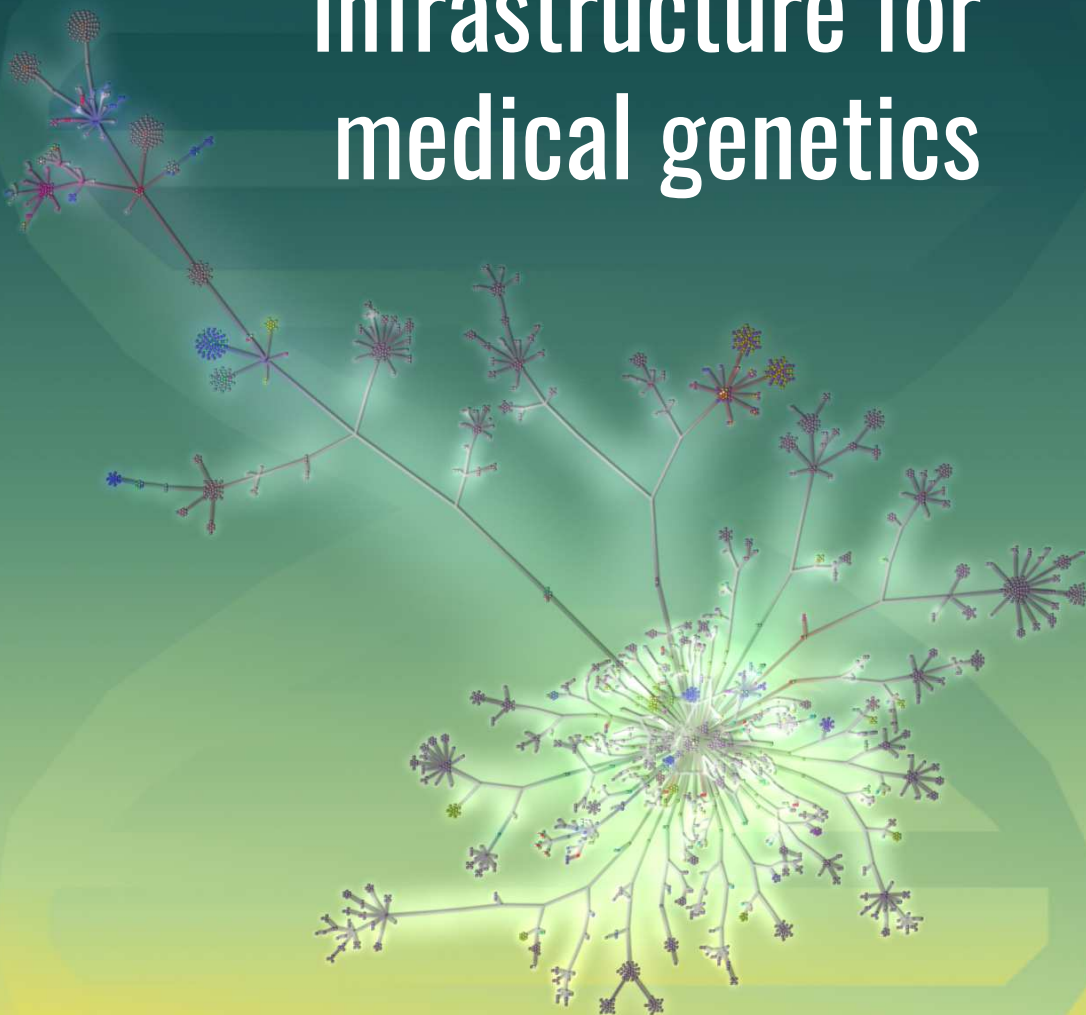


Translational software infrastructure for medical genetics



K. Joeri van der Velde

Translational software infrastructure for medical genetics

K. Joeri van der Velde

Kasper Joeri van der Velde. **Translational software infrastructure for medical genetics.** Thesis, University of Groningen, with summary in English and Dutch.

The research presented in this thesis was mainly performed at the Genomics Coordination Center, Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. The work in this thesis was financially supported by European Union Seventh Framework Programme (FP7/2007-2013) research projects BioSHaRE-EU (261433) and PANACEA (222936), BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007), and NWO VIDI grant number 917.164.455.

Printing of this thesis was financially supported by Rijksuniversiteit Groningen, University Medical Center Groningen, Groningen University Institute for Drug Exploration (GUIDE) and NWO VIDI grant number 917.164.455.

