuven, Belgium. [2]Department of Medical Genetics, Institute of Mother and Child, Warsaw, Poland. [3]Clinical and Molecular Genetics Unit, Institute of Child Health, London, UK. [4]Serviço de Genética Médica, Hospital Pediátrico de Coimbra, Coimbra, Portugal. [5]Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherlands. [6]Laboratory of Molecular Biology (Celgen), Center for Human Genetics, Catholic University Leuven, Leuven, Belgium. [7]Department of Molecular and Developmental Genetics, Flanders Institute for Biotechnology, Leuven, Belgium.[8]Department of Molecular Microbiology, Flanders Institute for Biotechnology, Leuven, Belgium. [9]Laboratory of Molecular Cell Biology, Catholic University Leuven, Leuven, Belgium. [10]Department of Electrical Engineering (ESAT), Catholic University Leuven, Leuven, Belgium. [11]Division of Medical Genetics, Department of Pediatrics, University of Mississippi Medical Center, Jackson, Mississippi, USA. [12]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands. [13]Department of Genetic and Laboratory Medicine, Geneva University Hospitals, Geneva, Switzerland. [14]Medical Genetics, Department of Molecular Medicine, Sapienza University, San Camillo-Forlanini Hospital, Rome, Italy. [15]Département de Génétique, Université Paris Descartes, Institut National de la Santé et de la Recherche Médicale (INSERM) U781, Hôpital Necker–Enfants Malades, Paris, France. [16]Division of Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [17]Division of Medical Genetics, Department of Biomedicine,

University Children's Hospital, Basel, Switzerland. [18]Department of Clinical Genetics, Alder Hey Children's Hospital, Liverpool, and Liverpool Women's Hospital, Liverpool, UK. [19]Medical Genetics, University of Pavia, Pavia, Italy. [20]Clinical Genetics Unit, Department of Obstetric and Pediatric, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS), Santa Maria Nuova Hospital, Reggio Emilia, Italy. [21]Institute of Human Genetics, Universitaet zu Luebeck, Luebeck, Germany. [22]Department of Pediatrics, University of Kentucky, Lexington, Kentucky, USA. [23]Institut für Medizinische Genetik, Humboldt-Universität, Berlin, Germany. [24]Department of Medical Genetics, The Children's Memorial Health Institute, Warsaw, Poland. [25]Institut für Humangenetik, Universitätsklinikum, Essen, Germany. [26]Department of Genetics, Children's Hospital of Eastern Ontario, Ottawa, Ontario, Canada. [27]Department of Pediatrics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. [28]North West Thames Regional Genetics Service, Kennedy-Galton Center, London, UK. [29]Northern Ireland Regional Genetics Service, Belfast City Hospital, Belfast, UK. [30]Department of Pediatrics, Amphia Hospital, Breda, The Netherlands. [31]Department of Clinical Genetics, Erasmus University Medical School, Rotterdam, The Netherlands. [32]Service de Génétique Médicale, Laboratoire Maladies Rares–Génétique et Métabolisme (EA 4576), Centre Hospitalier Universitaire (CHU) de Bordeaux, Bordeaux, France. [33]Department of Child Neurology, Medical University of Silesia, Katowice, Poland. [34]Department of Clinical Genetics, Great Ormond Street Hospital for Children, London, UK. [35]Clinical Genetics Service, City Hospital, Nottingham, UK. [36]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [37]Department of Clinical Genetics, Center of Human Genetics, Université Libre de Bruxelles, Brussels, Belgium. [38]Medical Genetics, IRCCS, Neurological Institute C. Mondino, University of Pavia, Pavia, Italy. [39]These authors contributed equally to this work. [40]These authors jointly directed this work. Correspondence should be addressed to R.H. (r.c.hennekam@amc.uva.nl) or J.R.V. (joris.vermeesch@uzleuven.be).

Abstract

Nicolaides-Baraitser syndrome (NBS) is characterized by sparse hair, distinctive facial morphology, distal-limb anomalies and intellectual disability. We sequenced the exomes of ten individuals with NBS and identified heterozygous variants in *SMARCA2* in eight of them. Extended molecular screening identified nonsynonymous *SMARCA2* mutations in 36 of 44 individuals with NBS; these mutations were confirmed to be de novo when parental samples were available. *SMARCA2* encodes the core catalytic unit of the SWI/SNF ATP-dependent chromatin remodeling complex that is involved in the regulation of gene transcription. The mutations cluster within sequences that encode ultra-conserved motifs in the catalytic ATPase region of the protein. These alterations likely do not impair SWI/SNF complex assembly but may be associated

with disrupted ATPase activity. The identification of *SMARCA2* mutations in humans provides insight into the function of the Snf2 helicase family. Nicolaides-Baraitser

Main

Nicolaides-Baraitser syndrome (NBS; MIM 601358) was first described in 1993 [96] but only recently has been well delineated [97]. Its main characteristics include sparse hair, typical facial morphology, short stature, microcephaly, brachydactyly, interphalangeal joint swellings, epilepsy and intellectual disability with marked language impairment (Figure 3.1). This syndrome occurs in individuals from various ancestry groups without substantial differences in the frequency of its occurrence between the sexes. No familial cases are known, with the exception of one pair of concordant monozygotic twins, and no parental consanguinity has been reported2, suggesting that NBS is caused by dominant de novo mutations. We collected DNA and clinical data from 22 of the 27 individuals previously described to have NBS and from 22 additional affected individuals (cases). Subjects were classified into two categories comprising 37 and 7 individuals that had high and low certainty of NBS diagnosis, respectively (Figure 3.1, Table 3.1 and Supplementary Table 1). Neither karyotype analysis in 36 cases nor microarray analysis in 28 cases revealed any notable abnormalities.

Initially, we sequenced the exomes of four individuals with a clinical diagnosis of NBS (NBS01–NBS04). Following targeted exome enrichment, we obtained 5.1–6.7 Gb of sequence data per individual by massively parallel sequencing. The mean exome coverage was 40-fold, with 80% of the exome covered at least 10 times. We focused on nonsynonymous variants, splice acceptor or donor site mutations and coding insertions and/or deletions (indels). Given the probable dominant, de novo nature of the disorder, we filtered out previously described SNPs using the dbSNP132 and 1000 Genomes Project databases and over 311 unpublished exomes. We selected genes in which at least three of the four affected individuals carried single previously undescribed nonsynonymous variants at different genomic positions, strengthening the likelihood that these variants were causative. Five genes were affected by distinct missense variants in three unrelated cases. To gather additional evidence, we performed exome sequencing in six additional individuals with NBS (NBS05–NBS10). *SMARCA2* (encoding SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin, subfamily a, member 2) was the only gene with previously undescribed nonsynonymous mutations present in eight of the ten affected individuals. The potential impact of the encoded amino-acid substitutions on protein structure and function was predicted to be damaging in all eight cases (Table 3.2). Sanger sequencing validated the heterozygous

Table 3.1: Summary of clinical findings in individuals with NBS examined in this study

Clinical features	**Cases with nonsynonymous *SMARCA2* mutation** ($n = 36$)	**Cases without nonsynonymous *SMARCA2* mutation** ($n = 8$)
Clinical diagnosis of NBS, reliable	34/36	3/8
Clinical diagnosis of NBS, possible	2/36	5/8
Prenatal growth retardation	10/34	2/7
Postnatal growth retardation	19/36	4/8
Intellectual disability (mild-moderate-severe/total)	3-9-24/36	0-1-7/8
Seizures	22/35	2/8
Microcephaly	19/35	3/8
Sparse hair	35/36	7/8
Increased skin wrinkling	18/36	2/8
Thick, anteverted alae nasi	32/36	6/8
Broad philtrum	31/36	8/8
Long philtrum	29/36	6/8
Large mouth	34/36	7/8
Thin upper vermillion	27/36	5/8
Thick lower vermillion	32/36	8/8
Prominent interphalangeal joints	28/35	4/8
Prominent distal phalanges	21/35	5/8
Short metacarpals and/or metatarsals	16/32	3/6



Figure 3.1: Photographs of four individuals with NBS in whom *SMARCA2* mutations were identified: (a,b) NBS08 (female), the first reported individual with NBS1, at 12 years (a) and 32 years (b). (c,d) NBS33 (male) at 9 years and 9 months (c) and 10 years and 11 months (d). (e–g) NBS21 (male) at 1 year (e), 1 year and 9 months (f) and 5 years and 3 months (g). (h–l) NBS13 (female) at birth (h), 1 year (i), 4 years (j), 14 years (k) and 21 years (l). Note the clinical variability and changing phenotype with age, with progressive facial coarsening. NBS13 shows many but not all classical features: note that sparseness of hair was present to a mild degree at a young age and was later no longer visible. We obtained written informed consent to publish clinical photographs from these individuals.

state of all detected variants and confirmed de novo origin in four individuals for whom parental samples were available.

SMARCA2 is located on chromosome 9p24.3 and consists of 34 exons. All eight mutations were located in the region of the gene encoding the ATPase domain (exons 15–25; 490 amino acids; Figure 3.2), which is 100% conserved in chimpanzee and mouse and 94.7% conserved in zebrafish compared to the human protein (Supplementary Fig. 1). We performed Sanger sequencing of exons 15 to 25 in 34 additional affected individuals and identified 27 nonsynonymous mutations and one splice-site mutation 2 bp downstream of exon 24 (Figure 3.2b and Table 3.2). None of these mutations was present in over 1,300 exomes of unaffected individuals (Supplementary Table 2). Taking these results together, we identified nonsynonymous *SMARCA2* mutations in a total of 36 of the 44 individuals studied. In each of the 15 individuals for whom DNA from both parents was available, the *SMARCA2* variant was confirmed to be de novo. The splice-site variant could only be assessed as absent in a single parent.

The 36 exonic mutations were clustered in exons 15 (n = 4), 18 (n = 11), 19 (n = 3), 24 (n = 4) and 25 (n = 14). We sequenced all remaining exons in mutation-negative cases, but the variants we detected in these individuals were also observed in the control exome group (Supplementary Table 2) and were therefore considered not to be causal. We could not distinguish any specific phenotypic differences between individuals with the different mutations, nor could we identify substantive specific differences between the mutation-positive and mutation-negative cases. Nevertheless, the overall subjective certainty of the clinical diagnosis was shown to be important, as mutations were found in 34 of 37 individuals with certain diagnosis and in only 2 of 7 individuals with uncertain diagnosis (Table 3.1).

SMARCA2 (previously known as *BRM*) is classified in the Snf2 family of helicase-related proteins that are characterized by the presence of the conserved Snf2 family ATPase domain [98] (Figure 3.2). These nuclear ATPases are the core enzymatic subunits of chromatin remodeling protein complexes. *SMARCA2* is the catalytic subunit of the main human *BRM*-associated factors (BAF) complex, with BAF belonging to the SWI/SNF family of ATPase-dependent chromatin remodelers that regulate gene expression, differentiation and development [99]. SWI/SNF proteins mediate gene expression by repositioning or removing nucleosomes, thereby making DNA more accessible to transcription factors and key cellular proteins, facilitating both the induction and repression of genes. In mammals, SWI/SNF family complexes are very diverse in subunit composition and, hence, are present in multiple forms. BAF is highly conserved in eukaryotes and is composed of multiple subunits, including either *SMARCA2* or *SMARCA4* (previously known as BRG1) as the central ATPase subunit. This subunit diversity suggests that different complexes may have tissue- and stage-specific roles during development and cell differentiation. The unique composition of BAF correlates grossly with the specific gene expression required for maintaining cell state. Although the exact mechanistic basis for this maintenance is unknown, exchange of *SMARCA4* for the *SMARCA2* subunit helps to drive the transition from pluripotency to multipotency [99, 100]. Somatic loss of both *SMARCA2* and *SMARCA4* has been reported in 15–25% of malignant cell lines and solid tumors, suggesting that they act as tumor suppressor genes [101, 102].

We evaluated the correlation of *SMARCA2* expression with some of the major NBS features (intellectual disability, microcephaly and hypotrichosis) by performing immunohistochemical studies of Smarca2 expression in the developing cerebral cortex and hair follicles of mouse embryos, using two independent antibodies to Smarca2 on paraffin-embedded sections. At embryonic day (E) 14.5 and E15.5, Smarca2 levels were high in the post-mitotic neurons of the cortical plate (Supplementary Figure 3.2a–c) and were generally low in the progenitor regions (ventricular and subventricular zones). Consistently,

high levels of Smarca2 gene expression have been reported in the cortical plate (GenePaint) at E14.5 (Supplementary Figure 3.2d). There was a decrease in expression around E16.5 (Supplementary Figure 3.2e,f), after which we found neurons that were intensely stained for Smarca2 in an area corresponding to layer 5 (Supplementary Figure 3.2g–i). In the adult cortex and hippocampus, Smarca2-positive cells resided in the cortex and hippocampus (Supplementary Figure 3.2j). Cells within the epithelium of developing early hair follicles produced Smarca2 (Supplementary Figure 3.2k,l), suggesting a possible function in hair follicle development. These data show that Smarca2 is expressed at sites consistent with the phenotypes observed in individuals with NBS [97].

Considering that deletions encompassing human *SMARCA2* do not cause NBS [103], that mice lacking functional Smarca2 do not present major developmental abnormalities [100, 104] and that none of the variants we identified were truncating, we predict that the mutations identified in NBS act in a dominant-negative or gain-of-function manner. All nonsynonymous *SMARCA2* mutations altered ultra-conserved amino acids (Figure 3.2b and Table 3.2), including seven helicase-related sequence motifs that define the SNF2 domain: motifs I, II, III and VI and the SNF2 domain D motif (Figure 3.2). Motifs I, II and III have been associated with nucleotide binding and motif D with DNA interaction and ATP hydrolysis [98]. In addition, seven mutations were localized to motifs that were not previously known to be highly conserved. The SNF2 domain is identical in both *SMARCA4* and *SMARCA2* and is highly conserved in yeast SNF2. In addition, a chimera of yeast SNF2 and human *SMARCA4* with the DNA-dependent helicase domain of SNF2 replaced by the corresponding human sequence restored normal mitotic growth and capacity for transcriptional coactivation in snf2-deficient yeast cells, proving functional conservation of this domain [105]. Mutations of yeast snf2 affecting conserved motifs resulted in dominant-negative activity in a functional assay [106], and two of these mutations are identical to *SMARCA2* mutations in individuals with NBS, while four map to the same motifs (Figure 3.2). This clustering of mutations provides genetic evidence that abolishing the ATP hydrolyzing engine, which provides energy directed toward the repositioning of histones on DNA, causes functional inactivation [107, 108]. Crystal structures of helicases have shown that all seven conserved motifs within the SNF2 domain lie in close proximity to the nucleotide-binding site and are thus probably involved in nucleotide binding and ATP hydrolysis [109, 110]. Structural and mutagenesis studies have shown that each of the conserved motifs has a role in the transformation of chemical energy from ATP hydrolysis to mechanical motion. These data, together with the mutations identified in individuals with NBS, support a model in which dysfunctional but structurally undamaged *SMARCA2* generates a BAF complex that is intact with respect to its composition and interacts properly with chromatin but is nonetheless functionally inactive, resulting in dominant-negative effects.

The identification of clustered mutations in a large number of individuals with NBS substantiates the notion that NBS is a distinct clinical entity. In addition, the collection of 44 individuals with NBS in a short time span since the syndrome's delineation shows that NBS is more common than was originally anticipated. Mutations in *SMARCA2* add it to a growing list of chromatin remodeling genes, including SMARCAL1 (causing Schimke immunoosseous dysplasia (SIOD); MIM 242900) [111] and CHD7 (causing CHARGE syndrome; MIM 214800) [99, 101, 112], that cause developmental abnormalities variably associated with intellectual disability. The CHD7 protein was recently shown to interact with *SMARCA2* in human neural crest-like cells [113]. We anticipate that mutations in genes encoding other helicase family members and BAF complex proteins will be identified as the cause of developmental disorders.

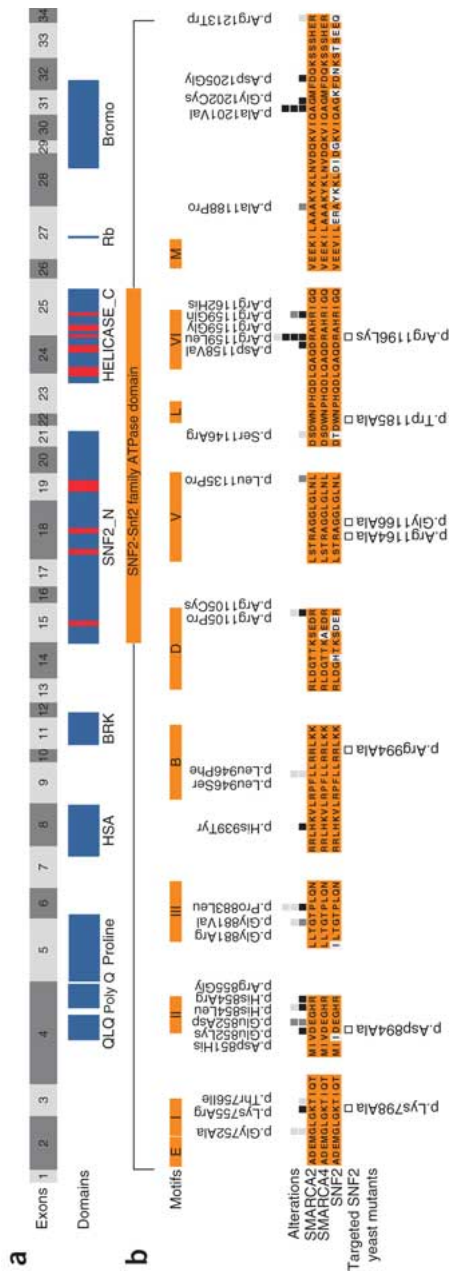


Figure 3.2: Exons, protein domains, conserved motifs and mutation spectrum for *SMARCA2*: (a) *SMARCA2* is composed of 34 exons that encode 9 conserved domains (blue) [102]. QLQ, glutamine-leucine-glutamine domain; proline, proline-rich domain; HSA, small helicase/SANT-associated domain; BRK, brahma and kismet domain; SNF2_ N and HELICASE_ C, DNA-dependent ATPase domains. The Snf2 family is characterized by seven helicase-related sequence motifs defining the ATPase domain. (b) Amino-acid alignment for the mutated conserved motifs in the catalytic ATPase domain for human *SMARCA2* and *SMARCA4* and yeast Snf2, showing the conserved structural motifs in Snf2 subfamilies [98]. The amino-acid substitutions in individuals with NBS are indicated with rectangles above the alignment (black, confirmed de novo; dark gray, mutation is absent in one available parent; light gray, no parental DNA available). Targeted alterations in yeast used for functional assays [106] are indicated by white rectangles below the alignment.

Material & Methods

Subjects

Subjects were recruited through genetics centers in Europe, Canada and the United States. Informed consent was obtained from all subjects, and the study was approved by the appropriate Institutional Review Boards involved in this investigation. The diagnosis of NBS was based on the typical morphology of face and limbs, radiographic data, growth pattern and cognitive development as published [97] (Supplementary Table 1). Affected individuals were classified into high- and lower-certainty groups by consensus among three clinical geneticists (R.H., K.D. and S.B.S.). The parental status was tested with the Profiler plus kit from Applied Biosciences.

Targeted enrichment and exome sequencing

Genomic DNA was sheared by sonication, platform-specific adaptors were ligated, and the resulting fragments were size selected. The libraries from subjects NBS01–NBS04 were captured using the SureSelect All Exon Target Enrichment System (Agilent, Human All Exon Kit, 38Mb) according to the manufacturer's protocols. The enriched libraries underwent paired-end (2 $\times$ 76 bp) sequencing on a Genome Analyzer IIx (Illumina). NBS05 and NBS06 libraries were enriched with the SeqCap EZ Human Exome Library v2.0 (Roche, NimbleGen), and 2 $\times$ 76-bp paired-end reads were generated on a HiSeq2000 (Illumina). The NimbleGen exome capture array (version 1) was used to produce the enriched libraries for NBS07–NBS10 that were sequenced on the Applied Biosystems SOLiD system.

Mapping and variant analyses

The paired-end sequence reads of NBS01–NBS06 were aligned to the human genome (hg19) with the Burrows-Wheeler Aligner (BWA; version 0.5.8a) [114] using default settings, and the read trimming parameter was set to 15. SAMtools (version 0.1.8) [43] was used for converting (SAM/BAM), sorting and indexing alignments. The quality metrics for mapping were calculated with Picard tools (version 1.38). Duplicate reads were marked with Picard tools and excluded from downstream analysis. The GATK framework (version 1.0.4974) [47] was used for performing the local realignment, base call recalibration and SNP calling. Variants with at least Q10 confidence were reported, but, if the confidence was less than Q50, it was reported as LowQual. Indels were called

Table 3.2: Annotation of *SMARCA2* mutations identified in individuals with NBS

					Confirmed *de novo*[b]				Predictions[d]			
Subject	Group	Exome sequenced	Mutation[a]	Exon	Mother	Father	Amino-acid change	Chr. 9 position[c]	S	L	M	P
NBS01	1	+	c.3637C>T	25	–	–	p.Arg1213Trp	2116002	D	D	D	D
NBS02	1	+	c.3604G>T	25	+	+	p.Gly1202Cys	2115969	D	D	D	P
NBS03	2	+										
NBS04	1	+	c.3476G>A	25	+	+	p.Arg1159Gln	2115841	D	D	D	P
NBS05	1	+										
NBS06	1	+	c.3473A>T	25	+	+	p.Asp1158Val	2115838	D	D	D	D
NBS07	1	+	c.3475C>G	25	+	+	p.Arg1159Gly	2115840	D	D	D	P
NBS08	1	+	c.2642G>T	18	+	–	p.Gly881Val	2086944	D	D	D	–
NBS09	1	+	c.3485G>A	25	+	+	p.Arg1162His	2115850	D	D	D	P
NBS10	1	+	c.3476G>T	25	–	–	p.Arg1159Leu	2115841	D	D	D	P
NBS11	2	–										
NBS12	1	–	c.2648C>T	18	+	+	p.Pro883Leu	2086950	D	D	D	–
NBS13	1	–	c.3476G>A	25	+	+	p.Arg1159Gln	2115841	D	D	D	P
NBS14	1	–	c.3602C>T	25	+	+	p.Ala1201Val	2115967	D	D	D	B
NBS15	1	–	c.2815C>T	19	+	+	p.His939Tyr	2088545	D	D	D	D
NBS16	2	–	c.2267C>T	15	–	–	p.Thr756Ile	2081914	D	D	D	P
NBS17	2	–	c.3456+2T>G	Intron[f]	+	–		2110419				
NBS18	1	–	c.3313C>T	24	+	+	p.Arg1105Cys	2110274	D	D	D	D
NBS19	1	–	c.2556A>C	18	+	–	p.Glu852Asp	2086858	D	D	D	P
NBS20	1	–	c.2641G>C	18	–	–	p.Gly881Arg	2086943	D	D	D	–
NBS21	1	–	c.3404T>C	24	+	–	p.Leu1135Pro	2110365	D	D	D	P
NBS22	1	–	c.3562G>C	25	+	–	p.Ala1188Pro	2115927	D	D	D	B
NBS23	1	–	c.3314G>C	24	–	–	p.Arg1105Pro	2110275	D	D	D	P
NBS24[e]	1	–	c.2255G>C	15	–	–	p.Gly752Ala	2081902	D	D	D	P
NBS25[e]	1	–	c.2255G>C	15	–	–	p.Gly752Ala	2081902	D	D	D	P
NBS26	1	–	c.2648C>T	18	–	–	p.Pro883Leu	2086950	D	D	D	–
NBS27	1	–	c.2554G>A	18	–	+	p.Glu852Lys	2086856	D	D	D	P
NBS28	1	–	c.2560C>A c.2562C>A	18	–	–	p.His854Leu	2086862	D	D	D	D
NBS29	1	–	c.2648C>T	18	–	–	p.Pro883Leu	2086950	D	D	D	–
NBS30	1	–	c.2837T>C	19	–	–	p.Leu946Ser	2088567	D	D	D	B
NBS31	1	no	c.3614A>G	25	–	–	p.Asp1205Gly	2115979	D	D	D	P
NBS32	1	no										
NBS33	1	no	c.2561A>G	18	+	+	p.His854Arg	2086863	D	D	D	D
NBS34	1	no	c.3602C>T	25	–	–	p.Ala1201Val	2115967	D	D	D	B
NBS35	1	no	c.2264A>G	15	+	+	p.Lys755Arg	2081911	D	D	D	P
NBS36	2	no										
NBS37	1	no	c.3436A>C	24	–	–	p.Ser1146Arg	2110397	D	D	D	P
NBS38	1	no			–	–						
NBS39	1	no	c.2838A>T	19	–	–	p.Leu946Phe	2088568	D	D	D	B
NBS40	2	no										
NBS41	2	no	c.3485G>A	25	+	–	p.Arg1162His	2115850	D	D	D	P
NBS42	1	no	c.3602C>T	25	+	+	p.Ala1201Val	2115967	D	D	D	B
NBS43	1	no	c.2551G>C	18	+	+	p.Asp851His	2086853	D	D	D	P
NBS44	1	no	c.2563C>G	18	+	+	p.Arg855Gly	2086865	D	D	D	P

[a]Mutation numbering is based on NM_003070. [b]For parental samples: +, confirmed not to be present in parental samples; –, no parental sample available. [c]Chromosomal positions based on build hg19 (Feb. 2009, UCSC Genome Browser). [d]Functional predictions retrieved from dbNSFP (S, Sift; L, LRT; M, MutationTaster; P, Polyphen2; D, probably damaging; P, possibly damaging; B, benign; –, no prediction). [e]Monozygotic twins. [f]Mutation in potential splice donor (GT) of intron 24.

with Dindel (version 1.01)22 using default parameters. Variants were annotated with ANNOVAR (version 2011) [115] and filtered against known variants by comparison to dbSNP131 and 1000 Genomes SNP data (release November 2010). The sequence reads for NBS07–NBS10 were aligned with BWA (version 0.5.7) to the reference genome (hg18) with default settings. SNPs and indels were called with SAMtools pileup (version 0.1.7) and Varscan (version 2.2), and annotated with ANNOVAR. Variants with an average base quality above 35, SIFT score below 0.05 and variant frequency above 15 were kept for further analysis. Variants that occurred in segmental, duplicated regions or were non-conserved (mce 44), synonymous or known (in dbSNP132 or 1000 Genomes, version July 2010) were excluded from the analysis.

Searches for known variants were performed on 1000 Genomes SNP data (release November 2010), 223 individuals from the Genome of The Netherlands Project (GoNL) and 86 in-house sequenced exomes). In the GoNL Project, individuals (father-mother-child) were sequenced at the Beijing Genomics Institute (BGI) with an average sequence depth of $>12\times$. The GoNL variants for all genes in all individuals were annotated with ANNOVAR on the Dutch grid (Big Grid) using the e-BioInfra framework [116]. The *SMARCA2* variants were extracted from this data set to verify whether the mutations found in NBS were previously unknown. In total, we found ten *SMARCA2* variants in the GoNL data set, of which eight were known from the public databases and two were new (Supplementary Table 2). None of the NBS-related mutations were present in the GoNL data set.

Variant confirmation

Primers were designed using Primer3 software. PCR products were purified with ExoSAP-IT (GE Healthcare) and sequenced using BigDye Terminator v3.1 chemistry (Life Technologies) on a 3730 DNA Analyzer (Life Technologies). Sequence traces were aligned to the *SMARCA2* reference sequence using SeqScape software (v2.6) (Life Technologies). Functional predictions for the amino-acid changes according to different models (SIFT, Polyphen2, LRT and MutationTaster) were retrieved from dbNSFP (database for predictions for nonsynonymous SNPs; Table 2) [117].

Immunohistochemistry and microscopy

Timed pregnant Swiss females (with E0.5 being the morning of the day of vaginal plug) were dissected and mouse embryos (E14.5) or brains (E15.5–E16.5) were collected and washed in ice-cold PBS, fixed overnight with

4% paraformaldehyde and submitted to progressive alcohol-assisted dehydration and paraffin embedding. Frontal (E15.5 and E16.5) or sagittal (E14.5) 6-µm thick sections were processed for immunohistochemistry using an automated platform (Ventana Discovery, Ventana Medical Systems; details of procedures can be obtained on request) and were dehydrated and mounted with Eukitt (Sigma). Studies were performed using two different primary rabbit antibodies to Smarca2, one from Abcam (ab15597; 1:60 dilution) and one from Sigma (HPA029981; 1:300 dilution). Photographs were taken on a Leica DM5500B microscope.

Accession codes

Exome sequencing data are available upon request to J.R.V. Exome variant data are available at the Annotate-it! website.

Acknowledgements

We thank the Genome of the Netherlands Project (GoNL) for providing their variant data. GoNL is one of the rainbow projects of the Dutch hub of the Biobanking and Biomolecular Research Infrastructure (BBMRI-NL). We thank S. Olabarriaga and M. Santcroos for their support in using the Dutch Grid for data analysis. We thank T.J.L. de Ravel and P. Brady for critical reading of the manuscript and G. Peeters for technical support. We are grateful to the subjects and their families for participating in this study. This work was made possible by grants from the Agency for Innovation by Science and Technology (IWT; SBO-60848), the Catholic University of Leuven (PFV/10/016 SymBioSys and GOA/12/015 to J.R.V., Y.M. and K.D.), the Queen Elisabeth Medical Foundation (GSKE 1113 to D. Hu and E.S.) and the type 3 large-infrastructure support InfraMouse by the Flanders Hercules Foundation (to D. Hu). B.A.N. is supported by a KOLUMB fellowship from the Foundation for Polish Science. S.B.S. was supported by the Fundação Para a Ciência e Tecnologia (SFRH/BD/46778/2008).

Supplementary Material

Supplementary Figures and Tables are available at
`http://http://www.nature.com/ng/journal/v44/n4/extref/ng.1105-S1.pdf`

Chapter 4

SNV data management, annotation and interpretation

> My dear, here we must run as fast as we can, just to stay in place.
> And if you wish to go anywhere you must run twice as fast as that.
>
> Lewis Carroll

4.1 Summary

The analysis process of tens of thousands of genetic variants, as discussed in Chapter 3, is a complex process. Filtering and cross-querying data-sets requires technical expertise and no easy-to-use tools existed at the time when this study was conducted. Additionally, because of the rapidly changing annotation sources and the need to integrate many samples and data sources, data management was (and still is) becoming one of the major issues in the analysis of exome-sequencing projects. In order to alleviate these hurdles we developed a tool called *Annotate-it* which handles data management aspects of exome-sequencing projects, centralizes annotation and provides an intuitive interface for querying and filtering the data, usable by clinical geneticists (Figure 4.1.)

4.2 SNV data management, annotation and interpretation

The PhD candidate conceptually designed the tool in close collaboration with clinical geneticists and bioinformaticians from the University Hospital in Leuven and in the context of the HeartRepair International Consortium. He served as the main developer of the tool, programming the annotation and database back-end, the web front-end, and several interactive visualization interfaces. He deployed and refined the tool to suit several currently ongoing exome-sequencing projects.