Cover design and layout by JA Bookdesign. The front cover features a Gource (<http://gource.io>) visualization of the MOLGENIS software repository (<http://github.com/molgenis/molgenis>) used throughout this thesis. The sunrise gradient and DNA symbolize the dawn of molecular genetics.

Printed by Ipskamp Drukkers, Enschede.

© 2017 K.J. van der Velde. All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means without permission of the author.

ISBN: 978-94-034-0351-9

ISBN (electronic version): 978-94-034-0350-2



umcg



rijksuniversiteit
 groningen





rijksuniversiteit
 groningen

Translational software infrastructure for medical genetics

Proefschrift

ter verkrijging van de graad van doctor aan de
 Rijksuniversiteit Groningen
 op gezag van de
 rector magnificus prof. dr. E. Sterken
 en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

maandag 8 januari 2018 om 14.30 uur

door

Kasper Joeri van der Velde

geboren op 24 mei 1986
 te Smalingerland

Promotores

Prof. dr. M.A. Swertz

Prof. dr. R.J. Sinke

Copromotor

Dr. Y. Li

Beoordelingscommissie

Prof. dr. R.K. Weersma

Prof. dr. V.V.A.M. Knoers

Prof. dr. P.L. Horvatovich

Paranimfen

Bart Charbon

Freerk van Dijk

Contents

1	Introduction	13
1.1	The origin of genetics	14
1.2	The genome in the clinic	18
1.3	Data interpretation challenges	20
1.4	Bioinformatic opportunities	22
1.4.1	Population reference genomes	24
1.4.2	Genomic association studies	26
1.4.3	Additional molecular data	28
1.4.4	Computational and 'big data' approaches	29
1.5	Thesis outline	30
1.5.1	New models to integrate life science data	31
1.5.2	New methods to translate research findings	31
1.5.3	New systems for medical genetics practice	33
2	XGAP model for genotype and phenotype experiments	35
2.1	Background	37
2.2	Minimal and extensible object model	41

2.3	Simple text-file format for data exchange	48
2.4	Easy to customize software infrastructure	49
2.4.1	Graphical user interface	51
2.4.2	Application programming interfaces	53
2.4.3	Import/export wizards	55
2.4.4	Customizing XGAP	57
2.5	Conclusions	58
2.6	Materials and methods	62
3	A scalable web environment for multi-level QTL analysis	67
3.1	Introduction	69
3.2	Features	69
3.2.1	Explore QTL profiles	70
3.2.2	Single and multiple QTL mapping	70
3.2.3	Add new analysis tools	70
3.2.4	Track analysis and monitor performance	72
3.2.5	Scalable data management	72
3.2.6	Customizable to research needs	72
3.3	Implementation	73
3.4	Conclusion	73
4	A web database for linking human disease to <i>C. elegans</i>	75
4.1	Introduction	77
4.2	Implementation	79
4.2.1	Tool 1: 'Disease2QTL'	83
4.2.2	Tool 2: 'Region2disease'	84
4.2.3	Tool 3: 'QTL2disease'	85
4.2.4	Tool 4: 'ComparePheno'	85
4.2.5	Software used	85
4.3	Results	86
4.3.1	Case 1: McGary <i>et al.</i>	86
4.3.2	Case 2: Li <i>et al.</i>	87
4.3.3	Case 3: Rodriguez <i>et al.</i>	88

4.3.4	Novel disease-gene associations	89
4.4	Discussion	93
5	Evaluation of CADD Scores in Mismatch Repair Genes	97
5.1	Introduction	99
5.2	Materials & Methods	102
5.2.1	Data processing	102
5.2.2	Cumulative link model	103
5.2.3	Data availability	106
5.3	Results	106
5.3.1	Exploratory data analysis	106
5.3.2	Discrepancy assessment	107
5.3.3	False positives	107
5.3.4	False negatives	110
5.3.5	Variants of unknown significance	112
5.4	Discussion	114
6	Variant interpretation for medical sequencing	129
6.1	Background	131
6.2	Results	132
6.2.1	Development of GAVIN	132
6.2.2	Performance benchmark	133
6.2.3	Added value of gene-specific calibration	134
6.3	Discussion	139
6.4	Conclusions	144
6.5	Methods	145
6.5.1	Calibration of gene-specific thresholds	145
6.5.2	Variant sets for benchmarking	148
6.5.3	Variant data processing and preparation	149
6.5.4	Execution of in silico predictors	150
6.5.5	Stratification of variants using ClinGenDatab.	152
6.5.6	Implementation	152
6.5.7	Binary classification metrics	153