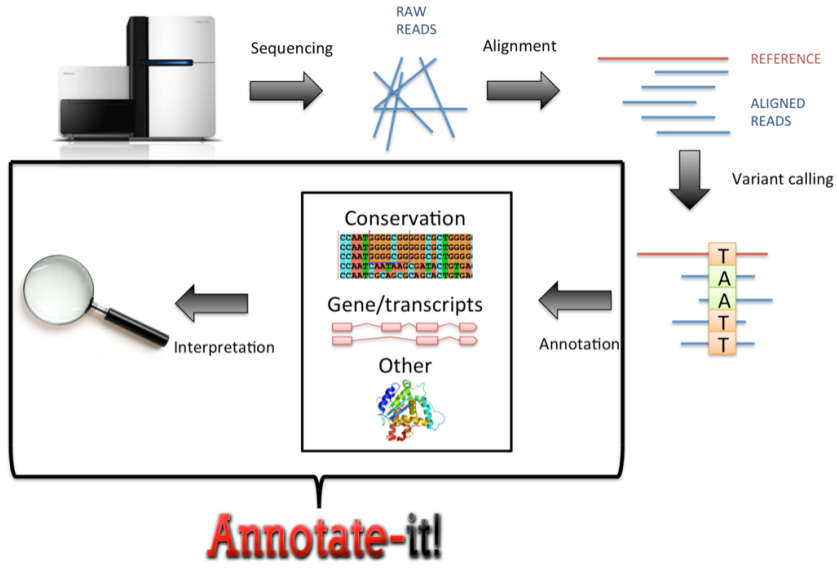


Figure 4.1: Overview diagram of Annotate-it: When a sequencing experiment is performed, first a preprocessing pipeline processes the raw information coming from the sequencer machine. In these steps the sequencer output, in the form of raw sequence reads, are aligned to the reference genome to determine their genomic coordinates. Once all reads are aligned, genomic variants (i.e. positions which differ from the reference genome) can be determined in a process called "variant calling". This variant information is usually summarized in a file in the Variant Calling Format (VCF). These files are difficult to handle and interpret due to their size and their lack of functional biological information. Annotate-it annotates such files with relevant information (e.g. evolutionary conservation, gene annotations, known disease associations and many more) and offers an easy-to-use point-and-click user interface which allows end-users to easily filter and compare variant sets according to different experimental designs, in order to facilitate the interpretation of these complex datasets.

This paper was published in *Genome Medicine*, 2012, Sept, 4:73 doi:10.1186/gm374

Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease

Alejandro Sifrim[1,2§], Jeroen KJ Van Houdt[3], Leon-Charles Tranchevent[1,2], Beata Nowakowska[3], Ryo Sakai[1,2], Georgios A Pavlopoulos[1,2], Koen Devriendt[3], Joris R Vermeesch[3], Yves Moreau[1,2], Jan Aerts[1,2]

[1]KU Leuven, Department of Electrical Engineering-ESAT, SCD-SISTA, Kasteelpark Arenberg 10, B-3001, Leuven, Belgium
[2]IBBT Future Health Department, Kasteelpark Arenberg 10, B-3001, Leuven, Belgium
[3]KU Leuven, Centre for Human Genetics, University Hospital Gasthuisberg, Herestraat 49, 3000 Leuven, Belgium
[§]Corresponding author

Abstract

The increasing size and complexity of exome/genome sequencing data requires new tools for clinical geneticists to discover disease-causing variants. Bottlenecks in identifying the causative variation include poor cross-sample querying, constantly changing functional annotation and not considering existing knowledge concerning the phenotype. We describe a methodology that facilitates exploration of patient sequencing data towards identification of causal variants under different genetic hypotheses. Annotate-it facilitates handling, analysis and interpretation of high-throughput single nucleotide variant data. We demonstrate our strategy using three case studies. Annotate-it is freely available and test data are accessible to all users at `http://www.annotate-it.org` . •

Introduction

Context

With the advent of massively parallel high-throughput sequencing technologies and the increasing availability of reference genomes, new opportunities emerge

for discovering genome-wide variation across individuals and populations, at both the large-scale level (deletions, duplications, and rearrangements) and the base-pair level (single nucleotide variants and small indels and repeats). As more and more sequence data are produced, accurate assessment of the frequency of variants in specific sub populations (patients versus controls or any phenotypically different populations) is vital to the interpretation of how these variants segregate across populations. International projects, such as the 1000 Genomes Project [16] and the Hapmap Project [118], have been set up to assess the genetic variation across large groups of 'normal' human individuals. Full-genome [119], whole-exome [120, 86, 85] or targeted gene panel [121, 122] sequencing studies of individuals struck by Mendelian disorders can aid in identifying the genetic cause for these diseases; for example, by leveraging publicly available data, under the assumption that these rare variants do not occur in the normal population. Furthermore, trio sequencing of an affected child and of his or her parents can identify de novo variants in sporadic cases of genetic disease [87, 123].

Because of the large size and complexity of next-generation sequencing data sets, new computational and statistical methods for analysing and interpreting the data are required to accurately find the variation of biological interest. To transform raw sequencing data into variation data, we need to undertake the following steps: (1) sequence alignment, (2) variant calling, (3) variant annotation, and (4) variant interpretation [40]. In this study, we mainly focus on the latter two steps. After sequence alignment and variant calling, we end up with a list of variants with their genomic coordinates and the variant alleles that differ from the reference sequence. Based on current knowledge of functional elements annotated in the human genome sequence, overlapping variants within the annotated features are found and the impact on RNA and protein sequence level is computed. Variant lists can be reduced further by applying functional impact prediction tools, such as Polyphen2 [71], SIFT [124], FoldX [125], and others [126]. Such tools are computationally intensive and require dedicated computing infrastructure to run variants in a high-throughput automated fashion. To tackle this problem, precomputed whole-genome predictions for popular tools are available in Ensembl and in dedicated databases [101]. Aggregated functional impact scores, such as CAROL [127], are also useful as these usually perform better in classification benchmarks for disease-causing mutations but are often not provided by current annotation tools.

These annotated variants can then further be interpreted by filtering based on either quality, functional, or genetic criteria. These filtered annotated data are then interpreted by either collapsing the data into the desired functional units, often genes, or performing formal statistical association analysis methods [128].

Several software tools have already been developed to tackle the aforementioned

analysis steps. Some of these tools, such as ANNOVAR [115], TREAT [129], VarioWatch [130], SeqAnt [131], and Anntools [132], have been specifically designed to handle the annotation task and use diverse data sources in order to achieve this (Table 4.1). Other tools focus solely on the interpretation step and facilitate tasks such as easy filtering of data (VarSifter [133]), ranking based on putative pathogenicity (Var-MD [134]) or do more complex types of analysis by performing association analyses or looking at different underlying genetic disease models (VAAST [135]). A variety of tools try to be more comprehensive and provide a streamlined approach by tackling both steps (KGGSeq [136], SVA [137]).

Although current tools have matured significantly, they still possess several limitations. First there is the 'red queen' annotation problem. As functional information is constantly evolving, end users are forced to keep data sources and variant annotation up to date with every release of new information. Although some annotation tools make it easy to update annotation sources, this can still be a problem when trying to use previously acquired data sets in a related setting without repeating the complete annotation process. Secondly, interpretation is often done at the level of the individual sample, whereas cross-sample analysis is often left to the end user. This requires computational expertise when dealing with large data sets. Additionally different biological contexts might require a different combination of analysis techniques (that is, case/control studies versus familial or trio studies). Also, these analysis techniques are often phenotype-naive and disregard existing a priori biological knowledge of the particular disease or phenotype under study. Thirdly, when working in a collaborative setting, keeping track of individual files, annotation versions, and variants can prove difficult and easy sharing of variation, validation, sample, and experiment metadata is usually crucial. This is especially the case in the analysis of Mendelian disorders where cases are often spread globally across research institutions.

Aim

To tackle these hurdles, we developed a framework, called Annotate-it, that provides experimentalists with a Swiss-knife approach for the interpretation of single-nucleotide variants, providing features such as automated annotation, prioritization of mutated genes, cross-sample querying and data management. We expand on existing comprehensive annotation and interpretation frameworks by integrating more annotation sources and different established state-of-the-art analysis techniques. We also integrate two different gene prioritization techniques: AGeneApart [138] and Endeavour [94]. AGeneApart mines MEDLINE abstracts to discover known genes or potentially new genes linked

Table 4.1: Comparison of features of different tools for next-generation sequencing annotation and interpretation

	ANNOVAR	TREAT	VAAST	VarSifter	Var-MD	KGGSeq	SVA	Anntools	VarioWatch	SeqAnt	Annotate-it
Interface	Command line	Command line	Command line	Graphical	Python script	Command line	Graphical	Command line	Web	Web	Web
Indels	x	x	x	x		x	x	x		x	
SNPs	x	x	x	x	x	x	x	x	x	x	x
miRNA Annotation	x							x			x
Regulatory Region Annotation	x	x				x		x			
Custom Annotation	x						x	x			
Gene annotation	x	x				x	x	x	x	x	x
Conservation scores	x	x				x				x	x
LRT	x					x					x
MutationTaster	x					x					x
Polyphen2	x	x				x					x
SIFT	x	x				x			x		x
Aggregate scores						x					x
Haploinsufficiency prediction											x
OMIM		x					x		x		x
NHLBI Exomes	x										x
200 Danish Exomes											x
1000Genomes	x	x				x	x	x	x	x	x
dbSNP	x	x				x	x	x	x	x	x
Gene Prioritization											x
Literature retrieval						x			x		x
Gene Ontology							x		x		x
PPI/Complexes						x					x
Pathway information		x				x	x		x		x
Gene expression		x									x
Statistical Analysis			x				x				x
Filter	x	x		x	x	x	x				x
Complex Filtering Rules				x		x	x				x
Cross-Sample Querying				x	x	x	x				x
Genetic Inheritance Models			x	x	x	x	x			x	x
Data sharing											x
Visualizations		x					x		x	x	x

to the phenotype whereas Endeavour ranks genes based on similarity profiling compared to known disease genes. The latter method is more suitable for finding novel disease genes as it does not require a direct relationship between gene and disease [139]. These methods allow us to rank mutated genes taking the phenotype into account. Finally, because of its web-based interface, we allow for a collaborative workflow and the management of shared or public data, as well as a wide range of visualizations.

Implementation

Annotation and filtering methodology

Multiple annotation sources are supported, which can be classified into gene level and variant level information. Genes are annotated using Gene Ontology terms, Online Mendelian Inheritance in Man (OMIM) and haploinsufficiency predictions [65]. We also provide pathway information from KEGG [140], BIOCARTA, and Reactome [141], as well as protein-protein interaction and protein complex data from STRING [142] and the Corum project [143]. Additionally, we provide statistical phenotypic associations with known chromosomal aberrations and phenotypic ontologies based on text-mining of recent MEDLINE releases [138]. Tissue-based expression information from the eGenetics/SANBI data set is also included to search for genes that are expressed in particular tissues of interest.

Information on the individual variants include PhastCons and Phylop scores (for primate, placental mammal and vertebrate multiple alignments) and presence in dbSNP, 1000 Genomes Project, and the 200 publicly available Danish exomes [144]. Also functional impact prediction scores from SIFT, Polyphen2, LRT [145], and MutationTaster [146] are extracted from dbNSFP. An aggregate score of SIFT and Polyphen2 is also computed with CAROL, which has been demonstrated to perform better than individual scores at predicting deleteriousness of nonsynonymous variants.

Annotate-it has a flexible query system that allows answering complex cross-sample questions in an easy and quick way without any need for computational expertise. Using this interface, users can quickly check different inheritance hypotheses (for example, recessive or dominant models, or de novo occurrence of variants in parents-child trio analysis) depending on the available familial information. It also permits filtering based on consequence type, minimum and maximum coverage, minimum and maximum variant allele frequency, heterozygosity call, presence in dbSNP, 1000 Genomes Project, the NHLBI exome variant server and 200 Danish exomes, variants present in other samples, and conservation scores (Additional file 1). Annotate-it also allows screening a

user-defined (or automatically generated) panel of genes of interest. Ultimately, these filtering steps will result in a list of genes (or exons) ordered by user-specified criteria (for example, present in given number of samples, number of unique alleles) (Additional file 2). Genes can then be further inspected by consulting all the available information that Annotate-it aggregates. This information contains associated phenotypes, automated literature retrieval, links to OMIM, region-based conservation plots, STRING or protein complex interaction partners and gene ontology terms (Additional file 3).

Interpretation and statistical analysis

In addition to the widely used collapsing method of disease-gene discovery, where variants across affected samples are collapsed at the gene level and counted, the user can also perform formal statistical analysis using the methodology proposed by Ionita-Laza for case/control studies [128]. This type of statistical methods has the added benefit of computing an approximate P-value for individual genes and of taking into account inherent background variation in genes due to factors such as gene size or hypermutable regions. However, they do require a significant amount of control samples in order to reach satisfying statistical power ($n > 50$). The analysis can be applied directly on the imported data without any reformatting or performing any command line actions.

The user can also provide a particular phenotype using Human Phenotype Ontology terms that can be used to derive genome-wide rankings produced by AGeneApart and Endeavour. These rankings are precomputed for all Human Phenotype Ontology terms for which at least five gene-phenotype associations are known, which can be used as a training set for the prioritization algorithm.

Furthermore the user can investigate mutation patterns of interaction partners of genes of interest. This is achieved by integrating the STRING and CORUM protein-protein interaction networks. This can be of particular interest when the cause of disease is expected to be oligogenic, such as diseases with some degree of phenotypic variability.

Metadata and collaboration

Additional metadata can be linked to samples and experiments. Samples are entities within experiments and can contain phenotypic information from different integrated ontologies (for example, European Paediatric Cardiac Code, London Dysmorphology Database, Human Phenotype Ontology [60]), gender, the date of sequencing, batch number, and other data. Metadata allow

organizing and querying samples across multiple experiments. The experiments themselves can also store metadata, such as the software and parameters used for alignment and variant calling. Within an experiment, variants can be marked for validation, and validation information can be stored and reviewed. Promising candidate genes can be organized into gene lists when dealing with many collaborators or complex study setup Gene lists can also be used as masking filters to either study gene specific gene panels or exclude sets of genes that are not of interest. An experiment-wide permission system with different roles (owner, editor, user) allows collaboration amongst different researchers and also allows the public sharing of data.

Visual Analytics

Visual analytics is gaining ground for data exploration in the context of big data. The integrated interactive graphical filter (Figure 4.2) visualizes the different distributions of variants (for example, coverage, consequence types, conservation scores, variant allele frequency) and shows the impact of different filtering thresholds in real time, allowing selection of optimal thresholds. This is particularly useful in an exploratory phase to get a grasp on the characteristics of the data set and on how filter settings will affect these characteristics. Once optimal thresholds have been selected, they are easily applied in a subsequent discovery phase where the main goal is to select candidate genes of interest.

Calculating statistics for such large data sets in real-time is computationally taxing; therefore, Annotate-it's interactive filter can currently only be run on a subset of the variants (maximum 200,000 variants). Further research in the development of advanced data structures and browser infrastructure is needed to cope with the running time complexity of these approaches. As user interaction and feedback increases, we aim to optimize the informative value of such interconnected graphs and to improve the throughput of these visualization techniques.

Technical specifications

Annotate-it is programmed in the Ruby programming language and makes use of the Ruby on Rails framework [45] for its web interface. Currently, Annotate-it supports both hg18 (NCBI36) and hg19 (NCBI37) genomic builds. The annotation process leverages existing libraries, such as the Ruby Ensembl-API [147] and BioRuby [148], to enable programmatic access to locally run mirrors of the latest Ensembl [149] versions for any given genomic build. MySQL was chosen to store variant information in a database. Nevertheless, Annotate-

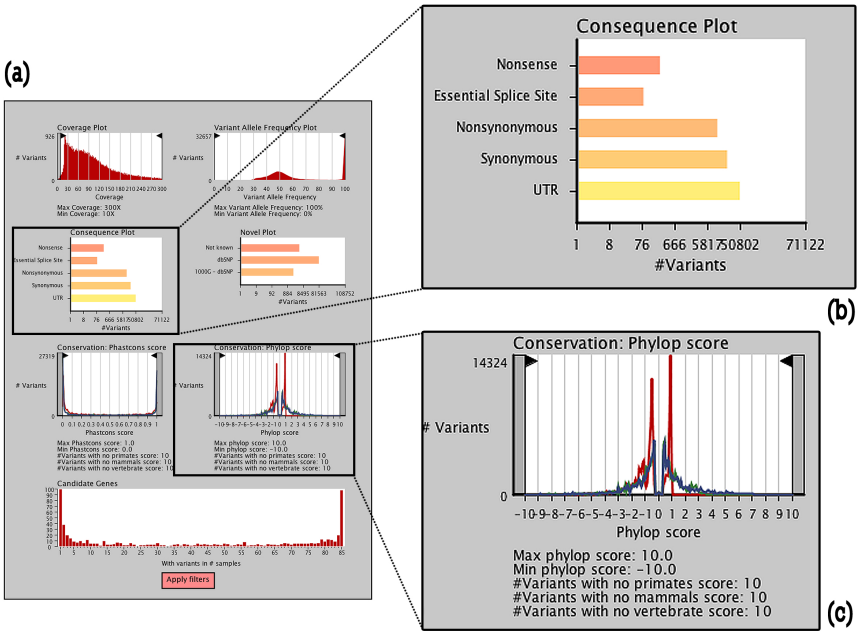


Figure 4.2: Screenshot of interactive graphical filter: (a) In Annotate-it, the graphical interactive filter allows the user to view the distributions of a number of parameters and interactively set filter settings. Filters can be adapted by dragging thresholds or clicking to select/deselect certain types of variants. Filtering on one parameter shows the impact on all other parameters in near real time. Plots included at the time of writing are (clockwise starting from upper left) coverage, variant allele frequency, presence in public databases, Phylop scores, the number of candidate genes hit in a given amount of samples, and Phastcons scores. Once satisfied the user can apply the filter settings and inspect the resulting gene lists. (b, c) Zoomed in portions of the interactive filter show a bar plot showing the proportions of variants of different consequence types (b) and the distribution of conservation scores for (red) primates, (blue) mammals, (green) vertebrates (c).

it's architecture is completely technology independent and therefore switching to different Rails-supported databases requires little effort. For large data sources containing billions of data points (such as base-wise conservation scores), we rely on Tokyo Cabinet [50] key-value stores for quick access. Interactive visualizations were implemented using Processing [33] and integrated into the web-interface as java applets.

Computationally intensive tasks (for example, complex predictions based on protein structure analysis) can be accomplished in a distributed manner by installing the Annotate-it client and the software one wishes to run on any computing node that has direct network access to the Annotate-it server. In this way, computing power can be scaled up on demand, either locally or by the use of cloud-based systems. Variant files can be uploaded and imported in the correct data structures at a pace of 500 variants per second using the current computing infrastructure. Previously unseen, and thus yet unannotated, variants are processed at a rate of approximately 85 variants per second. Computing infrastructure can be easily expanded to face increasing demand using elastic cloud technologies. This process is run completely at the server-side and requires no user interaction.

Input files of different formats are supported. So far, support includes Varscan [150], Atlas-SNP2 [151], Samtools variant calling format [43] , Genome Analysis Toolkit's VCF4 [47], Roche 454 GSMapper, and Illumina CASAVA output. Multiple files of different formats or sequencing technologies can be transparently combined for a single sample. Gene and variant lists can be exported to tab-separated value files based on the used filter-settings if needed.

Generation of Miller syndrome semi-synthetic data sets

We generated synthetic data sets of Miller syndrome by taking the following steps. We selected eight unrelated previously acquired exomes sequenced by Illumina GA2 and Illumina HiSEQ 2000. We aligned these exomes with BWA and called variants using the Genome Analysis Toolkit pipeline, both at default settings (Table 2). The sample identifiers of the variants were then randomly shuffled, maintaining the total number of variants, the number of variants per individual exome and the number of variants of each consequence type in the randomized data set (Figure 4.3). The eight synthetic exomes were then randomly assigned to either case or control groups so that each group contained four samples. Mutations found in the four published Miller syndrome cases were then added to each of the respective synthetic case samples. This synthetic case-control data set was then analysed with Annotate-it using the following filter settings: minimum 10× coverage, minimum variant allele frequency of 20%,

only 'novel' variants, variants not found in the 1000 Genomes Pilot project, and variants not occurring in the control samples. Only nonsynonymous, nonsense or essential splice site variants were considered. Ranking of the candidate genes was done based on the number of case samples in which the gene was considered a hit and the Phylop conservation score based on the multiple alignment of placental mammals. We calculated rank statistics based on 1,000 randomizations following the previous steps.

Table 4.2: Statistics of initial exomes used for randomizations: 8 previously acquired exomes were used to generate 8000 synthetic exomes by randomization. The initial exomes were sequenced by either Illumina GA2 or Illumina HiSeq 2000, then aligned with BWA and variants were called using the GATK pipeline.

Sample identifier	Platform	Essential Splice Site	Nonsense	Missense	Synonymous	UTR
1	GAII	178	99	8864	8409	3839
2	GAII	120	88	7468	7214	3216
3	GAII	137	95	8039	7756	3080
4	GAII	134	108	8246	8065	3264
5	HiSEQ	47	101	7925	8500	4264
6	HiSEQ	45	60	6715	7676	1443
7	HiSEQ	49	68	6487	7347	1421
8	HiSEQ	44	68	6560	7597	1367

Results

We validated the use of Annotate-it using three independent datasets: 1) a semi-synthetic Miller syndrome dataset, 2) a Schinzel-Giedion syndrome dataset and 3) a Nicolaides Baraitser syndrome dataset. In Miller syndrome as in Schinzel Giedion syndrome the causative gene was previously discovered and thus served to see if results could be reproduced using Annotate-it. Due to the synthetic nature of the Miller syndrome set we were able to statistically estimate the efficiency of commonly used filtering approaches in the discovery of Mendelian disease genes. In the Nicolaides Baraitser syndrome dataset the causative gene was previously unknown to us and illustrates how additional phenotype-specific information can aid in the discovery of causative genes.

Case study 1: semi-synthetic Miller syndrome data

To evaluate the impact of random neutral variation on finding the causative gene, we generated a synthetic data set modelled on published cases of Miller syndrome [85], a rare recessive disease caused by mutations in DHODH.

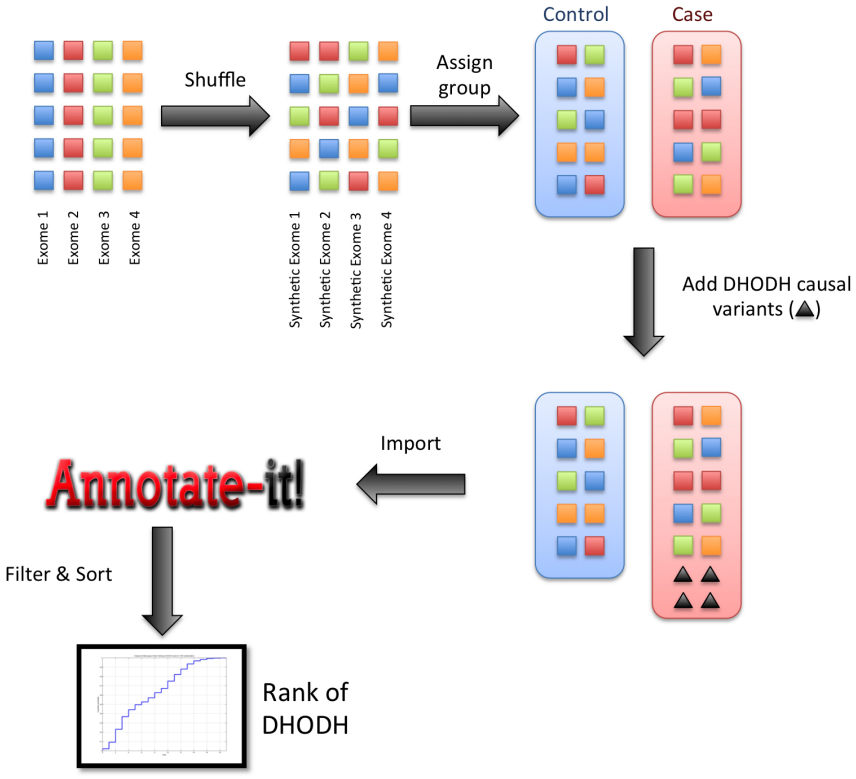


Figure 4.3: Randomization scheme used for the semi-synthetic Miller data set: We started from eight non-related previously acquired exome sequences and shuffled the sample identifiers linking a variant to a sample. By doing this the total number of variants and the number of variants per consequence was kept constant but the constitution of each of the exomes changed continuously. We then assigned four exomes to both control and case groups. In the case group we added two of the reported causal variants to each of the case samples. We then analysed each of the randomizations with Annotate-it and looked at the resulting rank of DHODH. By repeating this randomization cycle 1,000-fold we calculated rank statistics.

We considered two possible genetic models, which have been implemented in Annotate-it, when computing the rank of DHODH: the recessive model, meaning that a gene was only considered a hit in a sample if two different variants were found (assuming that a causal homozygous non-reference variant was highly unlikely), and the dominant model. Under the recessive model an average of 16 (± 2) candidate genes with hits in the 4 patients (41 (± 4), 105 (± 6),417 (± 12) candidate genes with hits in at least 3, 2, or 1 patient, respectively) were found. Out of these candidate genes, DHODH ranked in the top two in 95% of the randomizations (Figure 4.4). Under the dominant model an average of 113 (± 7) candidate genes with hits in the 4 patients (408 (± 12), 1132 (± 15),3417 (± 15) candidate genes with hits in at least 3, 2, or 1 patient, respectively) were found. Out of these candidate genes, DHODH ranked in the top seven in 95% of the randomizations. These data show that, despite random neutral variation, simple filtering and sorting criteria can dramatically decrease the number of genes to be validated in the case of Mendelian disorders.

Case study 2: Schinzel-Giedion syndrome

In a study by Hoischen et al. [84] exome sequencing was used to identify causative mutations in Schinzel-Giedion syndrome. Because of the rarity of the disorder, the absence of gender bias, its mostly sporadic occurrence and the absence of cytogenetic imbalances, the disorder was believed to be caused by variants in a single or perhaps a few genes resulting in an autosomal dominant inheritance pattern. Exome sequencing of four patients, and validation in an additional four patients by Sanger sequencing, has identified de novo mutations in SETBP1 to be causative for the disease. These mutations were clustered in a highly conserved region of the gene and are found in an 11-bp stretch. In the previous study, mutations were filtered based on dbSNP build 130 for novel variants and only nonsynonymous (mutations changing the amino acid composition of the protein), nonsense (mutations resulting in either a gain or loss of a stop codon) or essential splice site variants (mutations located within two bases of the intron/exon boundary) were considered. Previously generated in-house data were used to filter out variants found in samples not related to the phenotype. This was done under the assumption that because of the rarity of the disorder, it was unlikely that these control samples would contain the same causal variant.

We reanalysed the data using Annotate-it and applied default filters to validate the effectiveness of our methodology. We uploaded the variant files for annotation on the hg18/NCBI36 build of Annotate-it and filtered variants using two different settings: a cross-sample comparison at the gene level and a comparison at the exon level. The only major difference between both strategies is that instead of

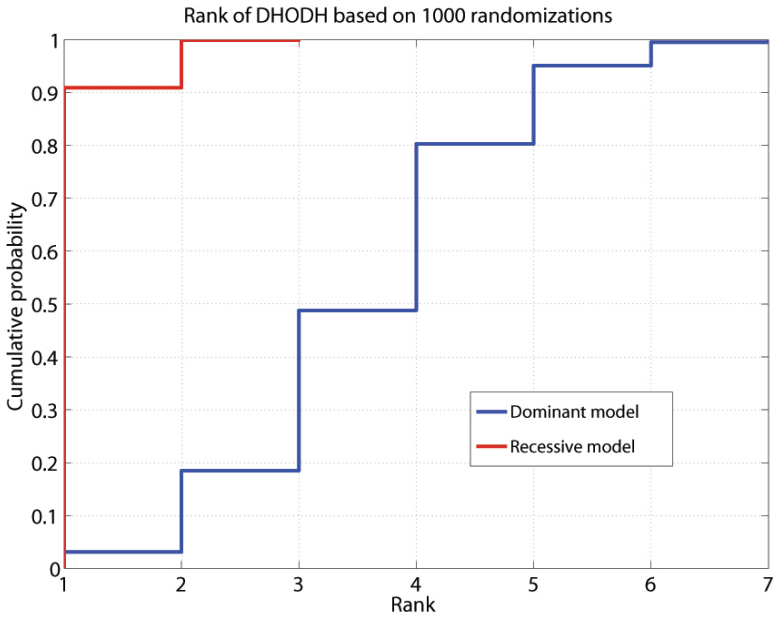


Figure 4.4: Rank statistics for DHODH based on 1,000 randomizations: The x-axis indicates the rank of DHODH, the y-axis indicates the cumulative probability of observing a rank lower than that marked on the x-axis.(red) Under the recessive model DHODH ranked in the top two candidate genes in 95% of the randomizations. (blue). Under the dominant model DHODH ranked in the top seven in 95% of the randomizations.

looking for candidate genes hit by variants across the patients, in the second approach we also look for specific exon hits across patients. For both strategies, we filtered out variants present in dbSNP (build 130) and the 1000 Genomes Project pilot studies (July 2010 release). We only selected variants with nonsense, nonsynonymous, and essential splice site consequences. We did not set any further filters, so as to recreate the scenario in which no further information was available to the researcher. Even though stricter filter settings would reduce the size of the resulting gene lists, we decided that manual inspection of all available information would allow us to select potential candidate genes without filtering a priori and potentially suffering false negative results that could lead to missing the implicated gene.

Both filtering approaches resulted in either a list of 11 candidate genes hit

across the 4 patients (or 30, 91, 793 candidate genes in, respectively, at least 3, 2, or 1 patient) or 13 candidate exons across the 4 patients (or 38, 81, 840 candidate exons in, respectively, at least 3, 2, or 1 patient). Sorting the gene list on number of samples implicated and maximum Phylop scores for placental mammals in the found variants, SETBP1 ranked at the top of the list as the prime candidate (Table 4.3). Further inspection of SETBP1 revealed the cluster of variants in a highly conserved region in the graphical view (Additional file 4), which might be indicative of gain-of-function mutations. Other genes that were hit across the four samples were also found but were deemed less likely because of either being hypervariable (for example, CDC27) or having variants in regions of lower conservation. In the exon-based approach one of the SETBP1 exons was ranked in the top ten exons hit across the four patients. Other high ranking exons were deemed unlikely candidates because of the high number of variants found and were likely hypervariable regions or pseudogenes. This approach is theoretically more sensitive to distinguish causative exons, as neutral variation is less likely to be clustered in individual exons across samples, even though it remains susceptible to the previously mentioned regions of increased variability.

Table 4.3: Top 10 candidate genes in Schinzel-Giedion case study: The data was analyzed using Annotate-it. Genes were considered a hit if they contained either nonsynonymous, nonsense or essential splice site variants which were not present in dbSNP or found in the 1000 Genomes Project. Genes were then sorted based on number of affected samples in which the gene was a candidate and conservation in placental mammals. Additionaly the total number of variants present across all samples and the number of variants unique to only one sample was given.