7	A bioinf. framework for downstream genome analysis	157
7.1	Introduction	159
7.2	Results	161
7.2.1	Framework for downstream genome analysis . .	161
7.2.2	Implementation for genome diagnostics	166
7.2.3	Validation tool: evaluation for diagnostics . . .	170
7.3	Discussion	176
7.3.1	Framework considerations	178
7.3.2	Implementation enhancements	179
7.3.3	Increasing diagnostic yield	180
7.4	Conclusion	182
7.5	Methods and Materials	183
7.5.1	MOLGENIS annotation tool	183
7.5.2	Population reference for false discovery analysis	184
7.5.3	Pathogenic variants for false omission analysis .	185
7.5.4	GAVIN+ interpretation tool	185
7.5.5	Running false omission analysis	186
7.5.6	Running false discovery analysis	187
7.5.7	Visualizing FOR and FDR analysis results . . .	188
7.5.8	MOLGENIS reporting tool	188
8	Discussion and Perspectives	191
8.1	Flexible models for life science omics data	193
8.1.1	Integration of heterogeneous omics data	194
8.1.2	Making omics data reusable across systems . .	198
8.1.3	Spreadsheets in the era of big complex data . .	204
8.1.4	Future perspectives of sharing life science data	208
8.2	Developing computational methods for medical genetics	211
8.2.1	Method dependance on high quality data	212
8.2.2	Benchmarking and characterization of methods	216
8.2.3	Finding appropriate methods in repositories . .	221
8.2.4	Integrating and running methods for evaluation	224
8.3	Towards better systems for (gen)omic medicine	225

8.3.1	Reusable and flexible DNA analysis workflows .	227
8.3.2	Community sharing of protocols and expertise .	230
8.3.3	Towards integrated multi-omics analyses	231
8.3.4	Future work on semantic analysis systems	234
8.4	Conclusion	237
	Bibliography	239
	List of Tables	297
	List of Figures	299
	Appendices	301
	A Summary	303
	B Samenvatting	307
	C Acknowledgements	311
	D About the author	315
	E List of publications	317
	F Other academic activities	323



Chapter 1

Introduction

1

2

3

4

5

6

7

8

1 Translational medical genetics is a cross-disciplinary field of research
2 that strives to advance genomic medicine using state-of-the-art find-
3 ings from life sciences. In this thesis, I contribute bioinformatic models,
4 methods and systems to improve the rate and precision of patient di-
5 agnoses by harnessing untapped molecular information.

6 I start this introduction by discussing the origins of the field of
7 genetics and how it evolved to its current state (1.1). I then zoom
8 in on medical genetics and explain how our growing understanding of
genetic disorders benefits patients (1.2). Recent revolutions in DNA
sequencing now allow us to complement genetics with genomics, which
is the physical characterization of the genome itself. I explain how this
shift presents medical practice with many exciting opportunities but
also with equally big challenges (1.3) for successful implementation.
These challenges that feed into the research questions addressed in this
thesis (1.4). The introduction ends with an overview of the chapters of
this thesis (1.5). Each chapter presents research that aims to translate
these opportunities into better understanding, diagnosis, and ultimately
treatment of genetic disorders.

1.1 The origin of genetics: solving the genetic riddle piece by “peas”

Just over 150 years ago, Gregor Mendel was the first to describe the basic rules of genetic inheritance[230]. He discovered striking patterns in how the traits of pea plants such as color and shape were transmitted from one generation to the next. His work established genetics as a science, even though he could not know the molecular basis of his observations.

It was only decades later that Mendel's theories were finally recognized, and the term *gene* was coined[170] as an innate unit of inheritance. Genes were defined as the effect observed on the traits (i.e. *phenotype*, these and other terms are clarified in Tables 1.1 and 1.2),

and were not yet measured on a molecular level.

Traits may be passed on to the next generation independently from each other, but some traits seemed to be passed on together with a certain frequency. The relative distance of the underlying genes could be estimated by analyzing this effect, called *linkage*[236], although it was only realized much later that this is related to physical distance on a DNA molecule. Further studies showed that genes seemed to direct enzyme synthesis[24] and that nucleic acid was the carrier of genetic material[19, 150], not proteins, as was the popular belief¹.