Rank	Gene	# Samples	# Unique variants	Total # variants	Haploinsufficiency	Max Phylop
1	*SETBP1*	4	2	3	0.721	3.455
2	*CDC27*	4	5	13	/	3.106
3	*CTBP2*	4	3	8	/	2.805
4	*PRB1*	4	2	5	/	1.458
5	*KIR2DL1*	4	1	2	/	1.05
6	*FLG*	4	5	6	/	1.034
7	*OR11H1*	4	0	1	/	0.856
8	*KIR2DL3*	4	4	6	/	-0.326
9	*CDCP2*	4	0	1	/	-0.365
10	*NBPF12*	4	0	2	/	-0.411

Case study 3: Nicolaides-Baraitser syndrome

Additionally we used Annotate-it to unravel the etiology of the previously unsolved Nicolaides-Baraitser syndrome (NBS) [152]. NBS is characterized by intellectual disability, sparse hair, distinctive facial characteristics, and distal-limb anomalies. The extremely rare sporadic recurrence of the disease and the lack of bias towards any gender suggested the cause of the disease to likely be autosomal dominant de novo mutations. We collected DNA from four unrelated cases and performed whole-exome sequencing. We ran the analysis in Annotate-it using default filters and parameters (nonsense, splice site, and nonsynonymous mutations not present in any of the available population databases). This resulted in a list of 296 candidate genes sorted based on the number of samples having mutations in the given gene and the number of mutations pertaining to a single sample for that gene. The top ranked gene, ACVR2A, was the only candidate gene with variants present in the four samples containing the same variant. Recurring mutations are more likely to be population-specific polymorphisms, and thus not captured by aspecific population databases, than de novo variation, which is believed to randomly occur throughout the genome and therefore has a very low probability of mutating the same position four times in non-consanguineous individuals. The second best ranking gene, SMARCA2, contained unique heterozygous nonsynonymous mutations in three of the four samples. Further investigation of associated gene ontologies and associated phenotypes computed by AGeneApart revealed this gene to be involved in chromatin remodelling and facial abnormalities, making it an interesting candidate for validation. Sanger sequencing of the samples and their respective parents revealed the three variants to be de novo events. Additional resequencing of SMARCA2 in 44 individuals revealed heterozygous mutations in the same carboxy-terminal helicase domain in 36 patients, hereby validating SMARCA2 as the causative gene for Nicolaides-Baraitser syndrome. Later published studies discovered the etiology of the phenotypically related Coffin-Siris syndrome to be caused by similar mutations in SMARCB1 and ARID1B, which are direct interaction partners of SMARCA2 in the SWI/SNF chromatin remodelling complex [153, 154]. Based on these findings we added functionality to Annotate-it to find causative mutations in protein complexes rather than in a strictly monogenic setting. By doing this we discovered that one of eight patients diagnosed with NBS for which we performed whole-exome sequencing carried a mutation in a gene identified in Coffin-Siris syndrome. Although the phenotype of NBS is strongly overlapping with that of Coffin-Siris syndrome, the latter differs by the presence of hypoplastic or absent fifth finger nails with or without hypoplasia of the terminal phalanges, the absence of swelling of the finger joints, short metacarpals and broad terminal phalanges and internal organ malformations are more common. Yet due to the large overlap of features,

both syndromes are hard to classify [97], especially at a young age, which could explain the misclassification of this particular sample.

Discussion

Context

A multitude of publicly available computational tools exist in order to assist bioinformaticians and geneticists by facilitating the analysis of single nucleotide variants at the level of annotation, interpretation or both. Yet most tools focus only on particular aspects of the analysis pipeline (Table 4.1) and are thus only usable under certain experimental circumstances. Furthermore, all described tools suffer from several drawbacks, such as 'red queen' annotation problems, cross-sample analysis, phenotype-naive interpretation and general data management problems, such as data compatibility and data sharing [155].

Annotate-it

In this paper, we describe Annotate-it, a versatile framework for the analysis of multisample single nucleotide substitution data generated by next-generation sequencing. Annotation of samples is performed on the server side, eliminating the need for the installation of complex tools and annotation sources by the end-user and automatically keeping those annotations up to date.

The query and filtering interface enables the geneticist to quickly test different genetic hypotheses (recessive, dominant, de novo) across multiple samples and aggregates available information at the gene and variant level, facilitating the manual revision of candidate gene lists. We also provide a suite of more complex analysis techniques, such as aggregate functionality scores, phenotype-specific gene prioritization, and statistical methods for disease-gene finding in case/control studies without any need for data reformatting or computational expertise. By merging different state-of-the-art analysis techniques we aim to deliver a Swiss-knife type tool that can be used in many different experimental contexts. By unifying the interface and centralizing the data the user is able to perform multifaceted analyses without any need for additional data management or formatting and without requiring any computational expertise.

Case studies

To validate our approach we simulated an experiment with randomized exomes and published variants found in Miller syndrome patients. We computed ranking statistics showing that the causative gene ranks highly, even with a large amount of random neutral variation. This happens even when considering a relatively small set of patient exomes. In a second benchmark, we reanalysed a published case study of the Schinzel-Giedion syndrome and identified the causative gene, SETBP1, as the top-ranked gene, using default parameters. Further inspection reinforced the previous finding that novel nonsynonymous variants across patients are clustered in a small 11-bp stretch. We applied the same approach in a previously unpublished study of NBS and found SMARCA2 as one of the top-ranking genes. This gene was then further validated to be the causative gene in this disease. Furthermore, we show in this case that gene prioritization methods can aid in the prioritization of mutated candidate genes by linking these genes to their associated phenotypes. Additionally, we show that in phenotypically overlapping disorders, leveraging protein-protein interaction data could prove useful in deciphering neighbourhoods of mutated genes.

Future Perspectives

Although the framework as presented here is fully able to support the discovery of causal genes of rare genetic disorders, our approach would benefit further from several critical features that we will investigate in future work. First, the confidentiality of human clinical data is a concern of many clinical geneticists and is still a topic of debate in most research centers. We aim to develop future releases of Annotate-it using a client-server architecture, so that sample and genotype information is dealt with on a locally installed client that automatically synchronizes genomic (but not clinical) data (that is, single nucleotide variant chromosome, position, and variant allele) with a server that holds the most up-to-date annotation information. A public client will be available for sharing data across research centers. Secondly, we designed Annotate-it with protein coding single-nucleotide substitutions in mind as proof-of-concept. We aim to further expand the scope of the methodology to encompass annotation of non-coding regions and other types of variation, such as insertions and deletions of varying lengths.

Conclusions

In this paper we describe the application of Annotate-it to study rare monogenic Mendelian disorders caused by rare, highly penetrant variants in the population. Because sequencing (as opposed to classical array-based genome-wide association studies) detects rare (less than 0.01% of the general population), infrequent, and frequent variants (over 1%), extensions of our strategy will be useful in resolving oligogenic diseases caused by a combination of infrequent variants of intermediate penetrance. Discovering the disease genes in such challenging cases will require good models of (1) the underlying population genetics (to avoid confounding effects from population structure), (2) the functional impact of variants on protein function and on regulation (to weed out passenger mutations), and (3) the biological pathways involved (to detect causative variants through their effect on pathways rather than individual genes). This is a primary research goal for the near future.

Current approaches to identifying causal variants by patient sequencing, including ours, can reliably interpret only the most straightforward subset of the variome (nonsense, missense, and splice-site variants). For example, synonymous variants, variants in regulatory, intronic, and intergenic regions, and variants in microRNAs remain highly challenging. Future experimental and in silico studies assessing the impact of such variants will allow us to develop computational strategies to bridge the gap in data interpretation between the exome and the complete genome. In particular, refinements in the mapping of regulatory regions and other functional elements through functional studies, evolutionary conservation, and in silico models will be essential to enable variant interpretation beyond the coding exome. Furthermore, the integration of large-scale structural variation will be crucial in a comprehensive study of the variome. Such integration will be challenging because structural variation inherently requires different types of data structure, visualization, and analysis methods. A high-level meta-analysis of variation at multiple scales should provide researchers with a complete picture of individual genomes. Making such complex, integrative analyses effective is a major challenge for the computational biology community.

Given the positive reactions from initial users, we believe that Annotate-it can play a significant role in the prioritization and identification of candidate genes in Mendelian disease. In the future, we aim at broadening its scope by increasing the annotation information available, the effectiveness and scope of our filtering analysis methods, and the robustness and ease of use of our web interface.

Availability

Annotate-it is freely available at `http://www.annotate-it.org`. Bug reports can be made and feedback accessed at `http://annotateit.userecho.com`.

Requirements

Annotate-it has been tested on the following browsers: Google Chrome (build 11.0.696.71 or later), Mozilla Firefox (version 3.6 or later), Safari (version 5.0 or later). Microsoft Internet Explorer is currently not supported.

Acknowledgements

The authors would like to thank Christian Gilissen en Joris Veltman for supplying the Schinzel-Giedieon data. Research supported by: Research Council KUL: ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys en KUL PFV/10/016 SymBioSys, START 1, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects, G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); G.0733.09 (3UTR); G.082409 (EGFR), IWT Grant No. IWT-SB/093289: PhD Grants, Silicos; SBO-BioFrame, SBO-MoKa, TBM-IOTA3, FOD:Cancer plans, IBBT, Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011); EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED. BN is supported by KOLUMB fellowship from the Foundation for Polish Science.

Supplementary Information

Supplementary figures and information can be found at
`http://genomemedicine.com/content/4/9/73/additional` .

Chapter 5

Phenotype-specific variant prioritization

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.
>
> Nick Bostrom

5.1 Summary

Although most of the analysis methods presented in previous chapters have been shown to be useful in identifying causative mutations in rare Mendelian disorders, sometimes these methods cannot be applied due to a lack of familial samples or due to limited sample sizes, a problem which frequently occurs when dealing with extremely rare disorders. In these cases, deleteriousness prediction methods had been proposed but suffer from high false-positive rates due to their inability to distinguish rare non-disease associated variation from the true causative alleles. In this chapter we propose the first published method which integrates phenotypical information, through gene prioritization, in order to predict causative variants in exome-sequencing studies of rare disease. We show how our method for phenotype-specific variant prioritization, called *eXtasy*, substantially outperforms the classical phenotype-naive methods. We offer the methodology in the form of an easy-to-use webtool and an open-sourced, stand-alone command-line tool suitable for standard UNIX-based systems.

5.2 Phenotype-specific variant prioritization

The PhD candidate jointly conceptually designed the methodology with joint first author Dusan Popovic. He generated both testing and training data sets and performed manual curation and ontological mapping of these data sets. The PhD candidate ported the prototype Random Forest model (developed by Dusan Popovic) from MATLAB to the R Statistical Framework and implemented the stand-alone tool. Jointly with co-author Amin Ardeshirdavani, he implemented the web-tool version of the tool. He is the current maintainer of the code-base and its main developer.

This paper was published in *Nature Methods*, Sept. 2013, 1083–1084, doi:10.1038/nmeth.2656

eXtasy: variant prioritization by genomic data fusion

Alejandro Sifrim[1,2,4], Dusan Popovic[1,2,4], Leon-Charles Tranchevent[1,2], Amin Ardeshirdavani[1,2], Ryo Sakai[1,2], Peter Konings[1,2], Joris R. Vermeesch[3], Jan Aerts[1,2], Bart De Moor[1,2], Yves Moreau[1,2]

[1]Department of Electrical Engineering, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium
[2]iMinds Future Health Department, Leuven, Belgium
[3]Laboratory of Molecular Cytogenetics and Genome Research, KU Leuven, Leuven, Belgium
[4]These authors contributed equally to this work
Corresponding author: Yves Moreau, yves.moreau@esat.kuleuven.be

Abstract

Massive parallel sequencing greatly facilitates the discovery of novel disease genes causing Mendelian and oligogenic disorders. However, many mutations are present in any individual genome, and identifying which ones are disease causing remains a largely open problem. We introduce a new approach to prioritize nonsynonymous single nucleotide variants (nSNVs) that substantially improves prediction of disease-causing variants in exome sequencing data by integrating variant impact prediction, haploinsufficiency prediction and phenotype-specific gene prioritization.

Main

Rare exomic variation identified by exome sequencing is particularly useful to discover the cause of many previously unresolved rare monogenic disorders. By filtering down the exome against nonsynonymous single nucleotide variants (nSNVs) and loss-of-function mutations that are not present in healthy populations or unaffected samples, we can discard a large proportion of the exomic variation as probably neutral. However, despite such aggressive filtering, several thousand candidate causal mutations remain and we need predictive

methods to prioritize variants for further validation. Several computational methods have been proposed that take into account biochemical, evolutionary and structural properties of mutations to assess their potential deleteriousness [71, 124, 146, 156, 145]. However, most of these methods suffer from high false positive rates when predicting the impact of rare nSNVs [156]. A plausible explanation for this poor performance is that many of the scrutinized variants are mildly deleterious and subject to weak purifying selection [157, 18], but not specific to the disease of interest. To assess this hypothesis and to further enhance variant prioritization, we propose a genomic data fusion methodology [139] that integrates multiple strategies to detect deleteriousness of mutations and prioritizes them in a phenotype-specific manner. A key innovation that we incorporate into our strategy is a computational method for gene prioritization [94], which scores mutated genes based on their similarity to known disease genes by fusing heterogeneous genomic information. We also integrate haploinsufficiency prediction scores [65] that predict the probability that the function of a gene is affected if present in a functionally haploid state. To integrate or fuse these data sources, we use random forest learning [158] and train our model on the Human Gene Mutation Database (HGMD) of human disease-causing mutations [159] compared to three control sets: common polymorphisms and two independent sets of rare variation (Figure 5.1).

After generating and annotating the different data sets, we inspected the distribution of the different deleteriousness prediction scores across the positive and control sets (Supplementary Fig. 1). All deleteriousness prediction scores seemed to score the positive set high and thus showed a high sensitivity. When looking at the control sets, we observed that most methods correctly classified common polymorphisms as benign - yet classified a substantial proportion of rare variation as deleterious, leading to a low precision. Control variants seemed to occur more often in genes predicted to maintain functionality in a haploid state (being haplosufficient), whereas disease-causing variants showed no clear pattern. Disease-causing variants were primarily found in top-ranked genes after gene prioritization. By contrast, control variants showed a homogeneous distribution of gene prioritization ranks, which is to be expected under the assumption that they are prioritized for randomly selected phenotypes.

By integrating these different scores, we aimed to boost our ability to discriminate between putatively mildly deleterious rare variants and actual disease causing variation. We evaluated several commonly used classification approaches and chose a Random Forest (RF) classification algorithm because it outperformed all other classification algorithms on this task (Supplementary Table 1). This classifier is trained by comparing our positive set of disease-causing variants to the rare variant control sets. When compared to classical deleteriousness prediction scores, we observe a considerable improvement in all

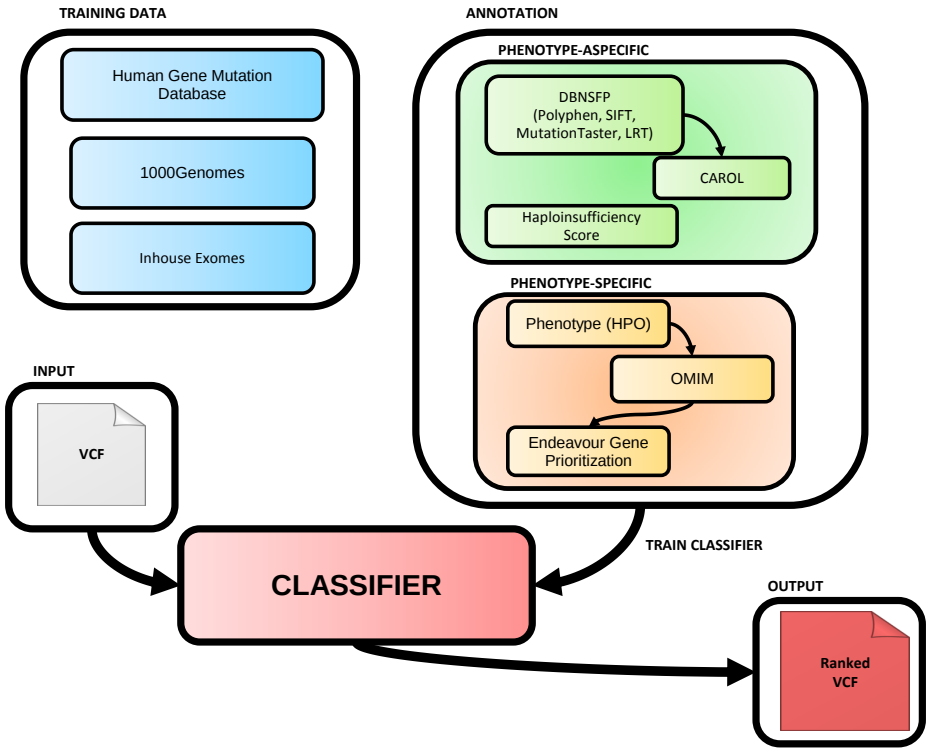


Figure 5.1: This diagram gives an overview of the presented method. First we annotate disease-causing variants and control, presumed neutral variants, with phenotype-specific and phenotype-aspecific annotations. These annotations such as deleteriousness prediction scores or phenotype-specific gene prioritization scores are used as features for training a machine learning classifier (e.g. a Random Forest classifier). This classifier can then be used to score a user-provided variant file in the standard Variant Calling Format (VCF) which can then be ranked based on the eXtasy score according to the computed likelihood of variants being causative for the observed phenotype.

performance measures. This is the case when distinguishing between disease-causing and rare control variants (Figure 5.2a, Supplementary Fig. 2a,b and Supplementary Table 2), as well as with distinguishing disease causing variants from common polymorphisms (Figure 5.2b, Supplementary Figure 2c,d and Supplementary Table 2). The performance against common polymorphisms is in line with published results for deleteriousness prediction tools because these tools were trained using common polymorphisms as controls. The performance of these tools against rare non-disease-causing variants is much lower than against common polymorphisms. Precision is the most improved performance measure, which is important when dealing with a prioritization task (Supplementary Table 1, Supplementary Note). However, performance measures obtained in retrospective benchmarks such as ours are usually optimistic estimates, because of the bias of prior information for gene prioritization predictions [139, 160]. This problem is difficult to address in an initial study and can only be truly resolved by long-term prospective benchmarks where predictions are made on the current state of knowledge and validated in future studies. To estimate to which degree such bias was present in our benchmark, we assessed classification performance based on the date of discovery of the causative variation using data for gene prioritization anterior to the year of discovery (Supplementary Fig. 3). If a bias is present, it would be less prominent in recent discoveries as these gene-disease associations would be less likely to be directly or indirectly incorporated in the gene prioritization data sources. Even though we see a slight decline in performance (in terms of Area Under the Curve) for more recent discoveries, the method still performs considerably better than classical deleteriousness prediction scores (Figure 5.2c, Supplementary Table 3). The performance of deleteriousness prediction scores is not affected by the year of discovery of the association, which is expected because these scores do not integrate time-dependent information. Finally, we looked at feature importance by shuffling each feature across disease-causing and control variants evaluating the increase in classification error (Supplementary Fig. 4). This analysis showed that all included features were informative and improved classification to some degree.

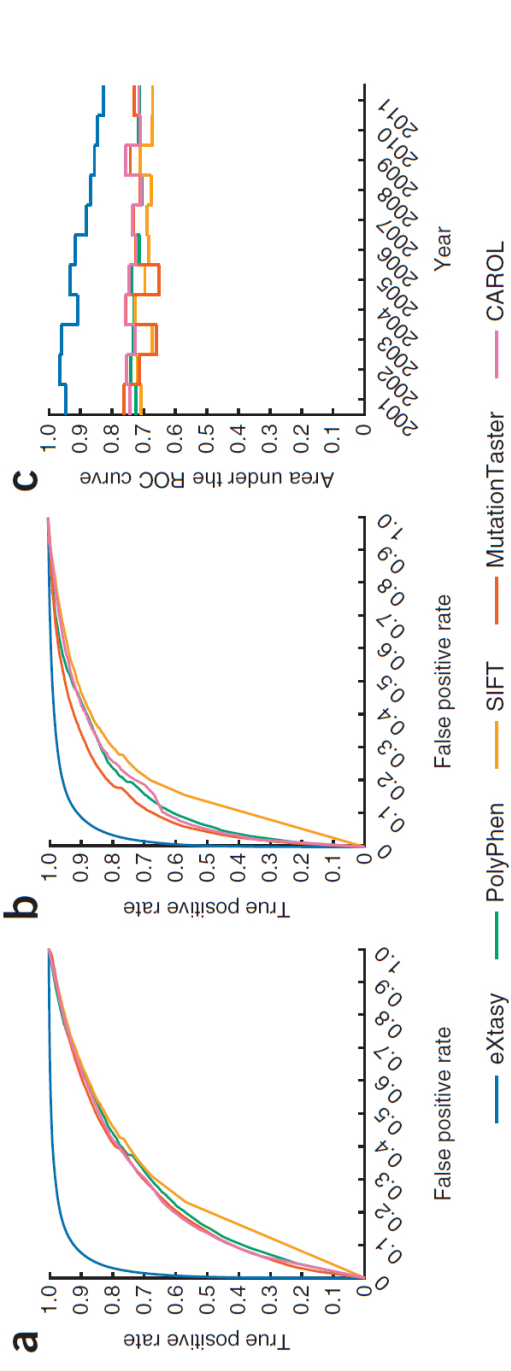


Figure 5.2: Receiver operating characteristic (ROC) curves comparing eXtasy and classical deleteriousness prediction scores: **(a–c)** ROC curves for disease-causing versus rare control variants (a) and disease-causing versus common polymorphisms **(b)**. In both cases, eXtasy outperforms other methods, as can be seen by an increase in the area under the curve (AUC). **(c)** We stratified disease-causing variants by year of discovery to test biases in our retrospective benchmark. More recent disease-variant associations show a decrease in AUC performance for eXtasy as biases decrease, but eXtasy always outperforms classical deleteriousness prediction scores.

In this study, we show a novel approach towards integrating biochemical, evolutionary and phenotype-specific information to prioritize mutations for follow-up validation studies. This approach helps to distinguish disease-causing mutations from neutral common polymorphisms as well as rare, potentially deleterious but phenotypically unrelated variation in the coding genome. We acknowledge that performance measures are likely overestimated because of the biased nature of retrospective benchmarks, yet we can clearly appreciate a marked improvement in prediction performance over frequently used deleteriousness prediction scores in recently published mutations. We envision that in the near future current initiatives such as the Critical Assessment of Genome Interpretation (CAGI), although currently focused on single phenotype benchmarks, will play a major role in providing unbiased prospective benchmarks to optimally assess the performance of methods such as the one described in this study. Future research and the availability of larger public disease-causing variation data sets will likely widen the scope of the method to other types of mutations (mitochondrial, noncoding, synonymous, nonsense, splice-site mutations, indels). Also, the addition of other data sources, such as locus-specific information (e.g., copy number variant prevalence, GWAS-associated loci), to our prioritization method and its integration into genetic association test across multiple samples [135, 128] will likely increase its power to discover the cause of genetic disease. We freely provide public access to the described methodology through a web tool and also provide an offline standalone version (`http://homes.esat.kuleuven.be/~bioiuser/eXtasy/`, source code: `http://github.com/asifrim/eXtasy`).

Methods

Data generation

To determine the effectiveness of our approach, we assembled and evaluated a data set consisting of disease-causing variants and different types of control variants. We used a positive data set consisting of 24,454 disease-causing nSNVs from the HGMD associated with 1,142 different Human Phenotype Ontology (HPO) terms [60]. Mapping between HGMD disease descriptors and HPO terms was performed using the Phenomizer tool [161]. Phenotypic terms for which fewer than three implicated genes were known were discarded (excluding the gene in which the variant was located) so as to allow meaningful subsequent gene prioritization. For the control data sets, we considered two classes of variants on the basis of their minor allele frequency (MAF) in the 1000 Genomes Project data set: common polymorphisms (MAF $> 1\%$) and rare variation (MAF $\leq 1\%$). Additionally, we compiled a third control set

of rare variants using in-house–sequenced exomes of healthy individuals and retaining only high-quality calls (coverage $> 30\times$) not present in any publicly available variant database (NHLBI EVS, dbSNP or 1000G). After randomly assigning groups of 500 control variants from each control set to each HPO term represented in the disease-causing variants, we annotated all variant-phenotype combinations with functional information. For every variant, we extracted PolyPhen-2, sorting intolerant from tolerant (SIFT), MutationTaster and likelihood-ratio test (LRT) deleteriousness prediction scores from the dbNSFP database [117]. We calculated combined annotation scoring tool (CAROL) aggregate deleteriousness scores [127] and added phyloP [162] and PhastCons [163] evolutionary conservation scores across vertebrates, primates and placental mammal subsets from the UCSC Genome Browser database. We also included precalculated gene haploinsufficiency prediction scores where available. Finally, we computed gene prioritization scores with Endeavor [94] using gene-phenotype associations obtained from HPO. (For additional information on the data generation and annotation process see Supplementary Note 1).

Classifier benchmarks

We set up a benchmark to determine the optimal classifier for genomic data fusion. To do this, we followed published guidelines [164] applicable to these types of studies. Initially, we removed all records containing missing values and split the disease-causing variants into training (two-thirds of the total number of genes) and testing sets (one-third of the genes) (Supplementary Fig. 5). The data records were stratified at the gene level (the highest level of granularity) to assure that the algorithm is not overfitting the gene level information and thus does not overestimate performance. Subsequently, we randomly subsampled the negative examples in the training set so as to balance their number with the number of positive examples. We then trained eight different classifiers using the same combinations of training and test subsets for all classifiers. In addition, we applied state-of-the-art deleteriousness prediction methods (PolyPhen, SIFT, MutationTaster and CAROL scores) using their respective published or documented decision thresholds (0.85, 0.95, 0.5 and 0.98, respectively). As there were several records per single variant in the data set (because of the different phenotypes per variant), we used the maximum score across phenotypes in testing to predict the outcome for each variant. We chose the maximum as this is robust to noninformative phenotypes (because of phenotypically variable diseases), which might be present in our benchmark data set of disease-causing variants. This did not apply for methods that produce a single score per variant because they are not phenotype specific (PolyPhen, SIFT, MutationTaster and CAROL).

The procedure was repeated 100 times on different random splits of data to obtain an estimate of the variance of the resulting performance measures. The results were computed in terms of accuracy, sensitivity, specificity, positive and negative predictive value, Matthews correlation coefficient, area under the receiver operating characteristic (ROC) curve and precision-recall (PR) curves. We observed that all the classifiers built using all heterogeneous data sources performed better than state-of-the-art deleteriousness prediction methods, proving that these data sources contain additional information that facilitates better predictions.

Among all classification algorithms, Random Forests (RF) outperformed all others in terms of all performance metrics, with the exception of sensitivity and negative predictive value (NPV), for which linear discriminant analysis (LDA) performed marginally better (Supplementary Table 1 and Supplementary Figs. 6 and 7). However, precision (positive predictive value or PPV) of the LDA was very low (0.35), indicating that the classifier is overly optimistic. The same observation holds for all of the state-of-the-art methods (precision between 0.20 and 0.23). Furthermore, the PR curve for the Random Forests shows that changing the default threshold (0.5) of the classifier results in a sharp increase in precision but a small loss in sensitivity. This suggests that most of the true positives are highly ranked by the method. Also, the reported accuracy and other aggregate performance measures of the state-of-the-art tools depend greatly on where the decision threshold is set and the skewness of the class distribution. Even though the AUC for some of the ROC curves is higher on one tool than on another, sometimes a point can be found where a certain measure is higher for the tool with the lower AUC.

For all classification algorithms, we use their respective Matlab 7.10 implementations: classregtree class for the decision trees, TreeBagger class for the Random Forests, function knnclassify for k nearest neighbors, function classify (with argument “linear” or “quadratic”) for the LDA and QDA, respectively, class NaiveBayes for naive Bayes classification, and functions svmtrain and svmclassify for the feed-forward neural networks. Most of the used functions and classes are part of the Statistical Toolbox, except for KNN and SVMs (Bioinformatics Toolbox). The details on parameter settings of particular classifiers are provided in Supplementary Table 5.

Control-set benchmarks

We set up an additional benchmark to assess the behavior of the final model under different classification schemes with regard to the different negative outcomes. We considered two different scenarios for training and testing. In a

first scenario, the Random Forests model is trained using a subset of the rare non–disease-causing variants as negatives and is tested against the rest of them (as in our standard setup). In a second scenario, the Random Forests model is trained using all of the rare non–disease-causing variants available as negatives and is tested against the data set containing common polymorphisms.

In the case of the first scenario, the validation scheme is identical to the classifier benchmark: all data, including all positive and the negative examples, have been repeatedly (100 iterations) divided at random into training and testing subsets, grouped gene by gene. In the case of the second scenario, only the positive examples were randomly assigned, whereas the two distinct groups of negatives, rare and polymorphisms, were used respectively for training and testing (Supplementary Fig. 8). We made sure that during a single iteration, a split of positives stayed the same across the two scenarios.

Temporal stratification analysis

To analyze the sensitivity of the method with regard to a priori gene or disease association biases, we set up an additional benchmark to estimate the effect size of such biases. Under the assumption that recently published genes would be less likely to be biased by our gene prioritization step, we stratified our positive testing set of disease-causing variants by year of publication (2000–2012) while training the model only on data published before 2000. In this way we measure performance before and after Endeavor's data sources were last updated (the last update occurred in 2008). This threshold applies to both variant and gene level of data granularity, so variants discovered after 2000 but that are associated with genes that are part of the training (i.e., for which the gene-phenotype association was discovered before 2000) are also removed from test sets. The negative examples (non–disease-causing variants) were randomly assigned to one given year between 2000 and 2012, in numbers matching the class distribution of the whole data (given the number of positives in a particular year) and with no overlap between training and test sets (see Supplementary Fig. 9). As before, the splits were performed genewise. After the training phase, the classifier was used on cases from the subsequent years (2000–2011). The whole procedure of randomly assigning negatives was repeated 100 times to get stable estimates of performance metrics.

We observed a slight temporal decline in performance throughout all testing years. Nevertheless, eXtasy still significantly outperformed all classical deleteriousness prediction methods across all years. We attribute this decline in performance to the fact that over time some disease-causing genes are better described in the gene-prioritization sources (e.g., literature mining) and are therefore easier

to classify. Although such effects are likely present and point to the fact that our main benchmark is an overoptimistic estimate of the real performance, this benchmark setting is a pessimistic estimate of the real performance because of various properties of the training/validation scheme. First, only a fraction of the positive training data can be used (data before 2000), leading to suboptimal learning of the features of disease-causing variants. Second, the training data themselves contain well-described genes and are thus themselves biased toward easier-to-classify examples. This leads to overly optimistic decision thresholds in the classifier and thus degrades the performance when faced with more difficult examples of less well–described genes in the testing set. Finally, owing to the genewise stratification of training and testing sets (to avoid overfitting specific genes), if gene data were published before 2000 but then later published in the light of a new phenotype, they were omitted from the test set. Well-described genes are often discovered to play important roles in new (and often related) physiological processes. This type of discovery can greatly benefit from gene-prioritization approaches but are excluded from this benchmark.

The real performance of the classifier depends greatly on the use case, which is unknown to the researcher applying the method. In the case where the cause of the phenotype is a novel mutation in a previously described gene, the performance is likely to resemble, or even exceed, that of the cross-validation benchmark. When the phenotype is caused by a novel mutation in a gene not previously associated with the phenotype, or in a gene associated with a different phenotype, the real performance likely lies between the overoptimistic cross-validation benchmark and the pessimistic temporal-analysis benchmark (the shaded light blue area in Supplementary Fig. 3).