How the information of genes was stored in nucleic acid was unclear until the physical structure of DNA was elucidated[361]. This was followed by the cracking of the genetic code[193], specifically how DNA codon triplets are translated via temporary RNA-based copies into amino acid sequences that fold into functional proteins. These proteins are the workhorses of the cell. They communicate with other cells (e.g. via excretions and receptors), process metabolic substrates (e.g. via glycolysis) and regulate cell homeostasis (e.g. via signal transduction).

The first completely sequenced genome was that of a virus, Bacteriophage MS2, which has just 3,569 DNA bases[101]. When the human genome sequence was completed in the year 2000[189], it turned out to have over 3,000,000,000 base pairs². With this milestone, the field of human molecular genetics gained huge momentum. Now, traits and diseases could be associated to the actual sequence of DNA instead of an abstract linkage map or approximate cytogenetic location (observable chromosomal aberrations leading to a disease phenotype).

¹Nevertheless, there are known mechanisms for protein-based inheritance[276, 52] and cell memory[51].

²The largest reliably measured genome currently known belongs to the Japanese Canopy plant *P. japonica*, which has 150,000,000,000 base pairs[258].

CHAPTER 1. INTRODUCTION

Term	Definition
Allele	A variant form of a gene or genetic locus
Amino acid	Small organic building block of proteins
Base	Building block of nucleic acid
Base pair	Two bases bound by hydrogen in the DNA double helix
Chromosome	Organizational unit of DNA, humans have 22 pairs plus XX or XY
Codon triplet	Sequence of three bases that codes for a specific amino acid
Complex disease	A disease caused by the joined effect of multiple environmental and genetic factors
Conserved loci	Genomic locations that have changed little in evolution
Diagnostic yield	The percentage of solved patient cases
DNA	Deoxyribonucleic acid, encodes the genetic information of an organism
Dominant disease	A disease caused by a single pathogenic allele on one chromosome of a pair
Enzyme synthesis	Production of proteins that act in, or execute chemical reactions
Exon	Short for 'expressed region', coding sections of a DNA sequence
Genetic code	Rules by which nucleic acid is translated into messenger RNA
Genetic inheritance	Transmission of inborn traits from parent to offspring
Genome sequencing	Determining the order of bases in a genome

Table 1.1: Glossary of key terms used in this introduction, pt. 1/2.

1.1. THE ORIGIN OF GENETICS

Term	Definition	
Homeostasis	Active regulation to maintain a stable equilibrium of variables in an organism	1
Mendelian inheritance	Set of rules for the basic modes of inheritance for single-gene diseases	2
Mutation	A change of the nucleotide sequence of the genome	3
Non-Mendelian inheritance	More complex inheritance patterns such as additive, co-dominance, polygenic, imprinting or heterosis	4
Nucleic acid	Biopolymer consisting of sugars, phosphates and nitrogenous bases	5
Oligogenic	A few genes controlling a trait	6
Penetrance	The proportion of individuals adversely affected by a pathogenic mutation	7
Phenotype	Collection of observable characteristics of an organism resulting from the interaction of its genotype with the environment	8
Polymorphism	A neutral and commonly present mutation	
Proteins	Large biomolecules with a variety of functions	
Recessive disease	A disease caused by pathogenic alleles on both chromosomes of a pair	
RNA	Ribonucleic acid, predominantly acts as a messenger carrying instructions from DNA for controlling protein synthesis	
Splicing	The process by which exons are joined to form messenger RNA	
Variant	General term for all mutations and polymorphisms	

Table 1.2: Glossary of key terms used in this introduction, pt. 2/2.

1.2 The genome in the clinic

1 Inheritance patterns of inborn disorders in humans have been stud-
2 ied since the rediscovery of Mendel's work, with study focusing mainly
3 on genes. The first inborn disorder to be described was alkaptonu-
4 ria[117], a recessive disease with a prevalence of 1:100,000 to 1:250,-
5 000[382] caused by mutations (i.e. *variants*, small genetic differences)
6 in the *HGD* gene on chromosome 3. There are now around 8,000 such
7 gene-related disorders catalogued in the OMIM[142], Orphanet[11] and
8 DECIPHER[102] databases. For about 4,300 of these disorders an as-
sociated gene has been discovered, of which 3,300 are characterized as
clinically actionable to some degree[319].

The majority of clinical genes currently known usually follow a Men-
delian inheritance pattern, because those are more straightforward to
discover and characterize. These Mendelian disease genes are tradi-
tionally discovered by investigating the transmission pattern of specific
mutations through a family pedigree. The top candidates for these con-
firmation studies are usually rare mutations at conserved genomic loci
that strongly coincide with being affected by the disease. After more
independent families or patients have been found with the same symp-
toms and the same mutation or affected gene[317], a causal relation is
established[61].