Feature importance analysis

The Random Forests algorithm has the intrinsic ability to estimate the importance of features used for training [158]. This is achieved by measuring the difference between the mean-square error of the prediction on out-of-bag (OOB) examples when values of a given feature are shuffled compared to the error on undisturbed OOB data. This procedure is repeated for each and every tree in the ensemble with its corresponding OOB examples, providing a global measure of feature importance. Here we analyzed how the different features in the data contributed to the overall classification. In particular, we ran 100 simulations (in line with previously described benchmarks) per feature, during which ensembles are built using different random subsamples of the negative data. The result in terms of mean-square error increase when shuffling the feature is displayed in Supplementary Figure 4. From the plot, it appears that all features contribute to the classification to some degree. The increase of total

mean-square error when one of them is randomized ranges from around 2% for PPI-HPRD to 12% for sequence similarity and functional annotations. Second, highly correlated features, such as various Endeavor scores or state-of-the-art methods (CAROL with SIFT or PolyPhen) usually form clusters of seemingly less important features with a low yet nonzero increase in mean-square error. This is expected, as it has been shown that feature importance measures for Random Forests are strongly affected by the presence of correlation between features [165]. In the absence of a particular feature, other correlated features partially 'take over' the role of former, reducing the impact of the shuffling on the classification error. Hence, these variables are still individually very important—especially if data records contain missing values.

Supplementary Material

Supplementary Note 1 - Data generation

1.1 Data sets of variants

Disease-causing variants

We acquired 24,454 nonsynonymous variants from the Human Gene Mutation Database (HGMD)1 Professional version 2012.1. We manually mapped HGMD disease descriptors to OMIM disease terms. Using the Phenomizer tool3, we mapped the OMIM disease descriptors to Human Phenotype Ontology (HPO) terms. Diseases that were mapped to multiple HPO terms were split into one record per HPO term. For example, "Miller syndrome" would be mapped to OMIM:263750 by searching OMIM, the Phenomizer tool would then translate this into individual phenotypes such as: cleft palate (HP:0000175), micrognathia (HP:0000347), hypoplasia of the radius (HP:0002984) and other HPO terms. A variant associated with Miller syndrome would then be associated to multiple records associated with the individual HPO terms. Records with HPO terms with less than 3 known associated genes (not including the gene where the variant was located in) were discarded. Cancer-related HPO term containing records were also removed from the data set. In total, we retained records for 1,142 distinct HPO terms.

Common polymorphisms

For every HPO term present in the disease-causing variant data set, we randomly selected 500 variants genomewide with a global minor allele frequency (MAF) greater than 1% in the 1000 Genomes Project (n=43,724).

Rare genomic variation

Analogous to the common polymorphism set, we selected 500 variants per HPO term but with a 1000 Genomes Project MAF lesser than 1% (n = 257,556). Using the same procedure, we generated a second rare variation data set based on an in-house data set of 68 in-house sequenced healthy control exomes of which the variants were unique to one individual, not present in any of the public variation databases (dbSNP, 1000Genomes, NHLBI Exome Variant Server) and sequenced at a coverage higher than 20-fold (n=25,429).

1.2. Variant features

Variant level features

To every record, we appended Polyphen25, SIFT6, MutationTaster7, and LRT8 precomputed deleteriousness prediction scores extracted from dbNSFP9

(v 1.3). Because of differences in the underlying tools, dbNSFP handles alternative splicing differently for each tool. SIFT and LRT do not use transcript information and are thus not insensitive to alternative splicing. For Polyphen2, the canonical transcript was chosen. In the case of MutationTaster, ANNOVAR was used to annotate the variant with transcript information from Ensembl. The first transcript ID was then submitted to MutationTaster. If the first transcript could not be processed by MutationTaster, the second ID was used. This process was repeated until MutationTaster was successfully run. Additionally, we computed an aggregated deleteriousness prediction score using CAROL10. We further added PhyloP11 and PhastCons12 46-way conservation scores for vertebrate, placental mammals, and primate groups acquired from the UCSC genome browser.

Gene level features

We added publicly available gene-wise haploinsufficiency prediction scores where available. Finally, we computed genome-wide prioritization scores using Endeavour for each of the phenotypes using gene-phenotype associations obtained from the HPO. If the gene in which the variant was located was present amongst these known associations, this gene was excluded from the Endeavour training genes. This mimics the case where the gene itself is not known to be involved. The Endeavour data sources included sequence and protein domain similarity, molecular function, expression data, protein-protein interactions, text mining, and pathway information. We appended a global prioritization score using order statistics, as well as a score per data source to each variant according to the phenotype-specific scores for each gene. A short description of the individual Endeavour data sources can be found below:

1. **GO**: This data source is based on associations between genes and functional attributes that describe molecular processes, biological functions and cellular compartments. Data comes from Gene Ontology.

2. **SWISSPROT**: This data source is extracted from the UniProt - TrEMBL/SwissProt database. In particular, the keywords that are associated to the genes by UniProt are collected and organized in a pseudo-ontology. These keywords describe in general the main function of the genes and their products.

3. **INTERPRO**: This data source is based on associations between genes and domains identified from the sequences of the associated gene products. These domains describe the functional elements of the proteins and are coming from the InterPro database.

4. **KEGG**: This data source is based on biomolecular pathways from the KEGG database. These pathways describe common molecular processes and list the genes playing a role in these pathways.

5. **STRING**: This data source comes from the STRING database and is organized as a gene network with edges connecting the genes that are functionally related (or predicted to be). STRING integrates several types of information including Protein-Protein Interaction (PPI) networks, gene expression data sets and text mining of the scientific literature.

6. **BIOGRID**: This data source is a PPI network and is based on the integration of several large-scale experimental data sets. It contains both direct physical interactions as well as indirect interactions, such as genetic interactions. This data comes from the BioGRID database.

7. **INTNETDB**: The IntNetDB database integrates various gene networks including PPI and functional networks into one global gene network that contains both known and predicted gene interactions.

8. **HPRD**: This data source is a manually verified human PPI network and is based on the integration of several large-scale experimental datasets. It contains both direct physical interactions as well as indirect interactions such as colocalization. This data comes from the BioGRID database.

9. **SU ET AL.**: This data set comes from a large scale analysis of the human transcriptome (Su et al., PNAS 2002). It was obtained by profiling gene expression from 91 human tissues, organs, or cell lines.

10. **OUZOUNIS**: This data set represents a priori disease probabilities (Lopez-Bigas et al., NAR 2004). It is based on the use of sequence features (e.g., sequence length, UTR length, number of introns, and intron length) and a statistical framework to discriminate the human disease causing genes from the rest of the genome.

11. **BLAST**: This data source is based on the sequence similarities between all human proteins. It is obtained by using NCBI Blast on the protein sequences from Ensembl.

12. **TEXT**: This data source associates genes with keywords found in publications associated with these genes (from GeneRIFs). This text-mining data was obtained using a framework derived from the TxtGate tool starting from MEDLINE abstracts.

Supplementary Note 2 - Performance measures and their interpretation

We use eight performance measures to compare classifiers based on genomic data fusion against state-of-the art variant prioritization scores. These are standard metrics that are routinely reported in classification benchmarking, and we provide them all to facilitate complete evaluation of the performance. However, we strongly believe that in context of variant prioritization some of them are more interesting than others, simply because of the nature of the problem.

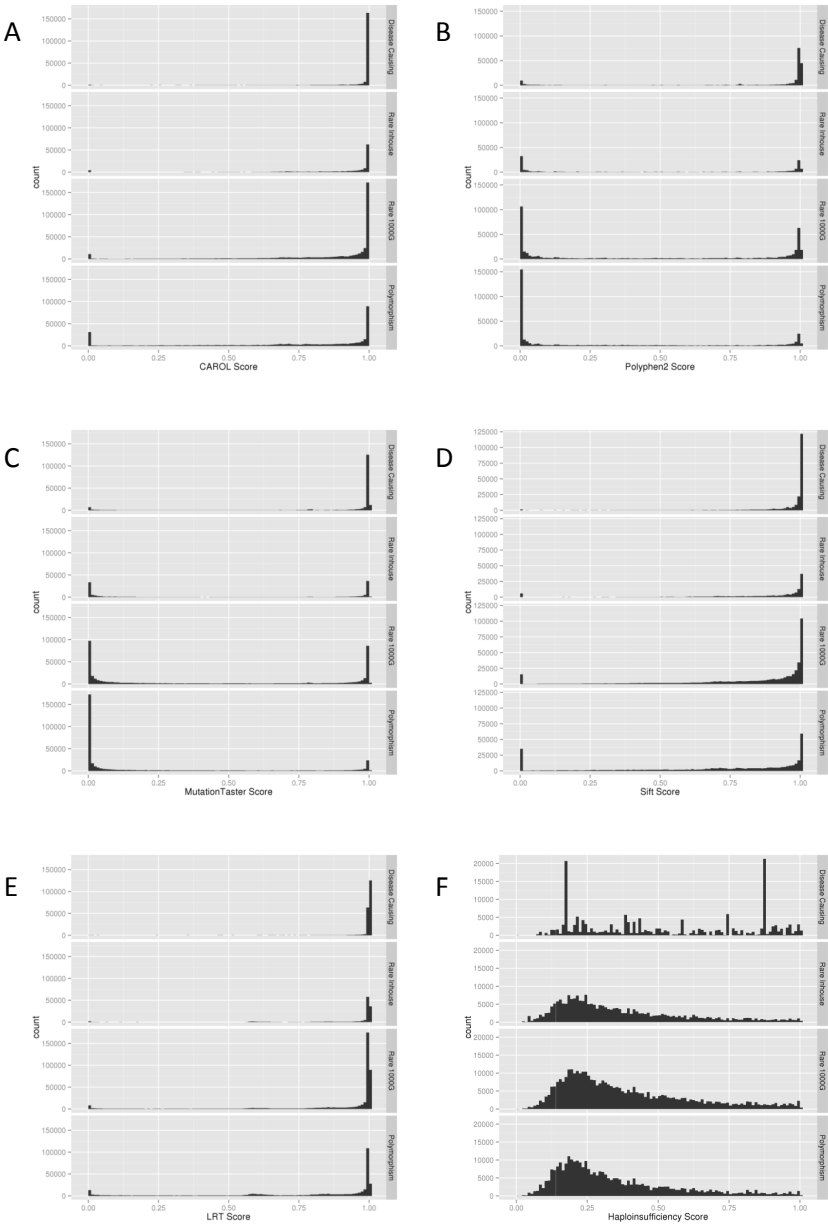
Firstly, there are usually big costs associated with experimental verification of highly prioritized variants, so the good precision of a prioritization method is a desirable feature. Precision (or positive predictive value) is the proportion of truly positive variants among all classified as positive. At the same time, one wants to capture as much as possible of the real disease-causing variants, which renders the sensitivity of the method important as well. Sensitivity (or recall) captures the proportion of the real positives that are classified positive by an algorithm, so it essentially gives information on the expected number of the real disease-causing variants that cannot be detected. Note that eXtasy outperforms all standard methods with regards to these two measures. Secondly, one might be interested in the relative scores produced by prioritization method, rather than in a binary classification based on a threshold of these scores. Downstream analyses are usually performed on the higher prioritized variants first, so the behaviour of the scores that reside in vicinity of the top of the list is of great importance. Examination of the graphical measures, such are ROC and PR curves, can provide quantified insights into this behavior.

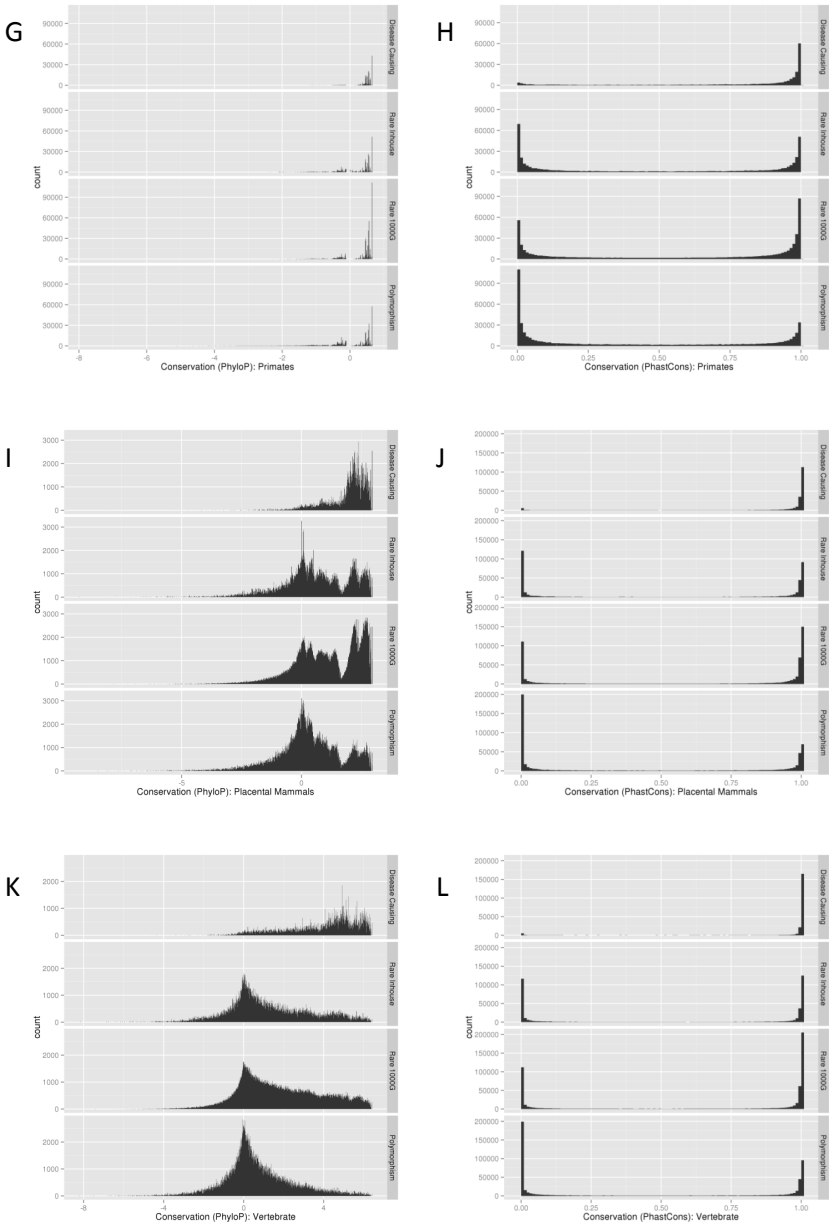
For example, from Supplementary Figure 3 (B) it appears that a big proportion of the real positives in case of the eXtasy prioritizations reside in region of highly prioritized variants, in contrast with the state-of-the-art methods. The precision across different sensitivity values is not just bigger in absolute terms, it also declines later. This is an indicator of increased density of the real positives while moving towards the top of the list.

In our example, this means that one can expect better precision/recall than that reported for the hard-threshold classification, provided that variants are verified in descending order of prioritization scores. Also, this result suggests that the method is less sensitive than previous methods to increase in class imbalance. The latter is desirable property for practical application as the real class inbalance across the whole genome is much more severe than in any benchmarking data set.

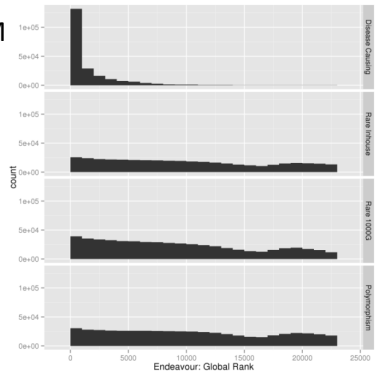
Finally, the aggregates of graphical measures also provide additional views on prioritization performance. For example, the area under the ROC curve could

be, and often is, interpreted as the probability that a real positive example will be ranked higher than a negative. That is, given two examples, eXtasy will prioritize a real positive over a real negative in 97% of the cases. Given the performance measures reported in Supplementary Table 1 and that an average human genome harbors roughly 9000 nonsynonymous mutations (Lupski et al., 2010) and assuming the disease of interest is caused by a single mutation we expect the real mutation to be found in 87% of the times using eXtasy (given the sensitivity. Out of the 9000 mutations and false positive rate of about 5% (FPR = 1-specificity) we expect to call roughly 450 non-disease causing variants as being causative. These numbers correspond to a single point in the ROC curve where we set the decision threshold for eXtasy scores at 0.5, one can easily reduce the FPR by shifting this decision threshold with only a moderate loss in sensitivity (due to the steepness of the ROC curve). Using the published thresholds for SIFT, Polyphen2, MutationTaster and CAROL we would correctly call the disease-causing variant respectively 71%, 78%, 84% and 80% of the time while facing 3150, 4050, 4500, 4050 false positive calls. Compared to MutationTaster which is the, best performing deleteriousness prediction tool reaching (in ROC AUC), eXtasy achieves a 10-fold reduction of false positives.). Although the candidate variant set will be ranked according to their probability of being disease-causing, additional manual inspection and filtering of population databases, expected zygosity, or other samples of the same phenotype can provide further means of prioritization.

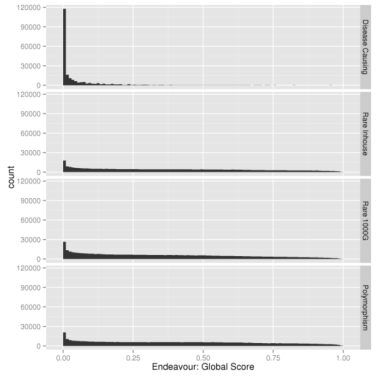

A
B
C
D
E
F
Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
CAROL Score
Polyphen2 Score
MutationTaster Score
Sift Score
LRT Score
Haploinsufficiency Score


G
H
I
J
K
L
Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Conservation (PhyloP): Primates
Conservation (PhastCons): Primates
Conservation (PhyloP): Placental Mammals
Conservation (PhastCons): Placental Mammals
Conservation (PhyloP): Vertebrate
Conservation (PhastCons): Vertebrate

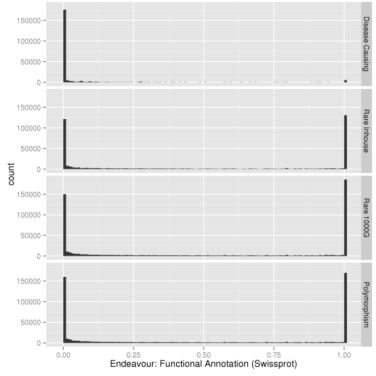
M


Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: Global Rank

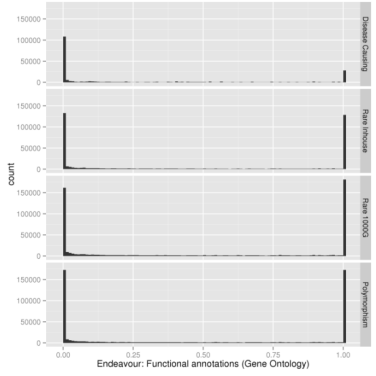
N


Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: Global Score

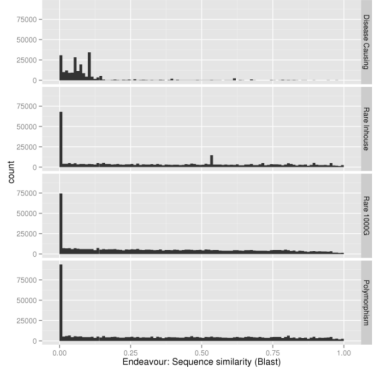
O


Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: Functional Annotation (Swissprot)

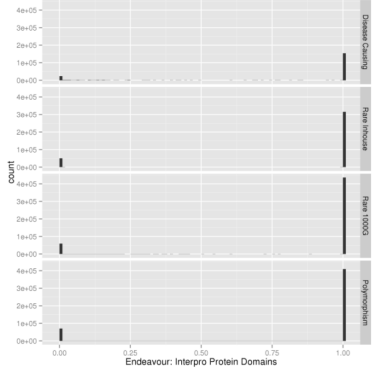
P


Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: Functional annotations (Gene Ontology)

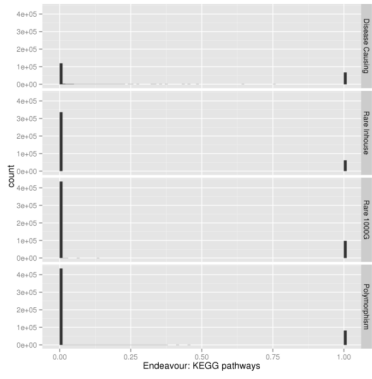
Q


Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: Sequence similarity (Blast)

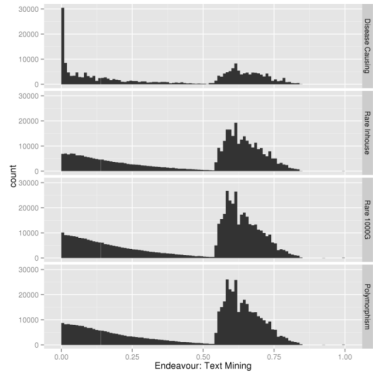
R


Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: Interpro Protein Domains

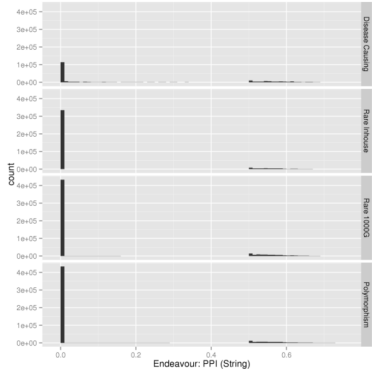
S


Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: KEGG pathways

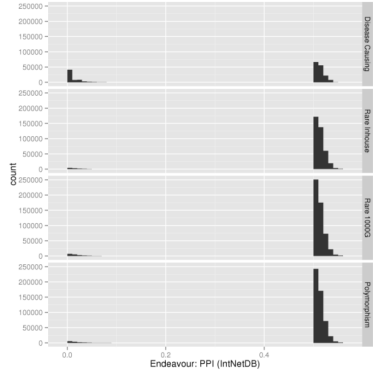
T


Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: Text Mining

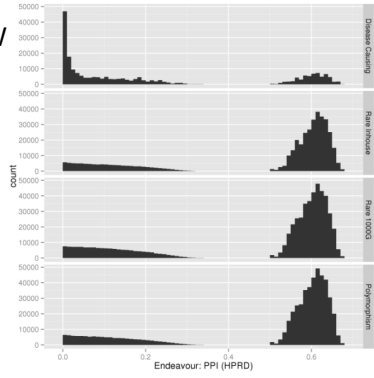
U


Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: PPI (String)

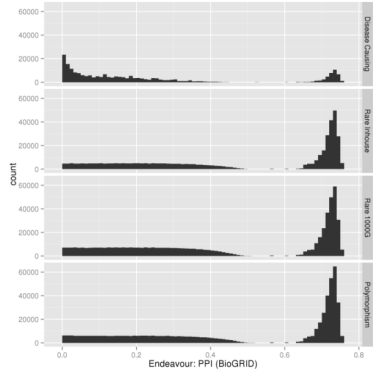
V

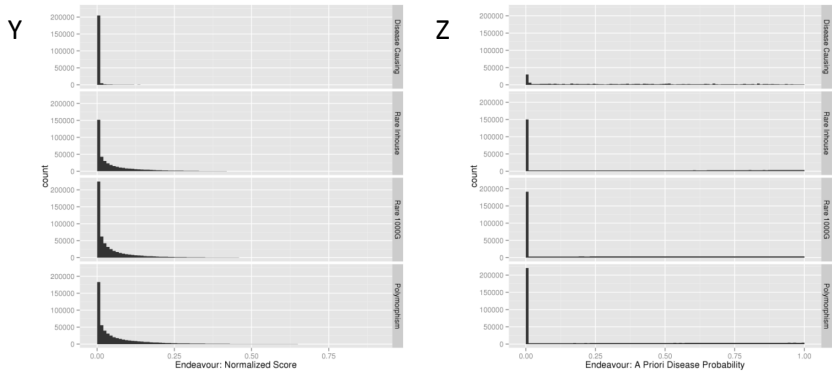

Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: PPI (IntNetDB)

W


Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: PPI (HPRD)

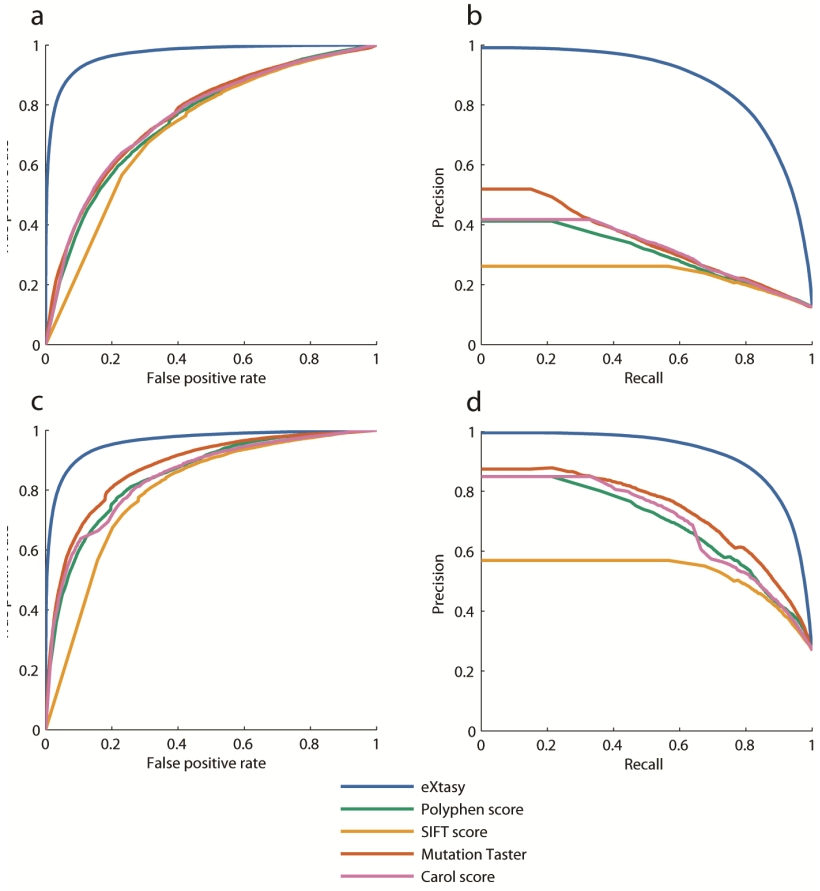
X


Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: PPI (BioGRID)

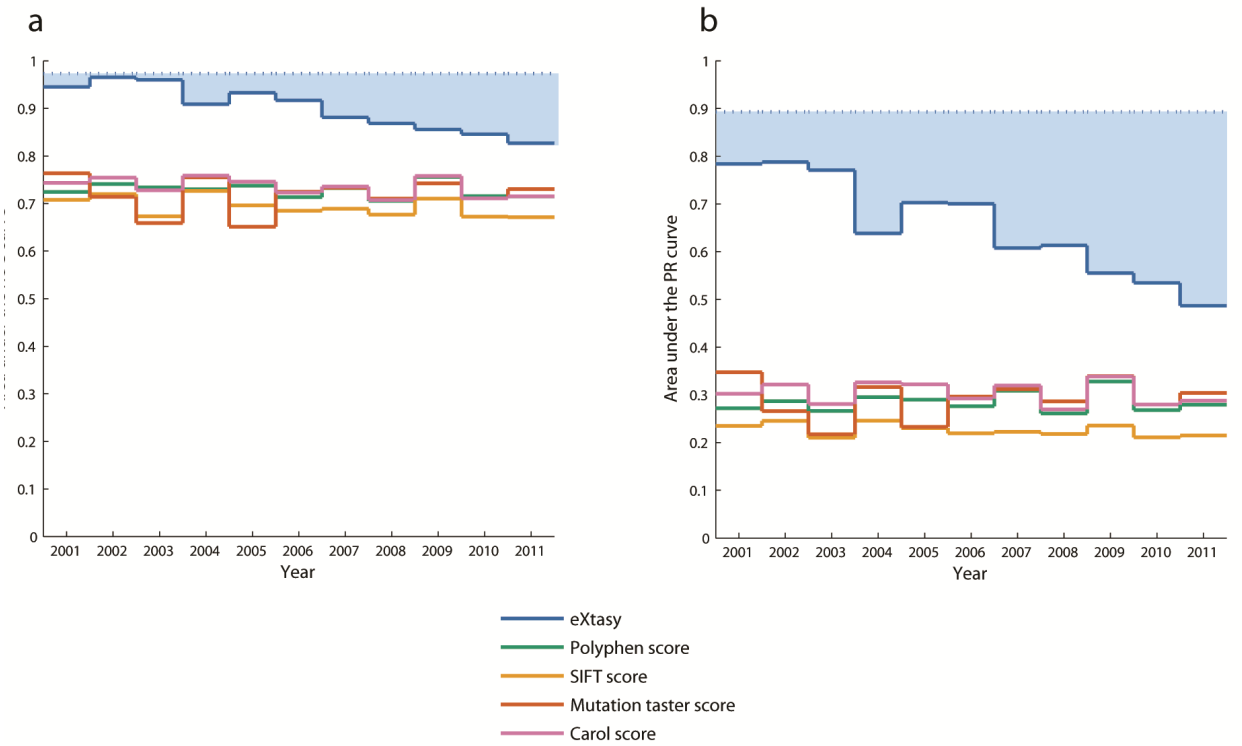


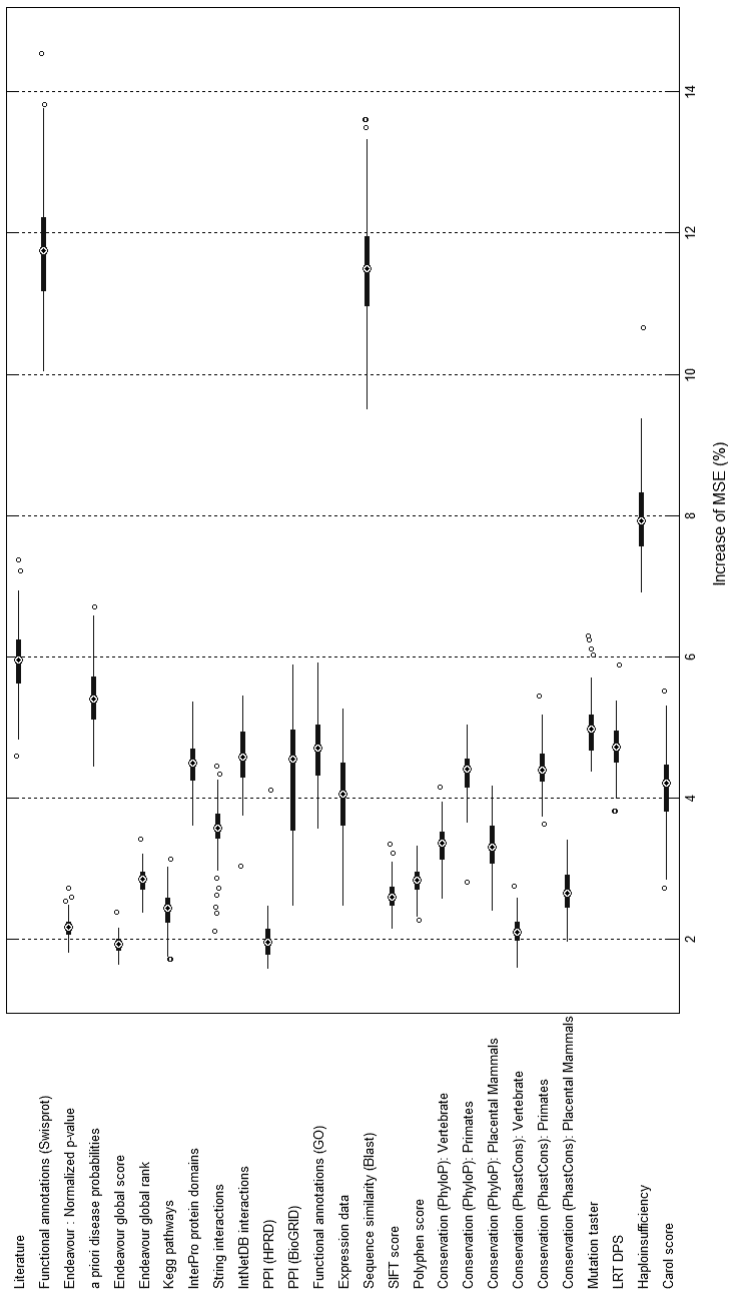
Supplementary Figure 5.1: Plots for the distributions of each of the used features are given for each of the data sets (the positive disease-causing variant set and the 3 control sets). Variables plotted are: A) Carol score B) Polyphen2 score C) MutationTaster score D) Sift score E) LRT score F) Haploinsufficiency scores G) Conservation (PhyloP) for primates H) Conservation (PhyloP) for placental mammals I) Conservation (PhyloP) for vertebrate J) Conservation (PhastCons) for primates K) Conservation (PhastCons) for placental mammals L) Conservation (PhastCons) for vertebrate M) Endeavour: Global rank N) Endeavour: Global score O) Endeavour: Functional Annotation: Swissprot P) Endeavour: Functional Annotation: Gene Ontology Q) Endeavour: Sequence similarity (Blast) R) Endeavour: Interpro Protein Domains S) Endeavour: KEGG Pathways T) Endeavour: Text Mining U) Endeavour: Protein-Protein Interactions: String V) Endeavour: Protein-Protein Interactions: IntNetDB W) Endeavour: Protein-Protein Interactions: HPRD X) Endeavour: Protein-Protein Interactions: BioGRID Y) Endeavour: Normalized score Z) Endeavour: A Priori Disease Probability score

Supplementary Figure 5.2: Receiver-operator (A and C) and Precision-Recall (B and D) curves for five classifiers tested against test sets comparing disease-causing variants to either rare non-disease-causing variants (A and B) or common polymorphisms (C and D). All curves represent the averages of 100 iterations comprising of different random splits of training/testing sets.



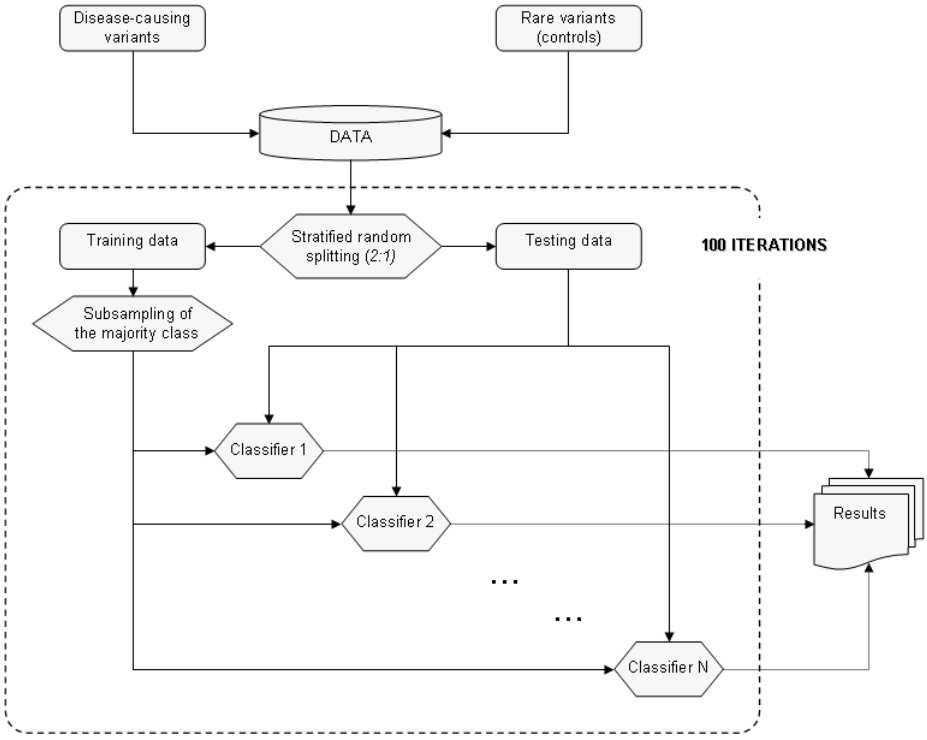
Supplementary Figure 5.3: Values of the area under the Receiver Operator (A) and Precision-Recall (B) curves for the five classifiers obtained during testing by stratifying the positive testing set according to the year of publication. The dotted blue line indicates performance of eXtasy obtained on the full classifier benchmark, while shaded area covers space of expected performances for different use-case scenarios (See Supplementary note 3). eXtasy was trained only on disease-causing mutations published prior to 2001 and all variants corresponding to genes present in both training and test sets were removed from the test set.



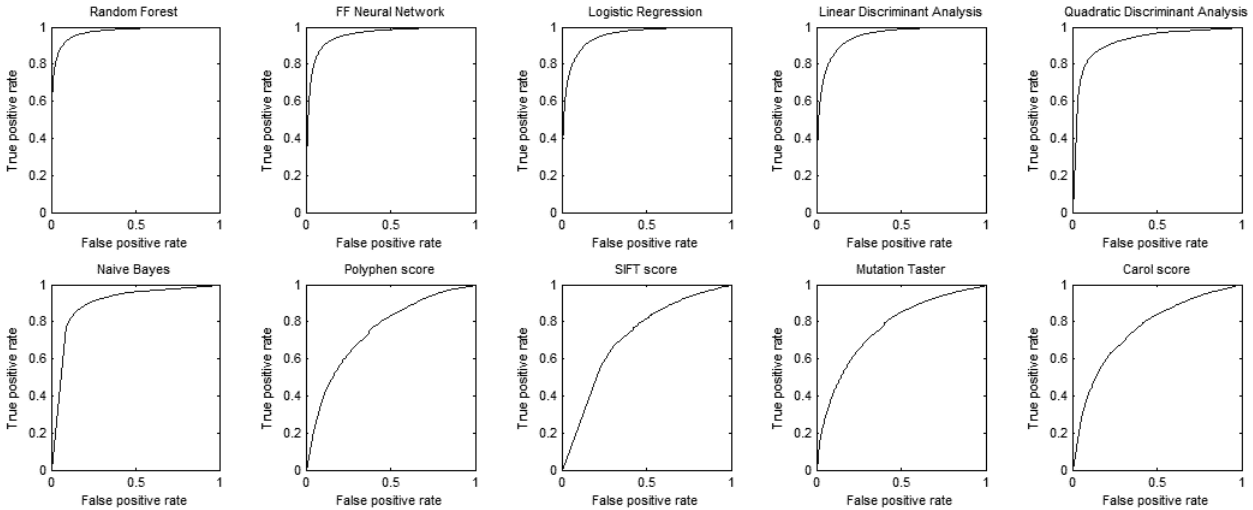


Supplementary Figure 5.4: The box plots depict the range of mean square error increase when an individual feature is shuffled across classes for 100 randomizations. The central mark (dot in circle) corresponds to the median, the box represents the range between the 25th and 75th percentiles, the line depicts the range of the error values and small circles correspond to potential outliers. Non-zero increases in mean square error mean that the feature is informative. However, the relative informativeness of each feature is hard to interpret because some features are not independent of each other (and thus removal/shuffling of the feature is in part compensated y the dependent features).

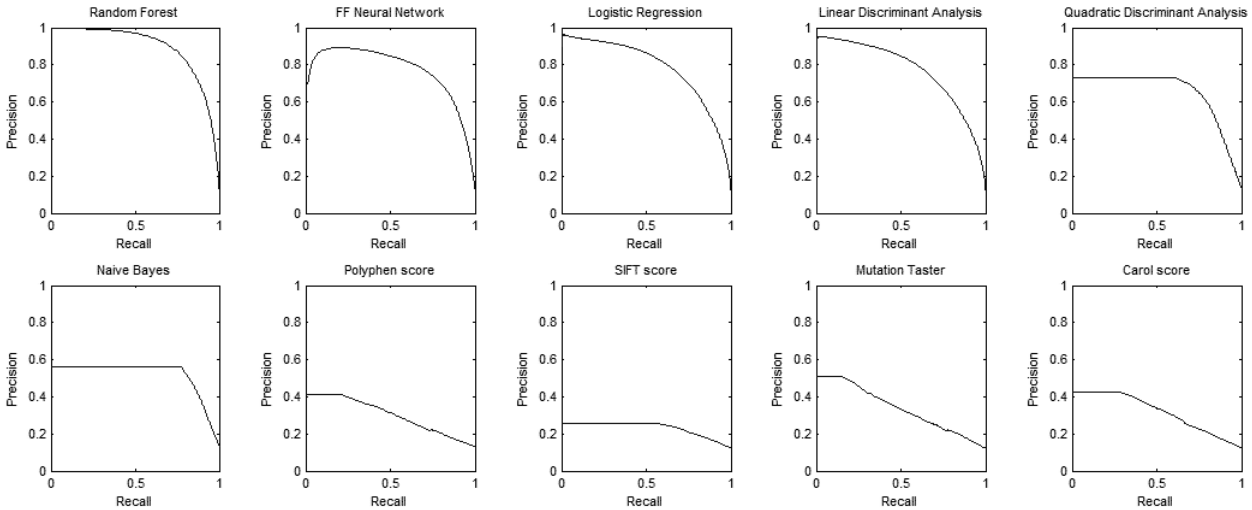
Supplementary Figure 5.5: The data set composed of disease causing and rare variants is randomly divided into two parts, where two-third of the examples are assigned to the training set and one third to the test set (stratified by outcome and divided along gene identifiers). The majority class in the training data set is subsequently sub-sampled to equalize the number of outcomes in each class. The eight classifiers are trained on this data and tested against test sets not containing genes or variants present in the training data. The whole process is repeated 100 times, with different random data splits in each iteration.



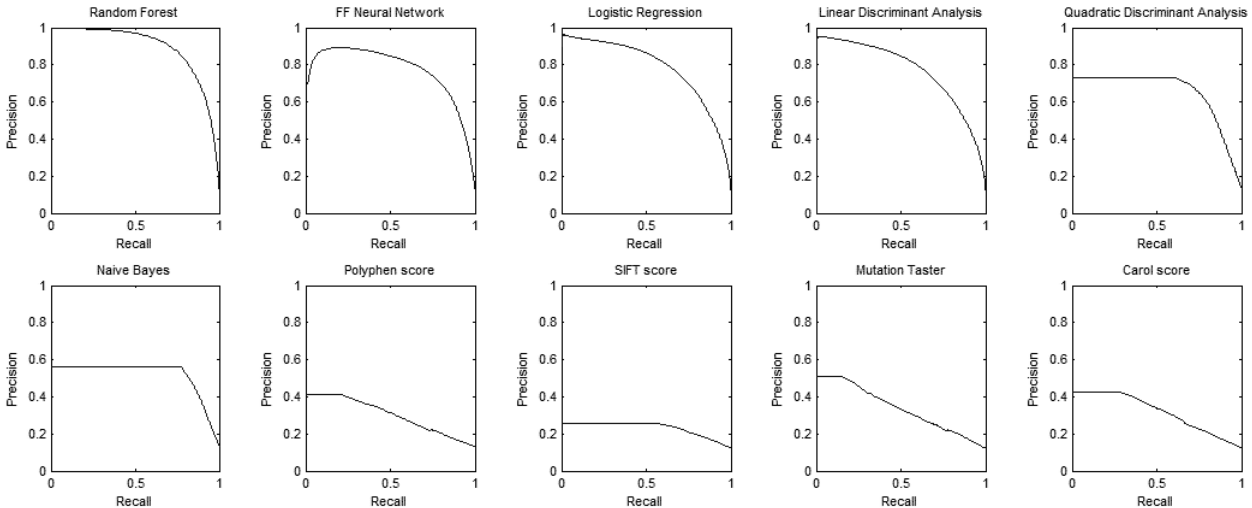
Supplementary Figure 5.6: The averaged ROC curves of 6 classifiers based on genomic data fusion (from Random Forest to Naive Bayes) and 4 deleteriousness prediction methods (from Polyphen to Carol score). The curves are averaged across 100 training vs. test splits of the benchmark data using threshold averaging. The genomic data fusion classifiers clearly outperform four deleteriousness prediction scores across all operating points.



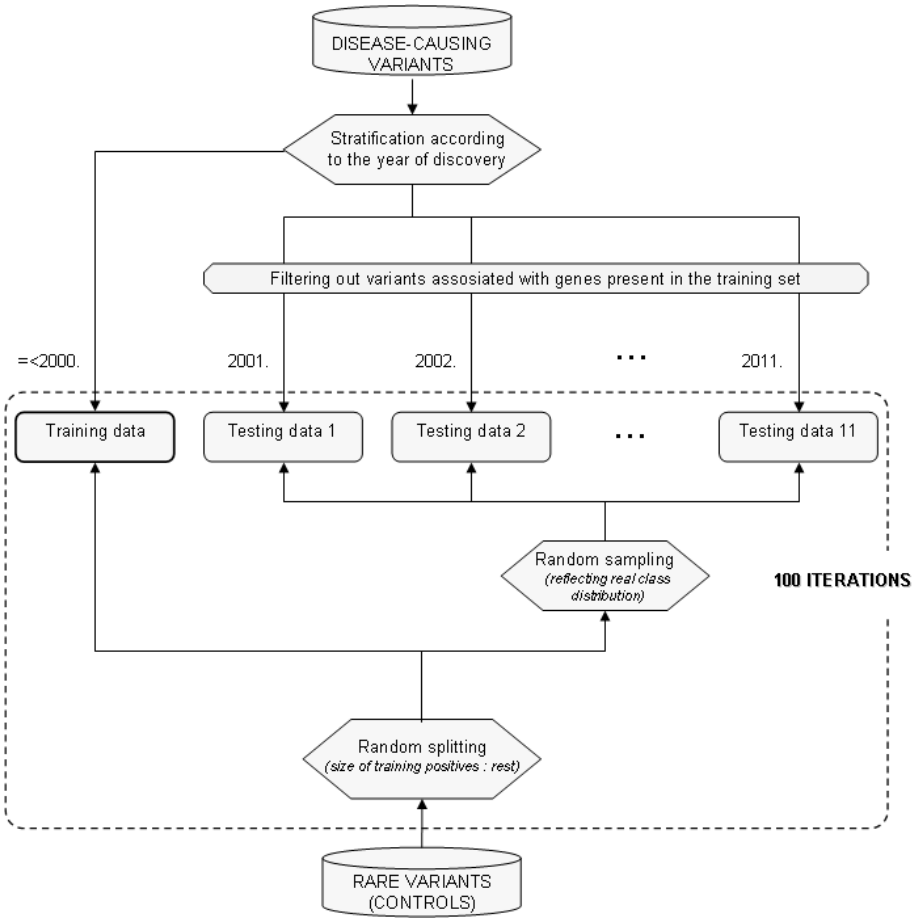
Supplementary Figure 5.7: The averaged PR curves of 6 classifiers based on genomic data fusion (from Random Forest to Naive Bayes) and 4 state-of-the-art methods (from Polyphen to Carol score). The curves are averaged across 100 training vs. test splits of the benchmark data using threshold averaging. The genomic data fusion classifiers clearly outperform four deleteriousness prediction scores across all operating points.



Supplementary Figure 5.8: The data set of disease causing variants is randomly divided into two parts with variants grouped gene by gene, where two-thirds of the examples are assigned to the training set and one third to the test set. The first training/test pair is augmented with negative examples from the rare variants data set using the same splitting procedure. The second pair is augmented with two distinct groups of negatives, rare and polymorphisms. Prior to classification, both training sets are balanced in terms of outcome distribution (not shown on the diagram for clarity). The whole process is repeated 100 times, with different random data splits in each iteration.



Supplementary Figure 5.9: The data set of disease causing variants is divided into twelve parts (one for training and eleven for testing) given the year of discovery. Variants associated with genes present in the training set are filtered out from all of the testing sets. The training data is then enriched with negative examples by random sampling from the rare variants data set (same amount as positives, grouped gene by gene). The rest of the rare variants are randomly assigned to one of the eleven testing sets, in amount that corresponds to the class distribution of the whole data given the number of positives in a year. The whole process of assigning negatives is repeated 100 times with different random splits in each iteration.



Supplementary Table 5.1: Accuracy, sensitivity, specificity, positive predictive values (PPV also referred to as precision), negative predictive values (NPV), Matthews correlation coefficients (MCC), area under the ROC (aROC) and PR (aPR) curves of 8 classifiers based on genomic data fusion and 4 state-of-the-art methods expressed in terms of the corresponding means and standard deviations across 100 training vs. test splits of the benchmark data. The best values of the performance measures across classifiers are indicated in bold.

Classifier/score		Accuracy		Sensitivity		Specificity		PPV		NPV		MCC		aROC	aPR
		mean	*std.*	*mean*	*std.*	*mean*	*std.*	*mean*	*std.*	*mean*	*std.*	*mean*	*std.*		
Genomic data fusion-based classification	Random forest	**0.9402**	0.0045	0.8725	0.0298	**0.9496**	0.0060	**0.7112**	0.0309	0.9815	0.0033	**0.7545**	0.0260	**0.9729**	**0.8929**
	Decision trees	0.8745	0.0075	0.8695	0.0311	0.8749	0.0103	0.4977	0.0341	0.9794	0.0042	0.5957	0.0293	N/A	N/A
	FF Neural Network	0.8547	0.0088	0.9291	0.0163	0.8439	0.0115	0.4590	0.0326	0.9883	0.0021	0.5877	0.0251	0.9541	0.7711
	Logistic regression	0.8118	0.0090	0.9410	0.0145	0.7931	0.0120	0.3934	0.0304	0.9896	0.0022	0.5300	0.0233	0.9509	0.7856
	k-Nearest neighbours	0.7882	0.0098	0.9272	0.0207	0.7681	0.0133	0.3631	0.0285	0.9868	0.0033	0.4929	0.0238	N/A	N/A
	LDA	0.7718	0.0098	**0.9526**	0.0134	0.7458	0.0126	0.3483	0.0289	**0.9911**	0.0021	0.4865	0.0229	0.9474	0.7707
	QDA	0.7814	0.0111	0.9106	0.0170	0.7627	0.0148	0.3539	0.0289	0.9837	0.0026	0.4764	0.0229	0.9216	0.6507
	Naive Bayes	0.7565	0.0118	0.9141	0.0152	0.7338	0.0156	0.3289	0.0269	0.9837	0.0025	0.4497	0.0214	0.8988	0.5169
State-of-the art methods	Polyphen score	0.6578	0.0041	0.7119	0.0207	0.6499	0.0029	0.2251	0.0251	0.9406	0.0061	0.2446	0.0218	0.7499	0.3013
	SIFT score	0.5779	0.0049	0.7855	0.0165	0.5482	0.0043	0.1989	0.0203	0.9470	0.0075	0.2204	0.0145	0.7180	0.2323
	Mutation Taster	0.5457	0.0076	0.8482	0.0196	0.5024	0.0051	0.1959	0.0215	0.9588	0.0059	0.2326	0.0200	0.7647	0.3390
	Carol score	0.5863	0.0052	0.8080	0.0112	0.5545	0.0037	0.2058	0.0215	0.9529	0.0055	0.2396	0.0154	0.7602	0.3187

Supplementary Table 5.2: The performance is expressed in terms of area under the ROC and PR curves. The best values of a performance measure across scores and scenarios are indicated in bold.

Method	**Area under the ROC curve**		**Area under the PR curve**	
	Rare non-disease causing	Common polymorphisms	Rare non-disease causing	Common polymorphisms
eXtasy	**0.9691**	**0.9616**	**0.8744**	**0.9244**
Polyphen score	0.7516	0.8504	0.3040	0.6888
SIFT score	0.7207	0.7974	0.2351	0.5275
Mutation Taster	0.7659	0.8774	0.3433	0.7350
Carol score	0.7627	0.8476	0.3199	0.6956

Supplementary Table 5.3: Numbers of disease-causing variants and corresponding unique genes in training and testing sets, after removing these associated with genes present in the training. The number of control variants per year is proportional to the number of disease-causing variants.

Data granularity level	**Training years**	**Testing years**										
	1980-2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Variants	5658	156	164	294	261	400	304	339	357	465	533	505
Genes	292	46	55	87	82	98	117	115	122	148	174	194

Supplementary Table 5.4: Area under the ROC and PR curves for five prediction scores for disease causing variants published during eleven consequent years (2001-2011); with standard deviations (in brackets). The best value of a performance measure given a score and a year is indicated in bold.

	Method	Year										
		2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Area under the ROC curve	eXtasy	**0.9456** (0.0084)	**0.9656** (0.005)	**0.9601** (0.0039)	**0.9088** (0.0096)	**0.9332** (0.0059)	**0.9170** (0.0055)	**0.8813** (0.0069)	**0.8685** (0.0088)	**0.8559** (0.0074)	**0.8460** (0.0083)	**0.8268** (0.009)
	Polyphen score	0.7243 (0.0081)	0.7408 (0.0088)	0.7342 (0.0066)	0.7302 (0.0068)	0.7374 (0.0057)	0.7134 (0.0067)	0.7325 (0.0063)	0.7057 (0.0063)	0.7561 (0.0046)	0.7155 (0.0046)	0.7146 (0.0051)
	SIFT score	0.7079 (0.0099)	0.7195 (0.0094)	0.6730 (0.0071)	0.7265 (0.0073)	0.6961 (0.0055)	0.6848 (0.0076)	0.6890 (0.0059)	0.6768 (0.0069)	0.7103 (0.0054)	0.6727 (0.0047)	0.6714 (0.0053)
	Mutation Taster	0.7634 (0.0087)	0.7141 (0.0091)	0.6590 (0.0066)	0.7554 (0.0074)	0.6514 (0.005)	0.7250 (0.0067)	0.7334 (0.0068)	0.7103 (0.0055)	0.7425 (0.0043)	0.7107 (0.0048)	0.7303 (0.005)
	Carol score	0.7436 (0.0086)	0.7543 (0.0091)	0.7280 (0.0064)	0.7586 (0.0066)	0.7459 (0.0056)	0.7230 (0.007)	0.7356 (0.0058)	0.7076 (0.0063)	0.7579 (0.0047)	0.7108 (0.0046)	0.7150 (0.005)
Area under the PR curve	eXtasy	**0.7840** (0.0276)	**0.7879** (0.0304)	**0.7712** (0.0233)	**0.6386** (0.0237)	**0.7027** (0.0258)	**0.7007** (0.0166)	**0.6077** (0.0201)	**0.6133** (0.0173)	**0.5553** (0.0166)	**0.5347** (0.0187)	**0.4869** (0.0136)
	Polyphen score	0.2717 (0.0116)	0.2868(0.012)	0.2663 (0.0092)	0.2951 (0.0094)	0.2900 (0.0092)	0.2760 (0.0102)	0.3085 (0.0098)	0.2610 (0.0076)	0.3278 (0.0092)	0.2678 (0.0064)	0.2793 (0.0073)
	SIFT score	0.2347 (0.0092)	0.2451 (0.0087)	0.2099 (0.0057)	0.2460 (0.0063)	0.2301 (0.0047)	0.2191 (0.0063)	0.2223 (0.0048)	0.2177 (0.0061)	0.2355 (0.0047)	0.2108 (0.0038)	0.2147 (0.0044)
	Mutation Taster	0.3475 (0.0179)	0.2659 (0.0144)	0.2173 (0.0071)	0.3160 (0.0142)	0.2327 (0.0072)	0.2963 (0.0117)	0.3114 (0.0121)	0.2863 (0.0103)	0.3392 (0.0117)	0.2794 (0.0076)	0.3040 (0.0079)
	Carol score	0.3020 (0.0126)	0.3216 (0.015)	0.2804 (0.0091)	0.3262 (0.0102)	0.3220 (0.009)	0.2922 (0.0102)	0.3195 (0.0105)	0.2694 (0.0082)	0.3387 (0.0082)	0.2794 (0.007)	0.2877 (0.0077)

Supplementary Table 5.5: Matlab functions or classes used for implementation of classifiers with corresponding parameters. Only the implementation-independent parameters of the classification methods are displayed, with the exception of logistic regression, LDA and QDA where Matlab-specific parameters are used to choose the classifier. All implementation-dependent parameters that are not enumerated in this table have been set to default values.

Classifier name	Matlab R2010a function/class	Parameters
Random forest	TreeBagger class method predict	Number of trees in ensemble : 100 Number of variables to select at random for each decision split. : square root of the number of variables (default) Minimal number of observations in a leaf : 1 (default)
Decision trees	classregtree class, method eval	Pruning : on (default) Minimal number of observations in a leaf : 1 (default)
Feedforward Neural Network	Functions newff, train and sim	Number of hidden layers : 1 Number of neurons in the hidden layer: 200 Number of training epochs : 100
Logistic regression	Functions glmfit and glmval	Distribution : binomial Link : logit (default for binomial)
k-Nearest neighbours	Function knnclassify	Distance measure : Euclidean (default) Number of neighbours considered : 15
Quadratic Discriminant Analysis	Function classify	Type : quadratic
Linear discriminant analysis	Function classify	Type : linear
Naive Bayes	NaiveBayes class, Methods fit, predict and posterior	/

Chapter 6

Comparison of aggregation methods for multiphenotype variant prioritization

> If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent. point), or one in a hundred (the 1 per cent. point). Personally, the writer prefers to set a low standard of significance at the 5 per cent. point, and ignore entirely all results which fail to reach this level.
>
> Ronald A. Fisher

6.1 Summary

Diseases are usually comprised of a conglomerate of several typical and atypical phenotypic clinical presentations. Instead of dealing with the more complex aggregate phenotypes, variant prioritization approaches usually separate the problem into individual prioritizations per phenotype and aggregate information a posteriori. In this chapter we investigate different aggregation strategies to merge individual variant prioritizations into a global prioritization across all of the presented phenotypes. We compare simple techniques such as taking the maximum or mean score, to more involved methodologies such as order statistics and parametric modelling. We show that all methods perform reasonably well and that the selection of an optimal aggregation method depends on the selection of the underlying phenotypes.

6.2 Comparison of aggregation methods for multi-phenotype variant prioritization

The PhD candidate implemented, performed and interpreted the discussed benchmarks and analyses. The PhD candidate, D. Popovic and Y. Moreau conceptually designed the presented parametric modelling approach. The PhD candidate wrote the draft of the manuscript.

Pre-print available at *bioRxiv* doi: http://dx.doi.org/10.1101/064899 (2016)

Comparison of aggregation methods for multiphenotype exomic variant prioritization

Alejandro Sifrim[1,2], Dusan Popovic[1,2], Leon-Charles Tranchevent[3], Joris R. Vermeesch[4], Jan Aerts[1,2], Bart De Moor[1,2], Yves Moreau[1,2]

[1]Department of Electrical Engineering, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium
[2]iMinds Department of Medical Information Technology, Leuven, Belgium
[3]Centre de Recherche en Cancérologie de Lyon (CRCL), Lyon, France
[4]Laboratory of Molecular Cytogenetics and Genome Research, KU Leuven, Leuven, Belgium
Corresponding author: Yves Moreau, yves.moreau@esat.kuleuven.be

Abstract

Recently, new computational methods have been proposed to prioritize variants by scoring them not only based on their potential impact on protein function but also on their relevance given the available information on the disease under study. Usually these diseases comprise several phenotypic presentations, which are separately prioritized and then aggregated into a global score. In this study we compare several simple (maximum and mean score) and more complex aggregation methods (order statistics and parametric modeling) to obtain the best possible prioritization performance. We show that all methods perform reasonably well (median rank below 20 out of more than 8000 variants) and that the selection of an optimal aggregation method depends strongly on the fraction of uninformative phenotypes. Finally, we propose guidelines as to how to select an appropriate aggregation method based on knowledge of the phenotype under study.

Introduction

In recent years, whole-exome sequencing has facilitated the discovery of genes underlying rare Mendelian disease by allowing the identification of disease-causing genomic variation [84, 152, 86, 85, 166]. Several different strategies have been proposed to sift through the thousands of variants identified in a

single human exome to retain a handful of meaningful candidate variants for further downstream functional validation studies [167]. These strategies usually start by filtering variants with a higher potential for disturbing gene function, such as gaining a stop-codon, resulting in a frameshift, changing the amino-acid sequence of the protein or interfering with a splice site. In this first filtering step intergenic, intronic and synonymous mutations are discarded. A second filtering step removes all variants present above a certain minor allele frequency (e.g. the variant occurs in less than 1% of the population) in commonly used population databases such as the 1000 Genomes Project [168] or the NHLBI Exome Sequencing Project [18]. This step relies on the assumption that variants causing rare disease, following either dominant or recessive inheritance patterns, would not be present in high frequencies in populations of healthy individuals. Even after such filtering several hundreds of potential candidate variants remain and additional filtering is required. At this point strategies diverge depending on the available data. If familial samples are available, linkage analysis or searching for de novo variants can sufficiently narrow down the list to only a couple of variants [87]. Another strategy is to sequence a number of unrelated individuals exhibiting very similar phenotypical presentations and looking for genes recurrently struck by variants in several of these individuals [152, 85]. This strategy usually assumes a monogenic inheritance pattern so as to have sufficient signal to overcome the noise of neutral genomic variation. Additionally, acquiring sufficient samples is difficult as patient recruitment usually has to happen across several clinical institutions because of the rarity of the disorder in any given population.

Several computational methods have been developed to predict the impact of a given variant based on various biochemical, structural and evolutionary properties [71, 146, 124]. Although these methods provide reasonable performance in predicting variants affecting protein function, they lack the specificity to identify true disease-causing variants because the human genome is riddled with mildly deleterious variation unrelated to the specific phenotype under study [156, 79]. Recently, new computational approaches have been proposed to address this shortcoming by considering not only functional impact but also phenotypic relevance [139, 169]. In a recent study we proposed a method, named eXtasy, that integrates deleteriousness and haploinsufficiency prediction with phenotype-specific gene prioritization through genomic data fusion [170]. This approach relies on the similarity across several data sources (e.g. gene expression, protein-protein interactions, sequence similarity) between known phenotype-associated genes and the mutated gene [94, 95]. This process helps discriminating variants in genes likely to be involved in the phenotype and variants in genes which might contain functional variants but for which there is no evidence of the genes being involved in the disease.

In this paper we elaborate on the problem of aggregating across different prioritizations when the observed clinical presentation is an agglomeration of different phenotypes, as is often the case in rare Mendelian disorders. For example, Miller syndrome, one of the first Mendelian diseases elucidated by massively parallel exome sequencing [85], involves a combination of craniofacial abnormalities, postaxial limb deformities and sometimes internal malformations. It is reasonable to assume that some of these phenotypes can be more informative than others when considering them for variant prioritization, because (1) some of them might be better understood in the current literature, (2) have simpler underlying molecular mechanisms or (3) have lower inherent phenotypic variability. Yet estimating the relative informativeness of phenotypes would require extremely in-depth knowledge of these properties. To circumvent this problem in eXtasy, we used as an *ad hoc* score the maximum score obtained from any single variant prioritization for a given variant. Thus, if one of the phenotypes delivered an informative prioritization, this approach garantueed that this information would be taken into account (increasing sensitivity of the method). However, this approach potentially also inflated our scores for non-disease causing variants, leading to increased false positive rates. Using the maximum essentially discards information when more than one informative phenotype is used for prioritization because only the best phenotype is considered. In this study, we propose three alternative ways of aggregating over multiple phenotype-specific prioritizations using either classical order statistics [171, 94], robust rank aggregation [172] and statistical modeling of score distributions. To estimate and compare their performance, we also adapt our benchmarking scheme to resemble a more real-life application of variant prioritization. We show that although in most cases the maximum surprisingly outperforms other aggregation methods, using more complex aggregation methods is usually the best when phenotypes are carefully selected. These methods also offer the added benefit of obtaining probabilistic outcomes which could be integrated further in other statistical frameworks, such as case/control association studies.

Material & Methods

Benchmark

To compare the performance of the different aggregation methods we set up a benchmark that closely resembles a real-life application of variant prioritization using whole-exome sequencing. We first generated synthetic exomes by randomly sampling nonsynonymous variants from the 1000Genomes Project. For every nonsynonymous variant in the 1000 Genomes Project (October 2012 release) a

random number between 0 and 2,184 (the maximum number of times an allele could be present in the 1,092 diploid individuals) was generated and divided by 2,184. If this randomly generated number was smaller than the observed global minor allele frequency for that variant, then this variant was included in the synthetic exome. We chose the Human Gene Mutation Database [159] as a set of disease-causing variants, and selected 16,272 nonsynonymous variants that were suitable for variant prioritization and studying phenotype aggregation. More specifically, we selected disease-causing variants having at least 3 different phenotypes with at least 3 known gene-phenotype associations each. We then injected each of these variants into a randomly generated synthetic exome. We then ran these synthetic exomes containing one disease-causing nonsynonymous mutation through eXtasy for each phenotype describing the disease. For each exome we aggregated the resulting scores of the different prioritization either using the maximum, mean, classic order statistics and robust order statistics and using a statistical modeling approach. We then obtained the rank of the injected disease-causing variant for each aggregation method (Figure 6.1). We additionally computed the fraction of informative phenotypes for a disease-causing variant as the number of phenotypes with an eXtasy score over 0.5 over the total number of phenotypes prioritized. This threshold is the same as the threshold at which a variant is considered disease-causing in the original eXtasy paper and corresponds to the majority vote of trees in the Random Forest model.

Non-parametric order statistics

Nonparametric order statistics have been previously used to aggregate rankings of genes obtained from different experiments or data sources [171, 94]. These methods methods compare sets of ranks indirectly by comparing them to the null hypothesis that the ranks were randomly drawn. These methods potentially have the benefit of being robust to non-informative rankings (for example adding an erroneous phenotype in the variant prioritization), without losing the added power of additional informative rankings (which are lost by aggregating using the maximum score). In an initial formulation of a gene rank aggregation algorithm, an algorithm was proposed which estimates the cumulative density function of the order statistic under the null hypothesis using a recursive formulation. This algorithm was computationally prohibitive for moderate numbers of ranked lists and was further refined by Aerts and colleagues (2006) decreasing the algorithmic complexity and thus computational time. Both formulations assume the null hypothesis that all ranked lists are potentially informative. Another computationally efficient formulation called Robust Rank Aggregation (RRA) [172] refines the model and allows a fraction of the ranked lists is allowed to be uninformative. In their benchmarks [172] showed that both RRA and

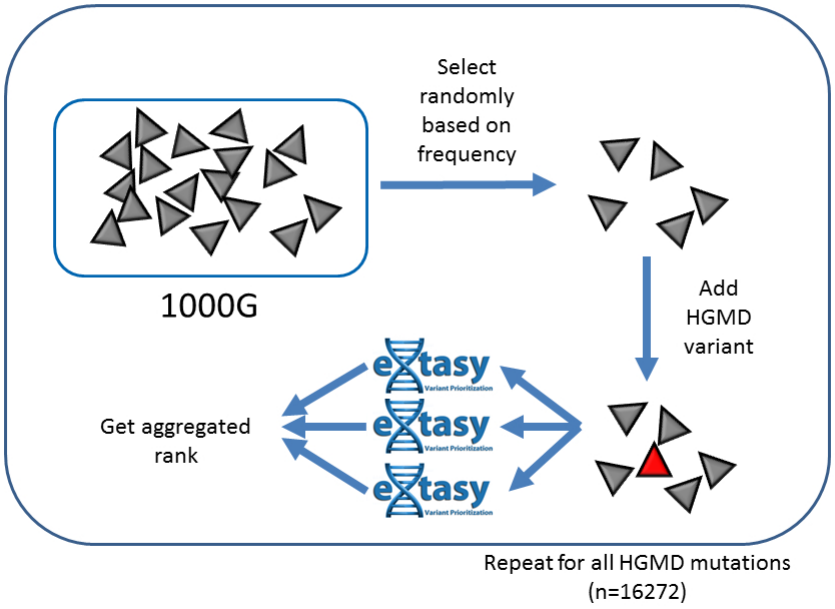


Figure 6.1: Schematic representation of the benchmarking scheme: Nonsynonymous variants are first sampled from 1000Genomes Project data based on their respective minor allele frequency, generating a synthetic control exome. Then, one disease-causing variant from HGMD is added to a single synthetic control exome and eXtasy prioritizations are performed for each phenotype annotated to the disease-causing variant in that exome. The different phenotype-specific prioritizations are then combined into a single score using different aggregation schemes and the ranks for the injected disease-causing variant is compared across the different methods. This process is repeated for each of the 16272 nonsynonymous HGMD variants.

the formulation of Aerts and colleagues perform similarly, but that the latter algorithm becomes unstable for large numbers of ranked lists ($N > 40$). In our situation it is unlikely that such numbers of phenotypes would need to be integrated in variant prioritization experiments. In this study, we use the R package (RobustRankAggreg) implemented by Kolde et al. which implements both algorithmic order statistic formulations and apply both of them to our benchmark to rank variants according to their estimated significance.

Parametric modeling

Order statistic approaches are extremely generalizable as they do not make any assumptions about the underlying score distributions by transforming the data into rankings. Although this is extremely useful in many situations, such nonparametric approaches are usually less powerful than their parametric counterparts, which take the underlying distributions into account. While studying the eXtasy output score distributions, we observed highly reproducible non-Gaussian density functions skewed towards 0 and with long tails towards higher scores, which can be best fitted using a Gamma distribution model. To devise an aggregate score across different eXtasy prioritizations we developed a parametrical statistical approach and applied it to our benchmark. Assume n phenotypes and the corresponding prioritizations $i = (1, ..., n)$, each containing a set of scores $s_{i,k}$ with $k = (1, ..., K)$ where K is the number of variants under study. For each prioritization of the variants against one given phenotype we robustly fitted a Gamma distribution Γ_i with parameters k_i and θ_i representing the shape and scale of the distribution using the robust package in R. We then computed the probability $p_{i,k}$ of seeing a score equal to or more extreme than $s_{i,k}$ given the cumulative density function of Γ_i. To aggregate these individual probabilities into a single probability of seeing a score, we use as null hypothesis that the $s_{i,k}$'s are drawn independently from the underlying Gamma distributions Γ_i across the n different prioritizations and asses how strongly $s_{i,k}$'s actually observed deviate from this null hypothesis. To this end, we use the p-values $p_{i,k}$ and compute the Fisher's omnibus meta-analysis statistic $r_k = \sum_{i=1}^{n} \ln p_{i,k}$ [173]. Under our null hypothesis, the $p_{i,k}$'s are drawn uniformly between 0 and 1, and our test statistic follows a χ^2 distribution with $2n$ degrees of freedom. Using critical values of the χ^2 distribution, we can assign a p-value to the observed r_k. Because of the large number of variants tested we correct these probabilities using Benjamini-Hochberg False Discovery Rate correction. We then rank the variants for each synthetic exome in our benchmark given the aggregated multiple testing corrected p-value.

Results

In contrast with the previously reported eXtasy case-control classification benchmark, our goal with the synthetic exome benchmark was to simulate a real-life situations where variant prioritization would be used and resembles the validation benchmark proposed by Robinson and colleagues [169]. Our exome generation algorithm generated exomes containing an average of 8399 nonsynonymous variants ($n = 16272$, $sd = 242$). This number is higher than the roughly 6000 variants reported from large-scale exome sequencing projects [18]. However, this is in line with results from whole-genome sequencing studies [174, 168] that do not face the difficulties related to coverage variability caused by exome-capture techniques. As an indication of real-life performance, we describe the rank of the injected disease-causing variant as a performance metric, as opposed to classical metrics (*e.g.*, precision and recall), to compare aggregation schemes for multiphenotype variant prioritization using eXtasy. Overall, we observe that all aggregation methods perform well (Table 6.1), having a median rank for the disease-causing mutation in the top 20 and a below the top 100 in 75% of the cases (with the exception of the median eXtasy score and RRA). Looking at the cumulative distribution functions of the ranks of the different methods (Figure 6.2), we can see that, surprisingly, using the maximum eXtasy score (as used in the original eXtasy publication) delivers the best rankings compared to the other methods. Aggregation methods using the average eXtasy score or the more involved parametric and nonparametric methods seem to show similar performance. Finally, RRA and the median eXtasy score provide the worst performance out of all the aggregation schemes. These results indicate that uninformative or incorrectly prioritized phenotypes might affect the global ranking significantly, which might explain why using the maximum score (being the most robust method) generally outperforms other aggregation schemes. To further investigate this hypothesis, we looked at how varying degrees of uninformative phenotypes affect the different aggregation schemes by stratifying our benchmark according to the fraction of individual phenotypes correctly prioritized (*i.e.*, phenotypes for which the disease-causing variants receives an eXtasy score over 0.5).

After stratification into low ($n = 5862$), moderate ($n = 8559$) and high ($n = 1837$) fractions of informative phenotypes we observe that no single aggregation method performs best across the three classes (Figure 6.3). When only less than a third of phenotypes are informative, the average rank of the disease-causing variants degrades significantly independently of the aggregation method (Table 6.2). In this scenario, aggregation using the maximum eXtasy score delivers the best ranking compared to all other methods, while all other methods show nearly equal performance (with the exception of RRA, showing significantly lower performance). When roughly half of the phenotypes are informative,

Table 6.1: Summary statistics for the overall comparison of different aggregation methods: This table shows the mean, median, 25%-quantile (Q1) and 75%-quantile (Q3) of the global ranks of disease-causing variants ($n = 16272$) using different aggregation schemes.

Aggregation Method	Q1	Median	Mean	Q3
Average eXtasy Score	3	10	59	39
Maximum eXtasy Score	1	5	41	28
Nonparametric Order Statistics: Stuart et al.	3	10	64	42
Nonparametric Order Statistics: RRA	5	20	102	126
Parametric Modelling	2	9	111	37

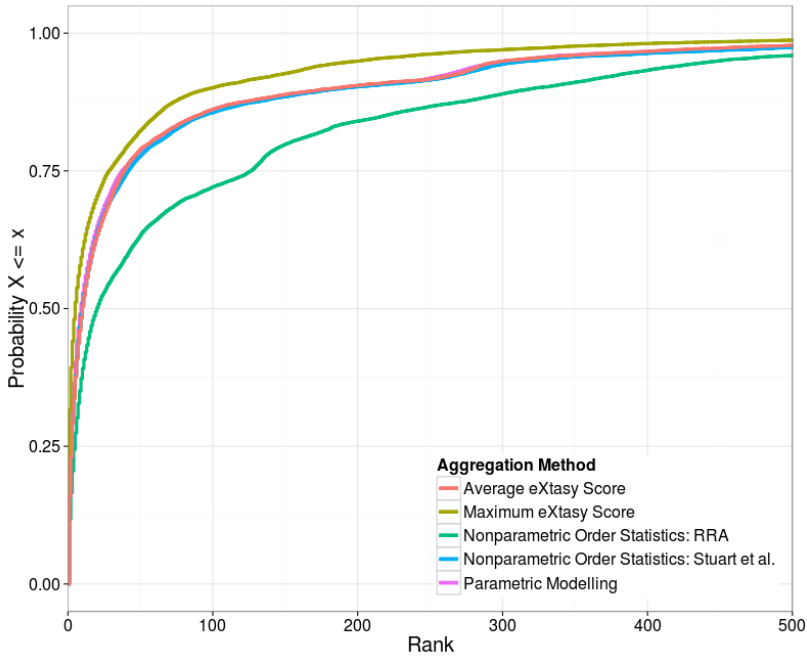


Figure 6.2: Empirical Cumulative Distribution Function (ECDF) of ranks comparing different aggregation schemes: This chart shows the proportion of variants out of all disease-causing variants ($n = 16272$) with global ranks after aggregation lower or equal than the rank shown on the x-axis.

all aggregation methods (except RRA) perform reasonably well (average rank lower than 10, and 75% of variants ranked below 12). Additionally, methods which leverage information across different phenotypes (mean score, order statistics, parametric modelling) perform slightly better than aggregation using the maximum score. In the case where more than two-thirds of the phenotypes are informative 75% of the disease-causing variants are in the top 3 ranked mutations for all methods except aggregation by maximum score (75% of variants in the top 10).

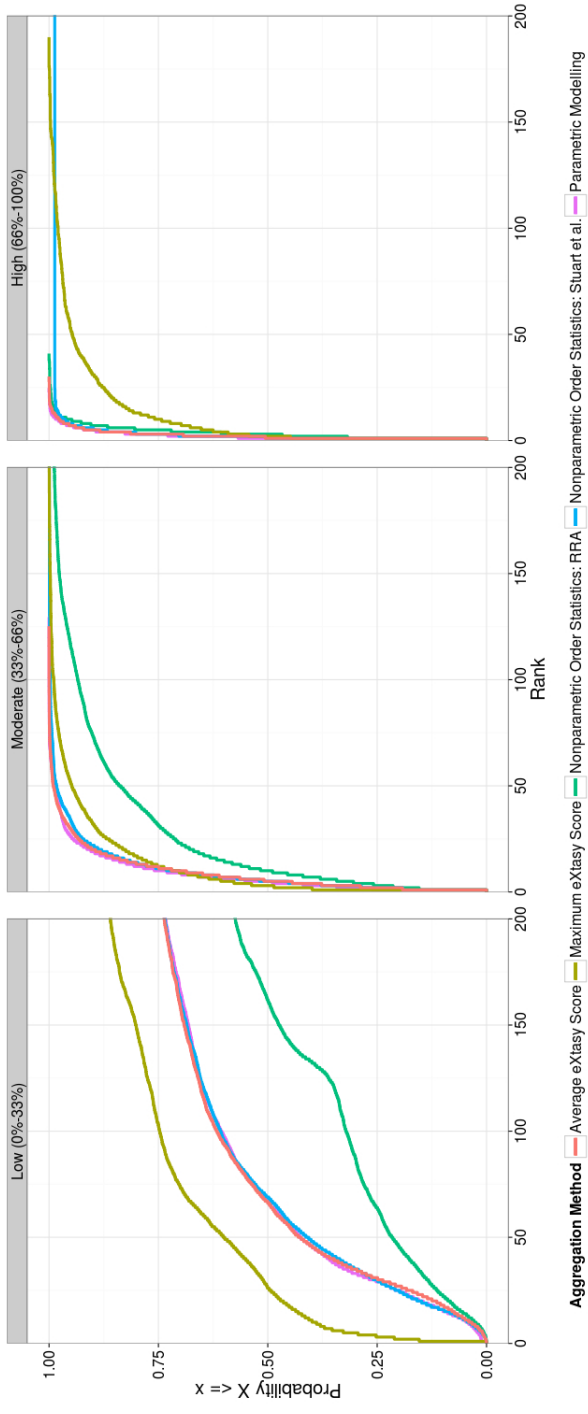


Figure 6.3: Schematic representation of the benchmarking scheme: Nonsynonymous variants are first sampled from 1000 Genomes Project data based on their respective minor allele frequency, generating a synthetic control exome. Then, one disease-causing variant from HGMD is added to a single synthetic control exome and eXtasy prioritizations are performed for each phenotype annotated to the disease-causing variant in that exome. The different phenotype-specific prioritizations are then combined into a single score using different aggregation schemes and the ranks for the injected disease-causing variant is compared across the different methods. This process is repeated for each of the 16272 nonsynonymous HGMD variants.

Table 6.2: Summary statistics for the comparison of different aggregation schemes by fraction of informative phenotypes: Analogous to Table 1, this table shows summary statistics for the ranks of disease-causing mutation but after stratification into 3 classes representing different fractions of informative phenotypes: low (0%-33%), moderate (33%-66%) and high (66%-100%) number of informative phenotypes showing an eXtasy score over 0.5 for the prioritized disease-causing variant.

Class	Aggregation Method	Q1	Median	Mean	Q3
Low(0%-33%)	Average eXtasy Score	31	66	152	218
	Maximum eXtasy Score	3	26	95	102
	Nonparametric Order Statistics: Stuart et al.	30	69	158	224
	Nonparametric Order Statistics: RRA	64	162	243	347
	Parametric Modelling	29	67	295	222
Moderate(33%-66%)	Average eXtasy Score	2	5	8	12
	Maximum eXtasy Score	1	3	11	12
	Nonparametric Order Statistics: Stuart et al.	2	5	10	11
	Nonparametric Order Statistics: RRA	3	10	26	31
	Parametric Modelling	1	5	8	10
High(66%-100%)	Average eXtasy Score	1	1	2	3
	Maximum eXtasy Score	1	2	11	10
	Nonparametric Order Statistics: Stuart et al.	1	1	11	3
	Nonparametric Order Statistics: RRA	1	3	3	5
	Parametric Modelling	1	1	2	3

Discussion

In this study we investigate the performance of different aggregation schemes when dealing with multiphenotype exomic variant prioritization. We benchmark these methods mimicking a real-life application by generating synthetic healthy exomes and injecting single known disease-causing nonsynonymous mutations from HGMD. By comparing different aggregation schemes, ranging from simple metrics to more involved nonparametric and parametric methods, we show that no single method is optimal for all cases and that performance depends on the expected fraction of informative phenotypes.

Not unsurprisingly, the optimal aggregation method is dependent on the signal-to-noise ratio originating from the underlying individual prioritizations. Any prioritization for a specific phenotype can add noise to the global prioritization for two reasons. First, the specific phenotype might not be biologically well-suited to describe the disease under study. This can for example be attributed to phenotypic variability of the disease (e.g. some of the phenotypes might not always be present in all patients showing the disease). This problem can relatively easily be managed by clinical knowledge of the disease, and thus be avoided in real-life applications. In our benchmark this is more difficult to address because of the sheer number of included diseases/phenotypes where disease terms are linked to phenotypic ontologies automatically. Secondly, the gene prioritization step can be noisy because of the selected set of training genes. For example, certain phenotypes could be overly general (*e.g.* intellectual disability, abnormality of the face) and have many heterogeneous underlying molecular mechanisms. Using all known genes in the gene prioritization step for such phenotypes results in uninformative prioritizations because of a dilution of the signal where similarity of a candidate is estimated against an extremely diverse group of genes and thus leads to a reduced performance. The inverse problem also exists where too little is known about the phenotype, biasing the prioritization towards certain training genes that might not necessarily be similar to the as-of-yet undiscovered causative gene. Predicting *a priori* the performance of the gene prioritization is difficult and requires a good understanding of the underlying data being used.

Because of these problems, it is important to use aggregation methods that can handle these uninformative or erroneous prioritizations robustly. Aggregating using the maximum score is a simple and intuitive robust solution and performs well in our previous and current benchmark, especially in low signal-to-noise situations. The drawback of using the maximum is that it ignores information from additional phenotypes, which could be quite powerful in increasing the performance of a global prioritization. This is indicated by the better performance of methods that take advantage of these phenotypes (mean,

order statistics and parametric modelling) when better signal-to-noise ratios are observed. It is important to remark that uninformative phenotypes are likely overrepresented in our benchmark, compared to real-life applications, because of the automatic mappings generated between disease-variants and phenotypic ontologies and the lack of clinical expertise for each evaluated disease. Considering this, it is likely that using these methods will result in better variant prioritizations if the phenotypes are carefully selected.

Here we show that choosing a suitable aggregation method in multiphenotype variant prioritization can affect the results of the resulting global prioritization. We demonstrate this by using simple aggregation methods such as the maximum score and the mean score. Additionally we show that more involved nonparametric and parametric statistical methods offer similar performance with the added benefit of providing estimates of significance which could in future studies could easily be used in broader statistical frameworks such as those used in weighted association or familial studies [135, 128]. We do remark that these methods assume independence of the individual prioritizations, which is an assumption that might not hold fully because of different phenotypes associated to the same disease might share similar molecular mechanisms and thus their ranking of variant might be correlated. This might not greatly influence the performance in prioritizing variants but result in biased estimates of significance. Several methods have been proposed which handle these dependencies [175, 176] but were not considered in this study because of the unavailability of usable implementations, but should be considered when true statistical significance is of importance.

Although currently we rely on users to select an appropriate aggregation method given their experience on how well-characterised the phenotype under study is, in future work we hope to partially or fully automate this process by trying to evaluate training gene set heterogeneity and gene prioritization performance through cross-validation. By gradually improving variant prioritization performance and usability we hope to aid clinical geneticists in discovering disease-causing mutations for their respective phenotypes of interest.

We currently offer all the proposed aggregation schemes in the current version of eXtasy (`http://homes.esat.kuleuven.be/~bioiuser/eXtasy/`). We also publicly provide the benchmark data for further development or benchmarking of other score/rank aggregation methodology (`http://homes.esat.kuleuven.be/~bioiuser/eXtasy/aggregation.tar.gz`).

Chapter 7

Conclusion

> Talk is cheap, show me the code.
>
> Linus Torvalds

The ability to identify rare variation in the human genome through the use of massively parallel sequencing technologies has greatly improved our ability to discover the cause of rare genomic disorders. A plethora of exome-sequencing studies have been reported, successfully identifying the cause of Mendelian disease in a variety of different inheritance models. These studies require advanced computational methodology in order to analyse the great amount of data produced by current sequencing technologies. As research moves further into more complex (e.g. oligogenic/polygenic) disease models and the larger sample sizes needed to identify their genetic causes, more sophisticated algorithms will be needed to manage and interpret the resulting data.

In this thesis we describe a successful application of exome-sequencing (Chapter 3), identifying the genetic cause of Nicolaides-Baraitser syndrome. We model many of the faced data analysis challenges (e.g. data management, cross-sample querying, annotation, data aggregation) into an easy-to-use application called *Annotate-it*(Chapter 4), targetted at the clinical geneticist. We then describe the first phenotype-specific variant prioritization methodology (in a tool called *eXtasy*), which combines the information aggregated by tools like *Annotate-it* by using a Random Forest machine learning approach (Chapters 5 and 6).

Before the availability of MPS technologies, the study of rare Mendelian disorders was restricted by the availability of knowledge concerning potentially involved genes. This prior knowledge could originate either from large familial pedigrees

and linkage-analysis (contrasting affected individuals against non-affected family members) or a deep understanding of the involved biological pathways resulting in a hopefully small set of candidate genes. These candidate genes could then be further assayed (e.g. by Sanger sequencing) in a larger cohort. Lack of large familial pedigrees (e.g. sporadic cases) and the shallow functional annotation for a large proportion of genes affects the power of these approaches greatly. With MPS technologies we're able to apply an unbiased assay of all genes, removing the need for such prior information. Interpreting the many thousands of variants in several hundreds to thousands of different genes is no trivial task. Even after rigorous filtering (e.g. minor allele frequency, functional consequence) 400 to 800 heterozygous and 1 to 50 homozygous rare coding variants per individual still remain. As illustrated in Chapter 3, these numbers can be reduced by looking at genes which contain several variants across unrelated individuals with the expected inheritance mode. In the NBS case, only 5 genes remained containing variants in 3 out of the 4 sequenced samples. Manual inspection of the available information on these genes (e.g. literature, expression, conservation, animal knock-out models) can then help to discern the true disease causing gene.

Several aspects of the described methodology make its real-life application not trivial. First, although there usually might be an assumption of the possible inheritance mode, this is usually not strictly known and several inheritance modes need to be considered. These different analyses often need different filtering criteria, for example recessive and compound heterozygous inheritance might allow for variants with higher frequencies in the population. Applying different filtering and querying schemes often requires several Secondly, annotation sources are subject to rapid changes (e.g. Ensembl had over 20 version changes over the course of this PhD) and keeping annotation information consistent across samples is often crucial. Thirdly, due to complexity of the processing pipelines and these rapid annotation changes it is important to keep track of experimental meta-data (e.g. version numbers, sample information) in order to assure reproducibility and to track potential problems. Fourth, manually gathering all relevant information for a variant or a gene can be extremely time-consuming, especially when several dozen genes have to be screened for interesting candidates. Finally, many tools have been developed over the years that tackle one or more of these problems, but few of these tools are accessible to users with limited computational expertise. In Chapter 4 we propose a versatile software tool, called *Annotate-it*, which aims to facilitate many of these hurdles by offering clinical geneticists an easy-to-use tool for annotation, interpretation and data management. Due its centralized software design, its architecture allows for the simultaneous updating of annotation data sources across all uploaded data sets. Its easy-to-use web interface gives users the possibility to quickly and flexibly filter their data according to several different prior hypotheses. After filtering and cross-querying, once a suitable

set of candidate genes is found, *Annotate-it* aggregates information from many different sources (e.g. OMIM, gene expression, biochemical pathways, gene prioritization, text mining, protein complexes, ...) to facilitate the manual review of candidates. Also, *Annotate-it* has built-in data management features allowing it to keep track of used software versions, validation study status and sample meta-data such as HPO-coded phenotypic descriptions. Annotate-it is currently in use in several in-house and external sequencing projects (e.g. [177]).

Exome sequencing study designs as the ones described in Chapter 3 and Chapter 4 have been successful but are not always feasible. Specially when dealing with vary rare syndromes, it can often prove difficult to find several cases with nearly identical phenotypical presentation. Furthermore, when dealing with more complicated inheritance modes (e.g. multi-allelic causation, or mono-allelic causation but different genes in different patients) these simple filtering approaches do not grant enough power to identify the causative gene/genes in small sample sizes. Computational methods have been previously described to predict the functional impact of mutation, potentially allowing us to reduce the search space by restricting analysis only to mutations predicted to be damaging and thus increasing the power of exome-sequencing studies. These deleteriousness prediction methods however suffer from a high false-positive rate. In Chapters 5 and 6 we propose a phenotype-specific variant prioritization method called *eXtasy*. Integration of phenotype-specific gene prioritization, haploinsufficiency predictions and classical deleteriousness prediction and conservation scores is done by the use of a Random Forest machine learning model. By doing so we can effectively distinguish rare phenotype-unrelated variation from true phenotype-causing variants, increasing the power of our method substantially over existing methods. Applications of our method in child-parent trio studies blinded for parental information shows that top-ranking mutations are significantly enriched (17% of variants remaining after filtering on eXtasy score) for *de novo* mutations (Unpublished Results), highlighting the power of the approach even when no parental information is available.

Although the approaches that we describe in this thesis could facilitate current exome-sequencing studies, many avenues can still be taken in order to enhance the current state of genomic variant interpretation. In the following sections we discuss several conceptual as well as technical improvements that could further benefit exome-sequencing studies of rare Mendelian disease.

7.1 Conceptual improvements

7.1.1 Annotate-it

Federated Design

Due to the potentially sensitive nature of genomic information, institutional policy commonly dictates strict rules on genomic data security. This view is reinforced by recent attacks on data anonymity have shown the successful identification of individuals partaking in genomic studies. Therefore, clinical centers might not be reluctant to upload whole-exome sequencing data to a third party data warehouse. Currently *Annotate-it*'s architecture is based on a centralized database linking annotation information to sample genotype data. By switching the data model to a server-client architecture, where the client stores all genotype information and only pulls annotation information from a centralized server, institutions would not have to migrate genotype information outside of secure zones. This design could maintain the benefits of a centralized annotation repository (as opposed to file-centric annotation), but would impose an additional maintenance cost due to need for additional data version tracking and specialized data query and transfer protocols. An additional benefit of such a federated design would be the possibility to perform cross-center queries without the need for migrating potentially large amounts of data. Such queries would be limited in scope (e.g. only few positions over a given time, anonymized and/or summarized results) in order to prevent identity attacks. A use case for such a system could be the search for additional hits in a potentially interesting candidate gene, or the determination of minor allele frequency of a given variant over a large cross-center cohort. After a match is found, both parties can come to a formal agreement to further share meta-data such as phenotype, experiment meta-data and/or raw data under appropriate consent agreements.

Indels

Annotate-it currently only stores single-nucleotide substitution information. Expanding its capabilities to handle small insertions and deletions would widen its scope thus ameliorating analysis. Although certain conventions exist concerning indel calling (e.g. left-alignment when several different calls are possible), variability between different indel callers makes matching indels across data sets and annotation sources a non-trivial task. An alternative approach to exact matching (ie.e. identical chromosome,

position, alternative allele) is relaxing the matching criteria to find a similar indel (i.e. insertion, deletion) within a certain window (e.g. 10bp). This windowed approach is prone to falsely matching non-identical indels, especially in highly repetitive regions of the genome. A thorough analysis would be needed in order to determine optimal window-sizes and allele similarities. An additional problem would be the translation of complex genomic rearrangements into a useful data format which can be implemented and queried in the database. Finally, although frameshift variants are believed to be extremely disrupting for the protein, the functional interpretation of inframe indel events remains a largely open problem.

Automated updating

In Chapter 4 we discussed the benefits of centralizing annotation information as it allows straightforward updating of all sample data simultaneously, as opposed to denormalized file-centric annotation. Due to the emergence of new available data sources and the constantly evolving existing data sources, the maintenance of such software architecture becomes non-trivial. This is especially the case with academically developed tools, which are primarily developed as proof-of-concept prototypes and where resources and staff are subject to frequent change due to their reliance on temporary funding. One solution to this problem would be the automation of data updates through web service protocols. Many of the larger data providers (e.g. Ensembl, UCSC) currently provide data exchange through the Distributed Annotation System (DAS) protocol [178], which could potentially be used to retrieve new annotation information automatically. However, DAS is designed for small queries and is not effective for the large scale information transfer needed by tools such as *Annotate-it*. Another option would be to use Semantic Web technologies where RDF tuple stores containing annotation information could be queried using a distributed query protocol (e.g. SPARQL). This would require proper standards in the scientific community for data dissemination using such protocols. This would allow the bioinformatics community to develop tools with higher longevity without the need to necessarily allocate additional resources to update data sources.

7.1.2 Variant Prioritization

Variant consequences

The variant prioritization scheme discussed in Chapter 5 shows a proof-of-concept for single nucleotide substitutions which result in amino-acid changes. These variants were specifically chosen as most of the

computational deleteriousness prediction tools focus on this class of variants. Expanding eXtasy to other variant types would significantly widen its scope and thus increase its usefulness in exome-sequencing studies. In order to achieve this, two components are needed for each consequence class: a suitable training set of variants of that class (known to be involved in disease) and a method to estimate its functional impact on the protein at hand. Some classes of variants are more elusive in these respects compared to others. For example, many examples of truncating mutations loss-of-function (e.g. nonsense and frameshift) variants have been reported to cause Mendelian disease and recent studies have painted the landscape of potentially neutral variants of this type [19, 179]. Simple metrics (e.g. relative distance to protein-end) or more involved prediction scores [180] can be used to assess their potential impact on the protein, making them straightforward mutation classes for expanding eXtasy. Variants which do not change the amino-acid sequence are much less well understood, and thus more difficult to interpret. For example, projects such as ENCODE [4] have sparked interest in functional non-coding variation and the first deleteriousness prediction method has just recently been reported [181]. Also synonymous variants, which are often regarded as the hallmark of neutral genomic variation, have been reported to sometimes lead to disease by affecting protein translation kinetics (leading to misfolding) and alternative splicing [12, 14, 13]. A computational prediction method to assess the functional impact of synonymous variants has also recently been reported [182]. One of the main problems with these classes of variants is the availability of large enough sets of disease causing variants of these types (i.e. because they are much more difficult to interpret) as to train classifiers and perform adequate benchmarking. One way to circumvent this problem is to train a single model for the different variant types given diverse genomic and evolutionary information, such as the recently published CADD-score [183]. Nevertheless, the phenotype-specific variant prioritization methodology proposed in this thesis can easily be adapted to other variant types given a reasonable well-performing method to assess their respective deleteriousness.

Model expansions

Currently, eXtasy integrates information concerning variant deleteriousness, haploinsufficiency and phenotype-specific gene prioritization. Although this model performs relatively well compared to classical tools, there is still room for improvement. One way of improving the score is by either adding additional complementary information. For example, shortly after publishing *eXtasy*, Robinson *et al.* proposed a phenotype-specific variant prioritization methodology, called *PHIVE*, using the phenotypic similarity between animal knock-out models and the the phenotype of

interest [169]. This approach is fully complementary to the approach proposed in *eXtasy* and both approaches could straightforwardly be integrated in order to increase power.

Another possible expansion to the current would be the integration of the inheritance mode (i.e. recessive, dominant) and/or mode of action (i.e. loss-of-function, gain-of-function, dominant negative). This follows the assumption that variants acting by different mechanisms (and thus following different inheritance patterns) possess different characteristics. For example, if gain-of-function mutations clustered around a specific locus in a gene are known to cause a specific phenotype, mutations near that locus could be prioritized more highly than mutations elsewhere in that gene. Currently we model these characteristics in implicitly by integrating haploinsufficiency scores. We chose this approach as high-quality information concerning inheritance mode was lacking from HGMD and because this information is often ambiguously represented in OMIM, complicating the generation of a reliable training set. Recent endeavours by the DECIPHER consortium to generate a manually curated database of genes causing Mendelian disorders (DDG2P) [184] including their respective mode of inheritance could pave the way to reliable integration of this information into variant priorization approaches. Which additional annotations are needed, or what the best possible training approach is (e.g. individual models for each specific mode of action or a single unified model) remain open questions and should be evaluated once suitable data sets have been generated.

Hierarchically unbiased training

Although flexible machine learning models, such as the Random Forest model used in eXtasy, have great capacity for capturing complex data patterns, they can suffer from high variance as they can easily overfit peculiarities in the training set. This is especially the case when dealing with data which is hierarchically organized (e.g. variants in genes, genes in chromosomes). Whenever a specific feature assumes identical values for all data-points of a higher-level granularity (e.g. all variants in the same gene have the same haploinsufficiency score), the feature's value can seem correlated with the response variable whereas in reality it is only correlated to the data's structure. If not accounted for, this phenomenon leads to overestimates of performance as learning to recognize genes based on the feature-space does not improve performance when variants are found in genes not present in the training set. In Chapter 5 we tackle this problem by stratifying training and testing sets as to not include overlapping genes, therefore any gene structures learned in training will have no effect on performance on testing sets. However, this problem not only affects estimates of performance but it inherently affects the training process

itself of any ensemble classifier trained by a classical bootstrap approach as this phenomenon will bias the individual classifiers. In an additional recently published study [185], we present an alternative hierarchical sampling approach effectively dealing with hierarchical structure bias when training ensemble classifiers. We show that this approach leads to a subtle drop in sensitivity while effectively doubling the algorithms precision. Implementing such an approach in the eXtasy tool will improve its performance substantially potentially resulting in halve the amount of false positive calls.

Phenotype selection and aggregation

One of the hallmarks of eXtasy is its use of phenotypic information. Yet selection of phenotypes suitable for variant prioritization is not only determined by their clinical presence. The successful application of the method depends on their being sufficient prior knowledge (e.g. phenotype-associated genes) in order to determine meaningful similarities. As a rule of thumb, for a phenotype to be considered for variant prioritization at least three phenotype-associated genes need to be known. A possible solution to this problem would be to propagate genes associated to higher-level HPO terms down the HPO ontology graph to more specific ontology terms. The extend and benefits of this propagation would need to be benchmarked as this approach will potentially also decrease in the signal-to-noise ratio when considering distant nodes on the HPO graph. Inversely, broad phenotypic categories (i.e. mental retardation), where many different genes are known lead to a diluted signal due to the computation of similarities to a heterogeneous set of genes. In this case, one possible approach would be to cluster the heterogeneous set of genes on different criteria (i.e. Endeavour's data sources) to obtain homogeneous subsets of training genes and then perform independent variant prioritizations on each subset.

Once variant prioritizations have been run for multiple selected phenotypes, it is often desirable to have one global score aggregating these independent prioritizations. In Chapter 6 we discuss several possible aggregation approaches and show that the optimal method depends on the fraction of informative phenotypes. Because this fraction is difficult to estimate *a priori*, it is often difficult to choose the optimal aggregation method for the phenotypes at hand. Estimating informativeness for each phenotype, through leave-one-out cross-validation of known phenotype-associated genes, could potentially aid in making an informed choice of aggregation method.

Tissue-specificity

Coding variants are frequently studied in the context of their respective

gene or their representative transcript but are rarely viewed with respect to all possible splicing products. Given that some splicing products might show a particular specificity for a given tissue or developmental stage, we hypothesize that the phenotypic impact might differ with respect to the consequence for each splicing product. Given a specific clinical presentation where several specific tissues are affected (e.g. heart defects), one could use this information to better prioritize likely causal variants. In order to achieve this, a tissue-wide map of alternative splicing is needed. Such an effort is currently being undertaken by the GTex-project [186]. GTex (Genotypes-Tissue Expression) is a recent project for the unravelling of the complex patterns of genetic variation and gene regulation across diverse human tissue types and is still in development. The goal is to create a resource database and associated tissue bank for the scientific community to study the relationship between genetic variation and gene expression in human tissues. By using RNA-Seq, eQTL-analysis and epigenomic assays on different tissues from donors post mortem, a data resource will be created that enables the systematic study of genetic variation and the regulation of gene expression in multiple reference human tissues. This project aims to assay 20000 tissues over 900 individuals through RNAseq by the end of 2015 (although pilot data is becoming available at the time of writing). Integration of this information in a variant prioritization pipeline will likely an extensive data management and statistical framework, usage of a tissue-descriptive ontology (e.g. BRENDA [187]) and a map between this ontology and the HPO.

Model interpretability

Currently eXtasy relies on a Random Forest model to prioritize disease-causing variants. Although this model outperforms other approaches, the drawback of this type of complex machine learning approaches is that the resulting model functions as a black box due to its highly dimensional decision boundary. Gaining insights into the decision-making process would greatly benefit the user when deciding to follow up a prediction made by eXtasy. Permutation-based approaches have been proposed in order to assess the relative importance of input features, but even these approaches fail to give insights into the model when features are intercorrelated (as is the case for many features in eXtasy). In a recent follow-up study [188] we proposed a methodology to obtain the same decision boundary as the Random Forest model utilizing a simpler model which can be readily interpreted. This is achieved by permuting the original training data set over all features to obtain data points describing the full input feature space. These simulated data points were then relabeled using by classifying them with the original Random Forest model. These relabelled permuted data set is then used to train a simple decision tree model

which is grown to overfit the simulated data set, effectively modelling the decision boundary of the original Random Forest model. By examining this tree we can indeed confirm that most of eXtasy's discriminatory power comes from a combination of gene prioritization information as well as deleteriousness prediction estimation. Although these kind of approaches improve transparency of the general decision making process, they do not offer insights at the level of individual predictions. Further work using interactive visualization of the machine learning model [189] could help in achieving more interpretable prioritization results.

Chapter 8

Postscript

If I'm doing an experiment to save the world, I better use my prior.

Andrew Gelman

Since leaving the KU Leuven and finishing this manuscript a little over two years have passed. Given that the field has evolved dramatically over this time, this short postscript aims to give a broad overview of the main developments in the field over the last two years (2014-2016) in relation to this thesis.

Decreasing costs of sequencing have paved the way for large-scale sequencing projects such as Deciphering Developmental Disorders (DDD) study [190, 191] and other studies in schizophrenia [81], autism [192] and intellectual disability [193]. These studies shift the experimental design from a phenotype-driven approach where cases are selected based on few, narrow, well-defined phenotypes, to a genotype-driven approach where large numbers of samples with broad phenotypic characteristics are analysed. By using a refined tri-nucleotide background mutation model [194], one is able to compute the probability of observing a certain enrichment (e.g. in a certain gene, or across gene sets) of *de novo* mutations compared the null hypothesis of random mutation. In this type of genotype-driven experimental design, if cases that share mutations in the same gene show more phenotypic similarity than expected by chance, the confidence in the gene-phenotype relationship is increased [195]. This shifts the use of phenotypic information from an a priori assumption to an a posteriori analysis, allowing the discovery of genes which cause broader phenotypic spectra and gaining insights into the genetic architecture and prevalence of the disease [196].

These large-scale studies can however be complemented by small or medium-scale studies focussing on a select phenotypic presentation. Several improvements have been made that facilitate the analysis of these type studies. First, the availability of much larger population variation databases such as that Exome Aggregation Consortium [197] provide information on over 60 000 presumably healthy individuals. As these control populations continue to grow the mutation-space of the human exome will start to saturate, indicating regions of the exome with varying degrees of tolerance for certain types of mutations. Secondly, the surge of global matchmaking initiatives [198, 199] facilitate finding additional cases with variants in candidate genes of interest. These collaborative platforms facilitate concerted efforts to increase the number of affected individuals carrying similar mutations in particular genes, effectively increasing confidence in genes with a reproducible clinical presentation. Thirdly, the advent of CRISPR/CAS9 genome editing biotechnology [200] have immensely benefited cellular and organismal disease modelling. These "knock-in" approaches tackle many of the criticisms associated with previously used "knock-down" models [201] and allow the modelling of exact allelic configurations observed in patients.

Even considering the lower number of candidate variants after using larger control populations, methods are still needed to prioritize candidate genes for matchmaking and/or functional studies. After eXtasy's publication, several

alternative variant prioritization methodologies have been proposed such as: Phevor [202], Phen-Gen [203], Exomiser [204] and PhenIX [205]. The relative performance of these tools was evaluated by Smedley and Robinson [206], showing that most tools perform considerably well (i.e. ranking the majority of disease causing variants in the top 50 variants), with newer methods outperforming eXtasy. However differences in performance depend on the exact benchmark and how the tools were trained (e.g. poor performance of Phen-Gen on non-1000Genomes data) or their specific use-case (e.g. PhenIX is tailored towards prioritizing known human disease-causing genes linked to the phenotype instead of novel associations). High performances notwithstanding, variant prioritization methods have been rarely used in real-life gene discovery applications due to the increased power offered by larger control populations. However, as the field moves towards from *de novo* and very rare variation to more common variation with patterns of reduced penetrance, new methods likely based on these first approaches will be needed to reduce candidate variant lists.

During the course of the last 6 years, clinical genetics studies have gone from gene-panel sequencing experiments in dozens of cases to tens of thousands of whole-exome trios, and will soon move into the realm of hundreds of thousands of whole-genome sequences and millions of whole-exome sequences in the next few years. These dramatic shifts in the field redefine the biological questions that we will be able to address such as polygenic risk, gene-environment interactions, the role of non-coding variation, epigenetics, mosaicism, etc. To tackle these questions and to deal with data sets on an unprecedented scale (e.g. the UK 100k Genomes Project will generate roughly 20 petabytes of raw sequence data by 2017) we will need new computational methodologies and infrastructures to solve these novel challenges.

Bibliography

[1] O. T. Avery, C. M. Macleod, and M. McCarty. "Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii." *The Journal of experimental medicine*, 79(2):pages 137–58 (1944). pages 6

[2] J. D. Watson and F. H. Crick. "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." *Nature*, 171(4356):pages 737–8 (1953). pages 6

[3] F. H. Crick. "The complementary structure of dna." *Proceedings of the National Academy of Sciences of the United States of America*, 40(8):pages 756–8 (1954). pages 6

[4] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter *et al.* "An integrated encyclopedia of DNA elements in the human genome." *Nature*, 489(7414):pages 57–74 (2012). pages 7, 10, 19, 124

[5] F. Sanger, S. Nicklen, and A. A. Coulson. "DNA sequencing with chain-terminating inhibitors." *Proceedings of the National Academy of Sciences*, 74(12):page 5463 (1977). pages 7

[6] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn *et al.* "Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction." *Cold Spring Harbor symposia on quantitative biology*, 51 Pt 1:pages 263–73 (1986). pages 7

[7] M. O'Connor, M. Peifer, and W. Bender. "Construction of large DNA segments in Escherichia coli." *Science (New York, N.Y.)*, 244(4910):pages 1307–12 (1989). pages 7

[8] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural *et al.* "The sequence of the human genome." *Science*, 291(5507):pages 1304–51 (2001). pages 7

[9] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.* "Initial sequencing and analysis of the human genome." *Nature*, 409(6822):pages 860–921 (2001). pages 7

[10] M. W. Nachman and S. L. Crowell. "Estimate of the Mutation Rate per Nucleotide in Humans." *Genetics*, 156(1):pages 297–304 (2000). pages 7, 22

[11] A. B. Stergachis, E. Haugen, A. Shafer, W. Fu, B. Vernot *et al.* "Exonic Transcription Factor Binding Directs Codon Choice and Affects Protein Evolution." *Science*, 342(6164):pages 1367–1372 (2013). pages 9

[12] C.-J. Tsai, Z. E. Sauna, C. Kimchi-Sarfaty, S. V. Ambudkar, M. M. Gottesman *et al.* "Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima." *Journal of molecular biology*, 383(2):pages 281–91 (2008). pages 9, 124

[13] J. B. Plotkin and G. Kudla. "Synonymous but not the same: the causes and consequences of codon bias." *Nature reviews. Genetics*, 12(1):pages 32–42 (2011). pages 9, 124

[14] Z. E. Sauna and C. Kimchi-Sarfaty. "Understanding the contribution of synonymous mutations to human disease." *Nature reviews. Genetics*, advance on (2011). pages 9, 124

[15] T. R. Mercer, M. E. Dinger, and J. S. Mattick. "Long non-coding RNAs: insights into functions." *Nature reviews. Genetics*, 10(3):pages 155–9 (2009). pages 10

[16] R. M. Durbin, D. L. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley *et al.* "A map of human genome variation from population-scale sequencing." *Nature*, 467(7319):pages 1061–1073 (2010). pages 11, 47

[17] G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin *et al.* "An integrated map of genetic variation from 1,092 human genomes." *Nature*, 491(7422):pages 56–65 (2012). pages 11, 20

[18] J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny *et al.* "Evolution and functional impact of rare coding variation from deep sequencing of human exomes." *Science*, 337(6090):pages 64–9 (2012). pages 11, 20, 21, 70, 106, 111

[19] D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris *et al.* "A systematic survey of loss-of-function variants in human protein-coding genes." *Science (New York, N.Y.)*, 335(6070):pages 823–8 (2012). pages 11, 19, 20, 124

[20] a. Goriely, G. a. McVean, M. Rojmyr, B. Ingemarsson, and a. O. Wilkie. "Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line." *Science*, 301(1095-9203 (Electronic)):pages 643–646 (2003). pages 12

[21] A. C. Allison. "Polymorphism and Natural Selection in Human Populations." *Cold Spring Harbor Symposia on Quantitative Biology*, 29(0):pages 137–149 (1964). pages 12

[22] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little *et al.* "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." *Nature Reviews Genetics*, 9(5):pages 356–369 (2008). pages 12

[23] S. Avigad, S. Kleiman, M. Weinstein, B. E. Cohen, G. Schwartz *et al.* "Compound heterozygosity in nonphenylketonuria hyperphenylalanemia: the contribution of mutations for classical phenylketonuria." *American journal of human genetics*, 49(2):pages 393–9 (1991). pages 12

[24] K. Ohno and K. Suzuki. "Multiple abnormal beta-hexosaminidase alpha chain mRNAs in a compound-heterozygous Ashkenazi Jewish patient with Tay-Sachs disease." *The Journal of biological chemistry*, 263(34):pages 18563–7 (1988). pages 12

[25] C. E. Pearson, K. Nichol Edamura, and J. D. Cleary. "Repeat instability: mechanisms of dynamic mutations." *Nature reviews. Genetics*, 6(10):pages 729–42 (2005). pages 14

[26] E. Buratti, A. Brindisi, F. Pagani, and F. E. Baralle. "Nuclear Factor TDP-43 Binds to the Polymorphic TG Repeats in CFTR Intron 8 and Causes Skipping of Exon 9: A Functional Link with Disease Penetrance." (2004). pages 14

[27] C. P. Austin, J. F. Battey, A. Bradley, M. Bucan, M. Capecchi *et al.* "The knockout mouse project." *Nature genetics*, 36(9):pages 921–4 (2004). pages 14

[28] J. Auwerx, P. Avner, R. Baldock, A. Ballabio, R. Balling *et al.* "The European dimension for the mouse genome mutagenesis program." *Nature genetics*, 36(9):pages 925–7 (2004). pages 14

[29] F. T. Merkle and K. Eggan. "Modeling human disease with pluripotent stem cells: from genome association to function." *Cell stem cell*, 12(6):pages 656–68 (2013). pages 14

[30] D. Dressman, H. Yan, G. Traverso, K. W. Kinzler, and B. Vogelstein. "Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations." *Proceedings of the National Academy of Sciences of the United States of America*, 100:pages 8817–8822 (2003). pages 16

[31] M. Fedurco, A. Romieu, S. Williams, I. Lawrence, and G. Turcatti. "BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies." *Nucleic acids research*, 34:page e22 (2006). pages 16

[32] J. H. Leamon, W. L. Lee, K. R. Tartaro, J. R. Lanza, G. J. Sarkis *et al.* "A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions." *Electrophoresis*, 24:pages 3769–3777 (2003). pages 16

[33] M. L. Metzker. "Emerging technologies in DNA sequencing." *Genome research*, 15:pages 1767–1776 (2005). pages 16

[34] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade *et al.* "A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning." *Genome research*, 18:pages 1051–1063 (2008). pages 16

[35] M. Margulies, M. Egholm, W. Altman, and S. Attiya. "Genome sequencing in microfabricated high-density picolitre reactors." *Nature* (2005). pages 16

[36] M. L. Metzker. "Sequencing technologies - the next generation." *Nature Reviews Genetics*, 11(1):pages 31–46 (2010). pages 16

[37] E. Hodges, Z. Xuan, V. Balija, and M. Kramer. "Genome-wide in situ exon capture for selective resequencing." *Nature genetics* (2007). pages 17

[38] G. J. Porreca, K. Zhang, J. B. Li, B. Xie, D. Austin *et al.* "Multiplex amplification of large sets of human exons." *Nature methods*, 4:pages 931–936 (2007). pages 17

[39] R. Tewhey, J. B. Warner, M. Nakano, B. Libby, M. Medkova *et al.* "Microdroplet-based PCR enrichment for large-scale targeted sequencing." *Nature biotechnology*, 27:pages 1025–1031 (2009). pages 17

[40] P. Flicek and E. Birney. "Sense from sequence reads: methods for alignment and assembly." *Nature methods*, 6(11 Suppl):pages S6–S12 (2009). pages 17, 47

[41] H. Li and N. Homer. "A survey of sequence alignment algorithms for next-generation sequencing." *Briefings in bioinformatics*, 11(5):pages 473–83 (2010). pages 17

[42] H. Li and R. Durbin. "Fast and accurate long-read alignment with Burrows-Wheeler transform." *Bioinformatics (Oxford, England)*, 26:pages 589–595 (2010). pages 17

[43] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics (Oxford, England)*, 25:pages 1754–1760 (2009). pages 17, 18, 38, 54

[44] B. Langmead and S. L. Salzberg. "Fast gapped-read alignment with Bowtie 2." (2012). pages 17

[45] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.* "The Sequence Alignment/Map format and SAMtools." *Bioinformatics (Oxford, England)*, 25:pages 2078–2079 (2009). pages 17

[46] S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk *et al.* "A survey of tools for variant analysis of next-generation genome sequencing data." *Briefings in bioinformatics*, 15:pages 256–78 (2014). pages 18

[47] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.* "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data." *Genome research*, 20(9):page 1297 (2010). pages 18, 38, 54

[48] C. A. Albers, G. Lunter, D. G. MacArthur, G. McVean, W. H. Ouwehand *et al.* "Dindel: accurate indel calls from short-read data." *Genome research*, 21(6):pages 961–73 (2011). pages 18

[49] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning. "Pindel : a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads." *Bioinformatics*, 25(21):pages 2865–2871 (2009). pages 18

[50] J. a. Neuman, O. Isakov, and N. Shomron. "Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection." *Briefings in bioinformatics*, 14:pages 46–55 (2013). pages 18

[51] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.* "The variant call format and VCFtools." *Bioinformatics (Oxford, England)*, 27:pages 2156–2158 (2011). pages 18

[52] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent *et al.* "Ensembl 2012." *Nucleic acids research*, 40:pages D84–90 (2012). pages 19

[53] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott. "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy." *Nucleic Acids Research*, 40:pages D130–5 (2012). pages 19

[54] P. A. Fujita, B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik *et al.* "The UCSC Genome Browser database: update 2011." *Nucleic acids research*, 39:pages D876–D882 (2011). pages 19

[55] K. Wang, M. Li, and H. Hakonarson. "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." *Nucleic acids research*, 38:page e164 (2010). pages 19

[56] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek *et al.* "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor." *Bioinformatics (Oxford, England)*, 26:pages 2069–2070 (2010). pages 19

[57] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen *et al.* "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." (2012). pages 19

[58] D. McCarthy, P. Humburg, A. Kanapin, M. Rivas, K. Gaulton *et al.* "Choice of transcripts and software has a large effect on variant annotation." *Genome Medicine*, 6(3):page 26 (2014). pages 19

[59] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler *et al.* "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nature genetics*, 25(1):pages 25–9 (2000). pages 19

[60] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn *et al.* "The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease." *American journal of human genetics*, 83(5):pages 610–5 (2008). pages 19, 21, 51, 74

[61] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese *et al.* "Database resources of the National Center for Biotechnology Information." *Nucleic acids research*, 36(Database issue):pages D13–21 (2008). pages 19

[62] K. F. Aoki-Kinoshita and M. Kanehisa. "Gene annotation and pathway mapping in KEGG." *Methods in molecular biology (Clifton, N.J.)*, 396:pages 71–91 (2007). pages 19

[63] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser *et al.* "The Reactome pathway knowledgebase." *Nucleic acids research*, 42:pages D472–7 (2014). pages 19

[64] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge *et al.* "The IntAct molecular interaction database in 2010." *Nucleic acids research*, 38:pages D525–D531 (2010). pages 19

[65] N. Huang, I. Lee, E. M. Marcotte, and M. E. Hurles. "Characterising and predicting haploinsufficiency in the human genome." *PLoS genetics*, 6(10):page e1001154 (2010). pages 19, 50, 70

[66] G. M. Cooper and J. Shendure. "Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data." *Nature reviews. Genetics*, 12(9):pages 628–40 (2011). pages 20

[67] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan *et al.* "dbSNP: the NCBI database of genetic variation." *Nucleic acids research*, 29(1):pages 308–11 (2001). pages 20

[68] G. M. Cooper, D. L. Goode, S. B. Ng, A. Sidow, M. J. Bamshad *et al.* "Single-nucleotide evolutionary constraint scores highlight disease-causing mutations." (2010). pages 20

[69] G. M. Cooper and C. D. Brown. "Qualifying the relationship between sequence conservation and molecular function." *Genome research*, 18:pages 201–205 (2008). pages 20

[70] J. D. McAuliffe, M. I. Jordan, and L. Pachter. "Subtree power analysis and species selection for comparative genomics." *Proceedings of the National Academy of Sciences of the United States of America*, 102:pages 7900–7905 (2005). pages 20

[71] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova *et al.* "A method and server for predicting damaging missense mutations." *Nature methods*, 7(4):pages 248–9 (2010). pages 21, 47, 70, 106

[72] J. M. Schwarz, D. N. Cooper, M. Schuelke, and D. Seelow. "MutationTaster2: mutation prediction for the deep-sequencing age." *Nature methods*, 11(4):pages 361–2 (2014). pages 21

[73] E. A. Stone and A. Sidow. "Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity." *Genome research*, 15:pages 978–986 (2005). pages 21

[74] P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak *et al.* "PANTHER: a library of protein families and subfamilies indexed by function." *Genome research*, 13:pages 2129–2141 (2003). pages 21

[75] C. Ferrer-Costa, J. L. Gelpí, L. Zamakola, I. Parraga, X. de la Cruz *et al.* "PMUT: a web-based tool for the annotation of pathological mutations on proteins." *Bioinformatics (Oxford, England)*, 21:pages 3176–3178 (2005). pages 21

[76] Y. Bromberg and B. Rost. "SNAP: predict effect of non-synonymous polymorphisms on function." *Nucleic acids research*, 35:pages 3823–3835 (2007). pages 21

[77] E. Capriotti, R. Calabrese, and R. Casadio. "Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information." *Bioinformatics (Oxford, England)*, 22:pages 2729–2734 (2006). pages 21

[78] A. González-Pérez and N. López-Bigas. "Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel." *American journal of human genetics*, 88(4):pages 440–9 (2011). pages 21

[79] G. V. Kryukov, L. A. Pennacchio, and S. R. Sunyaev. "Most rare missense alleles are deleterious in humans: implications for complex disease and association studies." *American journal of human genetics*, 80(4):pages 727–39 (2007). pages 21, 106

[80] N. Katsanis. "The oligogenic properties of Bardet-Biedl syndrome." *Human molecular genetics*, 13 Spec No:pages R65–R71 (2004). pages 22

[81] M. Fromer, A. J. Pocklington, D. H. Kavanagh, H. J. Williams, S. Dwyer *et al.* "De novo mutations in schizophrenia implicate synaptic networks." *Nature* (2014). pages 22, 23, 130

[82] S. M. Purcell, J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff *et al.* "A polygenic burden of rare disruptive mutations in schizophrenia." *Nature*, 506(7487):pages 185–90 (2014). pages 22, 23

[83] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond *et al.* "Exome sequencing as a tool for Mendelian disease gene discovery." *Nature reviews. Genetics*, 12(11):pages 745–55 (2011). pages 22

[84] A. Hoischen, B. W. M. van Bon, C. Gilissen, P. Arts, B. van Lier *et al.* "De novo mutations of SETBP1 cause Schinzel-Giedion syndrome." *Nature genetics*, 42(6):pages 483–5 (2010). pages 22, 57, 105

[85] S. B. Ng, A. W. Bigham, K. J. Buckingham, M. C. Hannibal, M. J. McMillin *et al.* "Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome." *Nature Genetics*, 42(9):pages 790–3 (2010). pages 22, 47, 55, 105, 106, 107

[86] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor *et al.* "Exome sequencing identifies the cause of a mendelian disorder." *Nature genetics*, 42(1):pages 30–5 (2010). pages 23, 47, 105

[87] L. E. L. M. Vissers, J. de Ligt, C. Gilissen, I. Janssen, M. Steehouwer *et al.* "A de novo paradigm for mental retardation." *Nature genetics*, 42(12):pages 1109–12 (2010). pages 23, 47, 106

[88] J. A. Veltman and H. G. Brunner. "De novo mutations in human genetic disease." *Nature reviews. Genetics*, 13(8):pages 565–75 (2012). pages 23

[89] N. G. C. Smith and A. Eyre-Walker. "Human disease genes: Patterns and predictions." *Gene*, 318:pages 169–175 (2003). pages 23

[90] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal *et al.* "The human disease network." *Proceedings of the National Academy of Sciences of the United States of America*, 104:pages 8685–8690 (2007). pages 23

[91] G. Jimenez-Sanchez, B. Childs, and D. Valle. "Human disease genes." *Nature*, 409(6822):pages 853–5 (2001). pages 23

[92] L.-C. Tranchevent, F. B. Capdevila, D. Nitsch, B. De Moor, P. De Causmaecker *et al.* "A guide to web tools to prioritize candidate genes." *Briefings in bioinformatics*, 12:pages 22–32 (2011). pages 23

[93] D. Börnigen, L.-C. Tranchevent, F. Bonachela-Capdevila, K. Devriendt, B. De Moor *et al.* "An unbiased evaluation of gene prioritization tools." *Bioinformatics (Oxford, England)*, 28:pages 3081–8 (2012). pages 23

[94] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens *et al.* "Gene prioritization through genomic data fusion." *Nature biotechnology*, 24(5):pages 537–44 (2006). pages 23, 48, 70, 75, 106, 107, 108

[95] L.-C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo *et al.* "ENDEAVOUR update: a web resource for gene prioritization in multiple species." *Nucleic acids research*, 36(Web Server issue):pages W377–84 (2008). pages 23, 106

[96] P. Nicolaides and M. Baraitser. "An unusual syndrome with mental retardation and sparse hair." *Clinical dysmorphology*, 2(3):pages 232–6 (1993). pages 31

[97] S. B. Sousa, O. A. Abdul-Rahman, A. Bottani, V. Cormier-Daire, A. Fryer *et al.* "Nicolaides-Baraitser syndrome: Delineation of the phenotype." *American journal of medical genetics. Part A*, 149A(8):pages 1628–40 (2009). pages 31, 35, 38, 62

[98] A. Flaus, D. M. A. Martin, G. J. Barton, and T. Owen-Hughes. "Identification of multiple distinct Snf2 subfamilies with conserved structural motifs." *Nucleic acids research*, 34(10):pages 2887–905 (2006). pages 34, 35, 37

[99] D. C. Hargreaves and G. R. Crabtree. "ATP-dependent chromatin remodeling: genetics, genomics and mechanisms." *Cell research*, 21(3):pages 396–420 (2011). pages 34, 36

[100] L. Magnani and R. A. Cabot. "Manipulation of SMARCA2 and SMARCA4 transcript levels in porcine embryos differentially alters development and expression of SMARCA1, SOX2, NANOG, and EIF1." *Reproduction (Cambridge, England)*, 137(1):pages 23–33 (2009). pages 34, 35

[101] G. Liu, S. Gramling, D. Munoz, D. Cheng, A. K. Azad *et al.* "Two novel BRM insertion promoter sequence variants are associated with loss of BRM expression and lung cancer risk." *Oncogene*, 30(29):pages 3295–304 (2011). pages 34, 36, 47

[102] D. Reisman, S. Glaros, and E. A. Thompson. "The SWI/SNF complex and cancer." *Oncogene*, 28(14):pages 1653–68 (2009). pages 34, 37

[103] L. A. Christ, C. A. Crowe, M. A. Micale, J. M. Conroy, and S. Schwartz. "Chromosome breakage hotspots and delineation of the critical region for the 9p-deletion syndrome." *American journal of human genetics*, 65(5):pages 1387–95 (1999). pages 35

[104] S. Bultman, T. Gebuhr, D. Yee, C. La Mantia, J. Nicholson *et al.* "A Brg1 null mutation in the mouse reveals functional differences among mammalian SWI/SNF complexes." *Molecular cell*, 6(6):pages 1287–95 (2000). pages 35

[105] P. A. Khavari, C. L. Peterson, J. W. Tamkun, D. B. Mendel, and G. R. Crabtree. "BRG1 contains a conserved domain of the SWI2/SNF2 family necessary for normal mitotic growth and transcription." *Nature*, 366(6451):pages 170–4 (1993). pages 35

[106] E. Richmond and C. L. Peterson. "Functional analysis of the DNA-stimulated ATPase domain of yeast SWI2/SNF2." *Nucleic acids research*, 24(19):pages 3685–92 (1996). pages 35, 37

[107] M. C. Hall and S. W. Matson. "Helicase motifs: the engine that powers DNA unwinding." *Molecular microbiology*, 34(5):pages 867–77 (1999). pages 35

[108] I. L. de La Serna, K. A. Carlson, D. A. Hill, C. J. Guidi, R. O. Stephenson *et al.* "Mammalian SWI-SNF complexes contribute to activation of the hsp70 gene." *Molecular and cellular biology*, 20(8):pages 2839–51 (2000). pages 35

[109] H. S. Subramanya, L. E. Bird, J. A. Brannigan, and D. B. Wigley. "Crystal structure of a DExx box DNA helicase." *Nature*, 384(6607):pages 379–83 (1996). pages 35

[110] H. Dürr, A. Flaus, T. Owen-Hughes, and K.-P. Hopfner. "Snf2 family ATPases and DExx box helicases: differences and unifying concepts from high-resolution crystal structures." *Nucleic acids research*, 34(15):pages 4160–7 (2006). pages 35

[111] C. F. Boerkoel, H. Takashima, J. John, J. Yan, P. Stankiewicz *et al.* "Mutant chromatin remodeling protein SMARCAL1 causes Schimke immuno-osseous dysplasia." *Nature genetics*, 30(2):pages 215–20 (2002). pages 36

[112] L. E. L. M. Vissers, C. M. A. van Ravenswaaij, R. Admiraal, J. A. Hurst, B. B. A. de Vries *et al.* "Mutations in a new member of the chromodomain gene family cause CHARGE syndrome." *Nature genetics*, 36(9):pages 955–7 (2004). pages 36

[113] R. Bajpai, D. A. Chen, A. Rada-Iglesias, J. Zhang, Y. Xiong *et al.* "CHD7 cooperates with PBAF to control multipotent neural crest formation." *Nature*, 463(7283):pages 958–62 (2010). pages 36

[114] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics (Oxford, England)*, 25(14):pages 1754–60 (2009). pages 38

[115] K. Wang, M. Li, and H. Hakonarson. "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." *Nucleic acids research*, 38(16):page e164 (2010). pages 40, 48

[116] S. D. Olabarriaga, T. Glatard, and P. T. de Boer. "A virtual laboratory for medical image analysis." *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 14(4):pages 979–85 (2010). pages 40

[117] X. Liu, X. Jian, and E. Boerwinkle. "dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions." *Human mutation*, 32(8):pages 894–9 (2011). pages 40, 75

[118] D. M. Altshuler, R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner *et al.* "Integrating common and rare genetic variation in diverse human populations." *Nature*, 467(7311):pages 52–58 (2010). pages 47

[119] J. R. Lupski, J. G. Reid, C. Gonzaga-Jauregui, D. Rio Deiros, D. C. Y. Chen *et al.* "Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy." *The New England journal of medicine*, 362(13):pages 1181–91 (2010). pages 47

[120] S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham *et al.* "Targeted capture and massively parallel sequencing of 12 human exomes." *Nature*, 461(7261):pages 272–6 (2009). pages 47

[121] A. V. Postma, K. van Engelen, J. van de Meerakker, T. Rahman, S. Probst *et al.* "Mutations in the sarcomere gene MYH7 in Ebstein anomaly." *Circulation. Cardiovascular genetics*, 4(1):pages 43–50 (2011). pages 47

[122] P. S. Tarpey, R. Smith, E. Pleasance, A. Whibley, S. Edkins *et al.* "A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation." *Nature genetics*, 41(5):pages 535–43 (2009). pages 47

[123] C. Gilissen, H. H. Arts, A. Hoischen, L. Spruijt, D. A. Mans *et al.* "Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome." *American journal of human genetics*, 87(3):pages 418–23 (2010). pages 47

[124] P. Ng and S. Henikoff. "SIFT: Predicting amino acid changes that affect protein function." *Nucleic acids research*, 31(13):page 3812 (2003). pages 47, 70, 106

[125] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau *et al.* "The FoldX web server: an online force field." *Nucleic acids research*, 33(Web Server issue):pages W382–8 (2005). pages 47

[126] M. Cline and R. Karchin. "Using bioinformatics to predict the functional impact of SNVs." *Bioinformatics (Oxford, England)*, pages 1–9 (2010). pages 47

[127] M. C. Lopes, C. Joyce, G. R. Ritchie, S. L. John, F. Cunningham *et al.* "A Combined Functional Annotation Score for Non-Synonymous Variants." *Human Heredity*, 73(1):pages 47–51 (2012). pages 47, 75

[128] I. Ionita-Laza, V. Makarov, S. Yoon, B. Raby, J. Buxbaum *et al.* "Finding disease variants in Mendelian disorders by using sequence data: methods and applications." *American journal of human genetics*, 89(6):pages 701–12 (2011). pages 47, 51, 74, 117

[129] Y. W. Asmann, S. Middha, A. Hossain, S. Baheti, Y. Li *et al.* "TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data." *Bioinformatics (Oxford, England)*, 28(2):pages 277–8 (2012). pages 48

[130] Y.-C. Cheng, F.-C. Hsiao, E.-C. Yeh, W.-J. Lin, C.-Y. L. Tang *et al.* "VarioWatch: providing large-scale and comprehensive annotations on human genomic variants in the next generation sequencing era." *Nucleic acids research* (2012). pages 48

[131] A. C. Shetty, P. Athri, K. Mondal, V. L. Horner, K. M. Steinberg *et al.* "SeqAnt: A web service to rapidly identify and annotate DNA sequence variations." In "BMC bioinformatics," volume 11, page 471 (2010). pages 48

[132] V. Makarov, T. O'Grady, G. Cai, J. Lihm, J. D. Buxbaum *et al.* "AnnTools: a comprehensive and versatile annotation toolkit for genomic variants." *Bioinformatics (Oxford, England)*, 28(5):pages 724–5 (2012). pages 48

[133] J. K. Teer, E. D. Green, J. C. Mullikin, and L. G. Biesecker. "VarSifter: Visualizing and analyzing exome-scale sequence variation data on a desktop computer." *Bioinformatics (Oxford, England)*, 28(4):pages 599–600 (2012). pages 48

[134] M. Sincan, D. Simeonov, D. Adams, T. C. Markello, T. Pierson *et al.* "VAR-MD: A tool to analyze whole exome/genome variants in small human pedigrees with Mendelian inheritance." *Human mutation* (2012). pages 48

[135] M. Yandell, C. Huff, H. Hu, M. Singleton, B. Moore *et al.* "A probabilistic disease-gene finder for personal genomes." *Genome research*, 21(9):pages 1529–42 (2011). pages 48, 74, 117

[136] M.-X. Li, H.-S. Gui, J. S. H. Kwan, S.-Y. Bao, and P. C. Sham. "A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases." *Nucleic acids research*, 40(7):page e53 (2012). pages 48

[137] D. Ge, E. K. Ruzzo, K. V. Shianna, M. He, K. Pelak *et al.* "SVA: software for annotating and visualizing sequenced human genomes." *Bioinformatics (Oxford, England)*, 27(14):pages 1998–2000 (2011). pages 48

[138] S. Van Vooren, B. Thienpont, B. Menten, F. Speleman, B. De Moor *et al.* "Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations." *Nucleic acids research*, 35(8):pages 2533–43 (2007). pages 48, 50

[139] Y. Moreau and L.-C. Tranchevent. "Computational tools for prioritizing candidate genes: boosting disease gene discovery." *Nature reviews. Genetics*, 13(8):pages 523–536 (2012). pages 50, 70, 72, 106

[140] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. "KEGG for integration and interpretation of large-scale molecular data sets." *Nucleic acids research*, 40(1):pages D109–14 (2011). pages 50

[141] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt *et al.* "Reactome: a knowledgebase of biological pathways." *Nucleic acids research*, 33(Database issue):pages D428–32 (2005). pages 50

[142] L. Jensen, M. Kuhn, M. Stark, S. Chaffron, and C. "STRING 8–a global view on proteins and their functional interactions in 630 organisms." *Nucleic Acids Research*, 37(October 2008):pages 412–416 (2009). pages 50

[143] A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach *et al.* "CORUM: the comprehensive resource of mammalian protein complexes–2009." *Nucleic acids research*, 38(Database issue):pages D497–501 (2010). pages 50

[144] Y. Li, N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, T. Jiang *et al.* "Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants." *Nature Genetics*, 42(11):pages 969–72 (2010). pages 50

[145] S. Chun and J. C. Fay. "Identification of deleterious mutations within three human genomes." *Genome research*, 19(9):pages 1553–61 (2009). pages 50, 70

[146] J. M. Schwarz, C. Rödelsperger, M. Schuelke, and D. Seelow. "MutationTaster evaluates disease-causing potential of sequence alterations." *Nature methods*, 7(8):pages 575–6 (2010). pages 50, 70, 106

[147] F. Strozzi and J. Aerts. "A Ruby API to query the Ensembl database for genomic features." *Bioinformatics (Oxford, England)*, 27(7):pages 1013–4 (2011). pages 52

[148] N. Goto, P. Prins, M. Nakao, R. Bonnal, J. Aerts *et al.* "BioRuby: Bioinformatics software for the Ruby programming language." *Bioinformatics*, 26(20):page 2617 (2010). pages 52

[149] P. J. Kersey, D. Lawson, E. Birney, P. S. Derwent, M. Haimel *et al.* "Ensembl Genomes: extending Ensembl across the taxonomic space." *Nucleic acids research*, 38(Database issue):pages D563–9 (2010). pages 52

[150] D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan *et al.* "VarScan: variant detection in massively parallel sequencing of individual and pooled samples." *Bioinformatics (Oxford, England)*, 25(17):pages 2283–5 (2009). pages 54

[151] Y. Shen, Z. Wan, C. Coarfa, R. Drabek, L. Chen *et al.* "A SNP discovery method to assess variant allele probability from next-generation resequencing data." *Genome research*, 20(2):pages 273–80 (2010). pages 54

[152] J. K. J. Van Houdt, B. A. Nowakowska, S. B. Sousa, B. D. C. van Schaik, E. Seuntjens *et al.* "Heterozygous missense mutations in SMARCA2 cause Nicolaides-Baraitser syndrome." *Nature genetics*, 44(4):pages 445–449 (2012). pages 61, 105, 106

[153] G. W. E. Santen, E. Aten, Y. Sun, R. Almomani, C. Gilissen *et al.* "Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome." *Nature genetics*, 44(4):pages 379–80 (2012). pages 61

[154] Y. Tsurusaki, N. Okamoto, H. Ohashi, T. Kosho, Y. Imai *et al.* "Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome." *Nature genetics*, 44(4):pages 376–8 (2012). pages 61

[155] G. Lyon and K. Wang. "Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress." *Genome Medicine* (2012). pages 62

[156] S. Kumar, M. Sanderford, V. E. Gray, J. Ye, and L. Liu. "Evolutionary diagnosis method for variants in personal exomes." *Nature methods*, 9(9):pages 855–6 (2012). pages 70, 106

[157] S. Asthana, W. S. Noble, G. Kryukov, C. E. Grant, S. Sunyaev *et al.* "Widely distributed noncoding purifying selection in the human genome." *Proceedings of the National Academy of Sciences of the United States of America*, 104(30):pages 12410–5 (2007). pages 70

[158] L. Breiman. "Random forests." *Machine learning*, pages 1–33 (2001). pages 70, 78

[159] P. D. Stenson, M. Mort, E. V. Ball, K. Howells, A. D. Phillips *et al.* "The Human Gene Mutation Database: 2008 update." *Genome medicine*, 1(1):page 13 (2009). pages 70, 108

[160] C. L. Myers, D. R. Barrett, M. A. Hibbs, C. Huttenhower, and O. G. Troyanskaya. "Finding function: evaluation methods for functional genomic data." *BMC genomics*, 7:page 187 (2006). pages 72

[161] S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken *et al.* "Clinical diagnostics in human genetics with semantic similarity searches in ontologies." *American journal of human genetics*, 85(4):pages 457–64 (2009). pages 21, 74

[162] M. Pertea, G. M. Pertea, and S. L. Salzberg. "Detection of lineage-specific evolutionary changes among primate species." *BMC bioinformatics*, 12:page 274 (2011). pages 75

[163] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou *et al.* "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." *Genome research*, 15(8):pages 1034–50 (2005). pages 75

[164] M. Vihinen. "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis." *BMC genomics*, 13 Suppl 4:page S2 (2012). pages 75

[165] A. Boulesteix, S. Janitza, J. Kruppa, and I. König. "Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics." (2012). pages 79

[166] M. Choi, U. I. Scholl, W. Ji, T. Liu, I. R. Tikhonova *et al.* "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing." *Proceedings of the National Academy of Sciences of the United States of America*, 106(45):pages 19096–101 (2009). pages 105

[167] C. Gilissen, A. Hoischen, H. G. Brunner, and J. A. Veltman. "Disease gene identification strategies for exome sequencing." *European journal of human genetics : EJHG*, 20(5):pages 490–7 (2012). pages 106

[168] R. M. Durbin, D. L. Altshuler, G. R. Abecasis, D. R. Bentley, A. Chakravarti *et al.* "A map of human genome variation from population-scale sequencing." *Nature*, 467(7319):pages 1061–1073 (2010). pages 106, 111

[169] P. Robinson, S. Köhler, A. Oellrich, K. Wang, C. Mungall *et al.* "Improved exome prioritization of disease genes through cross species phenotype comparison." *Genome research*, pages gr.160325.113– (2013). pages 106, 111, 125

[170] A. Sifrim, D. Popovic, L.-C. Tranchevent, A. Ardeshirdavani, R. Sakai *et al.* "eXtasy: variant prioritization by genomic data fusion." *Nature methods* (2013). pages 106

[171] J. Stuart, E. Segal, D. Koller, and S. Kim. "A gene-coexpression network for global discovery of conserved genetic modules." *Science* (2003). pages 107, 108

[172] R. Kolde, S. Laur, P. Adler, and J. Vilo. "Robust rank aggregation for gene list integration and meta-analysis." *Bioinformatics (Oxford, England)*, 28(4):pages 573–80 (2012). pages 107, 108

[173] R. Fisher. "Statistical methods for research workers." (1925). pages 110

[174] J. R. Lupski, J. G. Reid, C. Gonzaga-jauregui, D. R. Deiros, D. C. Y. Chen *et al.* "Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy." *New England Journal of Medicine*, 362(13):pages 1181–1191 (2010). pages 111

[175] M. Brown. "400: A Method for Combining Non-Independent, One-Sided Tests of Significance." *Biometrics* (1975). pages 117

[176] J. T. Kost and M. P. McDermott. "Combining dependent P-values." (2002). pages 117

[177] P. D. Brady, J. Van Houdt, B. Callewaert, J. Deprest, K. Devriendt *et al.* "Exome sequencing identifies ZFPM2 as a cause of familial isolated congenital diaphragmatic hernia and possibly cardiovascular malformations." *European journal of medical genetics*, 57(6):pages 247–52 (2014). pages 121

[178] R. Dowell, R. Jokerst, A. Day, S. Eddy, and L. Stein. "The Distributed Annotation System." *BMC Bioinformatics*, 2(1):page 7 (2001). pages 123

[179] D. G. MacArthur and C. Tyler-Smith. "Loss-of-function variants in the genomes of healthy humans." *Human Molecular Genetics*, 19 (2010). pages 124

[180] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan. "Predicting the functional effect of amino acid substitutions and indels." *PloS one*, 7(10):page e46688 (2012). pages 124

[181] G. R. S. Ritchie, I. Dunham, E. Zeggini, and P. Flicek. "Functional annotation of noncoding sequence variants." *Nature methods*, 11(3):pages 294–6 (2014). pages 124

[182] O. J. Buske, A. Manickaraj, S. Mital, P. N. Ray, and M. Brudno. "Identification of deleterious synonymous variants in human genomes." *Bioinformatics (Oxford, England)*, 29(15):pages 1843–1850 (2013). pages 124

[183] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper *et al.* "A general framework for estimating the relative pathogenicity of human genetic variants." *Nature Genetics*, advance on (2014). pages 124

[184] E. Bragin, E. A. Chatzimichali, C. F. Wright, M. E. Hurles, H. V. Firth *et al.* "DECIPHER: Database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation." *Nucleic Acids Research*, 42(D1) (2014). pages 125

[185] D. Popovic, A. Sifrim, and J. Davis. "Improving performance of the eXtasy model by hierarchical sampling." *Pattern Recognition in ...* (2014). pages 126

[186] The GTEx Consortium. "The Genotype-Tissue Expression (GTEx) project." *Nature genetics*, 45(6):pages 580–5 (2013). pages 127

[187] M. Gremse, A. Chang, I. Schomburg, A. Grote, M. Scheer *et al.* "The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources." *Nucleic acids research*, 39(Database issue):pages D507–13 (2011). pages 127

[188] D. Popovic, A. Sifrim, J. Davis, Y. Moreau, and B. De Moor. "Problems with the nested granularity of feature domains in bioinformatics: the eXtasy case." *BMC bioinformatics*, 16 Suppl 4:page S2 (2015). pages 127

[189] S. T. Teoh and K.-L. Ma. "PaintingClass." In "Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03," page 667. ACM Press, New York, New York, USA (2003). pages 128

[190] T. W. Fitzgerald, S. S. Gerety, W. D. Jones, M. van Kogelenberg, D. A. King *et al.* "Large-scale discovery of novel genetic causes of developmental disorders." *Nature*, 519(7542):pages 223–228 (2014). pages 130

[191] C. F. Wright, T. W. Fitzgerald, W. D. Jones, S. Clayton, J. F. McRae *et al.* "Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data." *The Lancet*, 385(9975):pages 1305–14 (2014). pages 130

[192] I. Iossifov, B. J. O'Roak, S. J. Sanders, M. Ronemus, N. Krumm *et al.* "The contribution of de novo coding mutations to autism spectrum disorder." *Nature*, 515(7526):pages 216–21 (2014). pages 130

[193] J. de Ligt, M. H. Willemsen, B. W. van Bon, T. Kleefstra, H. G. Yntema *et al.* "Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability." *New England Journal of Medicine*, 367(20):pages 1921–1929 (2012). pages 130

[194] K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo *et al.* "A framework for the interpretation of de novo mutation in human disease." *Nature Genetics*, 46(9):pages 944–950 (2014). pages 130

[195] N. Akawi, J. McRae, M. Ansari, M. Balasubramanian, M. Blyth *et al.* "Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families." *Nature Genetics*, 47(11):pages 1363–1369 (2015). pages 21, 130

[196] J. F. McRae, S. Clayton, T. W. Fitzgerald, J. Kaplanis, E. Prigmore *et al.* "Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation." Technical report (2016). pages 130

[197] E. A. Consortium, M. Lek, K. Karczewski, E. Minikel, K. Samocha *et al.* "Analysis of protein-coding genetic variation in 60,706 humans." Technical report (2015). pages 130

[198] A. A. Philippakis, D. R. Azzariti, S. Beltran, A. J. Brookes, C. A. Brownstein *et al.* "The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery." *Human Mutation*, 36(10):pages 915–921 (2015). pages 130

[199] O. J. Buske, M. Girdea, S. Dumitriu, B. Gallinger, T. Hartley *et al.* "PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases." *Human mutation*, 36(10):pages 931–40 (2015). pages 130

[200] J. D. Sander and J. K. Joung. "CRISPR-Cas systems for editing, regulating and targeting genomes." *Nature biotechnology*, 32(4):pages 347–55 (2014). pages 130

[201] S. Schulte-Merker and D. Y. R. Stainier. "Out with the old, in with the new: reassessing morpholino knockdowns in light of genome editing technology." *Development*, 141:pages 3103–3104 (2014). pages 130

[202] M. V. Singleton, S. L. Guthery, K. V. Voelkerding, K. Chen, B. Kennedy *et al.* "Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families." *American Journal of Human Genetics*, 94(4):pages 599–610 (2014). pages 131

[203] A. Javed, S. Agrawal, and P. C. Ng. "Phen-Gen: combining phenotype and genotype to analyze rare disorders." *Nature Methods*, 11(august):pages 1–7 (2014). pages 131

[204] D. Smedley, J. O. B. Jacobsen, M. Jäger, S. Köhler, M. Holtgrewe *et al.* "Next-generation diagnostics and disease-gene discovery with the Exomiser." *Nature protocols*, 10(12):pages 2004–2015 (2015). pages 131

[205] T. Zemojtel, S. Köhler, L. Mackenroth, M. Jäger, J. Hecht *et al.* "Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome." *Science translational medicine*, 6(252):page 252ra123 (2014). pages 131

[206] D. Smedley and P. N. Robinson. "Phenotype-driven strategies for exome prioritization of human Mendelian disease genes." *Genome Medicine*, 7(1):page 81 (2015). pages 131

Curriculum Vitae

Alejandro Sifrim was born in Buenos Aires, Argentina on June 4th, 1985. In the year 2007 he obtained a Bachelor's degree in Science (major in Biochemistry and Biotechnology, minor in Molecular Biology) from the university K.U. Leuven, Belgium with Cum Laude. In the year 2009 he obtained his Master's degree in Science (major Biochemistry and Biotechnology) from the university K.U. Leuven, Belgium with Magna Cum Laude. He fulfilled his master's dissertation titled "Alternative experimental design with an applied normalization scheme can improve statistical power in 2D-DIGE experiments" under supervision of Prof. dr. ir. Kathleen Marchal and Prof. dr. ir. Lutgarde Arckens. In the year 2009, he was awarded an IWT specialization scholarship and started the PhD program at the Department of Electrical Engineering (SISTA/SCD) at the Bioinformatics group. His research topic is the "Interpretation and prioritization of single nucleotide variation through genomic data fusion" under supervision of Prof. dr. ir Yves Moreau (co-promotors: Prof. dr. ir Joris Vermeesch, Prof. dr. ir. Jan Aerts).

List of publications

Journal papers

- Engelen K.[*], **Sifrim A.**[*], van der Plas B., Laukens K., Arckens L., Marchal K., "Alternative Experimental Design with an Applied Normalization Scheme Can Improve Statistical Power in 2D-DIGE Experiments", *Journal of Proteome Research*, vol. 9, no. 10, pp. 4919-4926 (2010)

- Pavlopoulos G.A., Hooper SD., **Sifrim A.**, Schneider R., Aerts J., "Medusa: A tool for exploring and clustering biological networks.", *BMC Res Notes.* , vol. 4, no. 1, pp. 384-390. (2011)

- Van Houdt J., Nowakowska B., Sousa S., van Schaik B., Seuntjens E., Avonce N., **Sifrim A.**, Abdul-Rahman O., van den Boogaard M., Bottani A., Castori M., Cormier-Daire V., Deardorff M., Filges I., Fryer A., Fryns J., Gana S., Garavelli L., Gillessen-Kaesbach G., Hall B., Horn D., Huylebroeck D., Klapecki J., Krajewska-Walasek M., Kuechler A., Lines M., Maas S., MacDermot K., McKee S., Magee A., de Man S., Moreau Y., Morice-Picard F., Obersztyn E., Pilch J., Rosser E., Shannon N., Stolte-Dijkstra I., Van Dijck P., Vilain C., Vogels A., Wakeling E., Wieczorek D., Wilson L., Zuffardi O., van Kampen A., Devriendt K., Hennekam R., Vermeesch J., "Heterozygous missense mutations in SMARCA2 cause Nicolaides-Baraitser syndrome", *Nature Genetics*, vol. 44, no. 4, pp. 1-6. (2012)

- Popovic D., **Sifrim A.**, Pavlopoulos G.A., Moreau Y., De Moor B., "A Simple Genetic Algorithm for Biomarker Mining", *Lecture Notes in Bioinformatics* (PRIB 2012, Tokyo, Japan), pp. 222-232. (2012)

- **Sifrim A.**, Van Houdt J.K.J, Tranchevent L., Nowakowska B., Sakai R., Pavlopoulos G.A., Devriendt K., Vermeesch J.R., Moreau Y., Aerts J., "Interpretation of single nucleotide variation in human disease : a

Swiss-knife approach to annotation and analysis", *Genome Medicine*, vol. 4, no. 9 p. 73 (2013)

- Pavlopoulos GA, G., Kumar, P., **Sifrim, A.**, Sakai, R., Lin, M., Voet, T., Moreau, Y., Aerts, J. "Meander : visually exploring the structural variome using space-filling curves". *Nucleic acids research*, 41 (11), e118. (2013)

- Sakai, R., **Sifrim, A.**, vande Moere, A., Aerts, J. "TrioVis : a visualization approach for filtering genomic variants of parent-child trios". *Bioinformatics*, 29 (14), 1801-1802 (2013)

- Pavlopoulos GA, Oulas A, Iacucci E, **Sifrim, A.**, Moreau Y, Schneider R, Aerts J, Iliopoulos I. "Unraveling genomic variation from next generation sequencing data". *BioData Mining*, 6:13 (2013)

- **Sifrim A.**[*], Popovic D.[*], Tranchevent L.-C., Ardeshirdavani A., Sakai R., Konings P., Vermeesch J., Aerts J., De Moor B., Moreau Y., "eXtasy: Variant Prioritization By Genomic Data Fusion", *Nature Methods*, vol. 10, no. 11, pp. 1083-1084 (2013)

- Popovic D, **Sifrim A**, Davis J, Moreau Y, De Moor B. "Problems with the nested granularity of feature domains in bioinformatics: the eXtasy case." BMC Bioinformatics 16 Suppl 4: S2. (2015)

- The Deciphering Developmental Disorders Study, "Large-scale discovery of novel genetic causes of developmental disorders." *Nature*, 519: 223–228. (2015)

- Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, Van Kogelenberg M, King DA, Ambridge K, Barrett DM, Bayzetinova T, Bevan P, Bragin E, Chatzimichali EA, Gribble S, Jones P, Krishnappa N, Mason LE, Miller R, Morley KI, Parthiban V, Prigmore E, Rajan D, **Sifrim A**, Swaminathan GJ, Tivey AR, Middleton A, Parker M, Carter NP, Barrett JC, Hurles ME, FitzPatrick DR, Firth HV 2015. "Genetic diagnosis of developmental disorders in the DDD study: A scalable analysis of genome-wide research data." *Lancet*, 385: 1305–1314 (2015)

- Akawi N, McRae J, Ansari M, Balasubramanian M, Blyth M, Brady AF, Clayton S, Cole T, Deshpande C, Fitzgerald TW, Foulds N, Francis R, Gabriel G, Gerety SS, Goodship J, Hobson E, Jones WD, Joss S, King D, Klena N, Kumar A, Lees M, Lelliott C, Lord J, McMullan D, O'Regan M, Osio D, Piombo V, Prigmore E, Rajan D, Rosser E, **Sifrim A**, Smith A, Swaminathan GJ, Turnpenny P, Whitworth J, Wright CF, Firth HV, Barrett JC, Lo CW, FitzPatrick DR, Hurles ME, The Deciphering

Developmental Disorders study. "Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families." *Nature Genetics* 47, 1363–1369 (2015)

- Isrie M, Breuss M, Tian G, Hansen AH, Cristofoli F, Morandell J, Kupchinsky ZA, **Sifrim A**, Rodriguez-Rodriguez CM, Dapena EP, Doonanco K, Leonard N, Tinsa F, Moortgat S, Ulucan H, Koparir E, Karaca E, Katsanis N, Marton V, Vermeesch JR, Davis EE, Cowan NJ, Keays DA, Van Esch H. "Mutations in Either TUBB or MAPRE2 Cause Circumferential Skin Creases Kunze Type." *Am. J. Hum. Genet.* 97, 790–800 (2015)

- ElShal S, Tranchevent L-C, **Sifrim A**, Ardeshirdavani A, Davis J, Moreau Y. "Beegle: from literature mining to disease-gene discovery." *Nucleic Acids Res* 44: e18. (2016)

- **Sifrim A**, Hitz M-P, Wilsdon A, Breckpot J, Al Turki SH, Thienpont B, McRae J, Fitzgerald TW, Singh T, Swaminathan GJ, Prigmore E, Rajan R, Abdul-Khaliq H, Banka S, Bauer UMM, Bentham J, Berger F, Bhattacharya S, Bu'Lock F, Canham N, Colgiu I-G, Cosgrove C, Cox H, Daehnert I, Daly A, Danesh J, Fryer A, Gewillig M, Hobson E, Hoff K, Homfray T, the INTERVAL Study, Kahlert A-K, Ketley A, Kramer H-H, Lachlan K, Lampe AK, Louw JJ, Manickara AK, Manase D, McCarthy KP, Metcalfe K, Moore C, Newbury-Ecob R,Omer SO, Ouwehand WH, Park S-M, Parker MJ, Pickardt T, Pollard MO, Robert L, Roberts DJ, Sambrook J,Setchfield K, Stiller B, Thornborough C, Toka O, Watkins H, Williams D, Wright M, Mital S, Daubeney PEF, Keavney B, Goodship J, the UK10K Consortium, Abu-Sulaiman RM, Klaassen S, Wright CF, Firth HV, Barrett JC, Devriendt K, FitzPatrick DR, Brook JD, the Deciphering Developmental Disorders Study and Hurles ME "Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing" *Nature Genetics*, 48, 1060–1065 (2016)

- The Deciphering Developmental Disorders Study "Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation" *Nature* (in press) (2016)

Pre-prints

- King DA, **Sifrim A**, Fitzgerald TW, Rahbari R, Hobson E, Homfray T, Mansour S, Mehta SG, Shehla M, Tomkins SE, Vasudevan PC, Hurles ME, The Deciphering Developmental Disorders Study "Detection of structural

mosaicism from targeted and whole-genome sequencing data" *bioRxiv*, *doi: http://dx.doi.org/10.1101/062620* (2016)

- **Sifrim A**, Popovic D, Vermeesch JR, Aerts J, De Moor B, Moreau Y "Comparison of aggregation methods for multiphenotype exomic variant prioritization" *bioRxiv*, *doi: http://dx.doi.org/10.1101/064899* (2016)

Conference papers

- Popovic, D., **Sifrim, A.**, Pavlopoulos, G., Moreau, Y., De Moor, B. "A simple genetic algorithm for biomarker mining." Pattern Recognition in Bioinformatics: Vol. 7632. PRIB 2012. Tokyo, Japan, pp. 222-232 (2012).

- Popovic D., **Sifrim A.**, Moschopoulos C., Moreau Y., De Moor B., "A Hybrid Approach to Feature Ranking for Microarray Data Classification", in Engineering Applications of Neural Networks (Communications in Computer and Information Science) Volume 384 (EANN 2013), Chalkidiki, Greece, Sep. 2013, pp. 241-248. (2013)

- Moschopoulos C., Popovic D., **Sifrim A.**, Beligiannis G., De Moor B., Moreau Y., "A genetic algorithm for pancreatic cancer diagnosis", in Engineering Applications of Neural Networks (Communications in Computer and Information Science) 384 (EANN2013), Chalkidiki, Greece, Sep. 2013, pp. 222-230 (2013)

- Popovic, D., **Sifrim, A.**, Moreau, Y., De Moor, B. (2013). "eXtasy simplified - towards opening the black box." Proc. of IEEE International Conference on Bioinformatics and Biomedicine. IEEE BIBM 2013. Shanghai, China, Dec. 2013 (pp. 24-28) (2013)

Awards

- Best Presentation Award for the talk entitled "Computational annotation and interpretation of single-nucleotide variation to identify disease causing variants by next-generation sequencing", *Belgian Society of Human Genetics, 12th Annual Meeting*, 2012, Liege, Belgium

- Best Flash Presentation Award for the talk entitled "Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease", *Benelux Bioinformatics Conference 2012 (BBC12)*, Nijmegem The Netherlands



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING (ESAT)
BIOINFORMATICS GROUP
Kasteelpark Arenberg 10, postbus 2440
B-3001 Heverlee
alejandro.sifrim@kuleuven.be