Finding causal genes is not a trivial effort and additional difficulty
may be introduced by oligogenic inheritance[119], incomplete penetran-
ce[306, 240], or variants that have an unexpected effect[93]. Causal
mutations and genes are catalogued in databases such as Clinical Ge-
nomics Database[319] and ClinVar[190].

Knowing which genes are responsible for disorders provides many
opportunities to improve patient care through applications such as im-
proved disease diagnosis, carrier screening, personalized medicine and
life course advice.

Firstly, we can use genomic knowledge for more accurate **disease
diagnosis**. For example, there are five subtypes of cardiomyopathies,

with over 60 genes involved[173]. Gene sets specific for a given disease type are called *panels*. Gene screening panels have been created for conditions including dystonia, dermatology, autoinflammatory diseases, epilepsy, familial cancer, intellectual disability and metabolic disorders. Finding a pathogenic mutation in one of these genes may lead to a diagnosis on the molecular level, which is more precise than a diagnosis based only on symptoms.

The second opportunity to improve healthcare is through **carrier screening**, which assesses whether a person is carrying a specific pathogenic mutation present in their family. A special case of carrier screening is preconception screening where it is determined whether both parents carry alleles in the same gene known to cause severe recessive disease. In this case the parents are not at risk, but a potential child has - following Mendelian laws - a 25% chance of inheriting both alleles, and thus being affected.

Lastly, **personalized medicine** refers to better tailoring of medication and treatment based on the genotype of the patient. A well-known example is the adjustment of the starting dose of warfarin depending on the patient's *CYP2C9* and *VKORC1* genes[207, 349], which affect their ability to metabolize this drug.

While DNA analysis was very costly until recently, which limited analysis to one or a few genes, clinical geneticists can now perform DNA-sequencing on their patient using a panel of many genes, or choose to look at all 26,000 genes at once using whole-exome sequencing (WES), or even consider whole-genome sequencing (WGS), thanks to *next-generation DNA-sequencing* techniques (NGS), which has largely replaced more traditional techniques such as Sanger sequencing[300]. The cost of sequencing a genome with NGS has dropped dramatically, from \$10 million in 2007 to only \$1,000 in 2015³, paving the way for a genomic revolution.

WES allows us to investigate thousands of genes at once in research or diagnostics[378]. This technique is useful for making a genome-

³<https://www.genome.gov/sequencingcostsdata/>

driven diagnosis when symptoms are hard to assess, for example in newborns[348] and other isolated cases[365]. WES allows analysis of exons and corresponding splice-site regions.

1 Using WGS we can look at 'non-coding' DNA, which is not tran-
2 scribed to protein but is still involved in the regulation of genes. From
3 application of WGS we now know that non-coding variants are also
4 implicated in disease[168] and that the genome is organized in topolog-
5 ical domains[81], structural changes in which are linked to pathogenic
6 effects[107]. This relatively new area of genomic research is already
7 becoming of diagnostic relevance[313].

8 Regardless which technique is used, a molecular diagnosis provides
an unprecedented ability to help patients. Most notably, a diagno-
sis can be established long before symptoms have developed, allowing
recognition and sometimes intervention that may prevent permanent
damage[302]. In all cases, a more informed diagnosis will lead to a
clearer prognosis and more appropriate treatment plan based on the
molecular etiology of the disease. However, when using genome-wide
screening approaches, incidental findings have to be dealt with[134], as
they can cause serious issues including a high patient opt-out rate[146].

1.3 Data interpretation challenges

Although genetic screening is successfully employed in many clinics around the world, we now effectively use only a minute amount of the genetic knowledge contained within the data generated. For most of the genes, and for almost all of the non-coding genome, we do not know the clinical relevance. A genetic cause has been established for only about half of the known Mendelian disorders, and we are only just starting to understand complex diseases[224]. Even within the known genes, it is not always clear if a mutation is harmful[67, 54]. As a result, the interpretation and subsequent classification of DNA variants is a major challenge.

The difficulty of this challenge is shown by the diagnostic yields currently achieved, which vary from 15 to 80%[356, 221, 74, 380] depending on factors such as disease type, patient inclusion criteria and sequencing technique used. This challenge is further shown by the discordant results given by direct-to-consumer genomic analysis companies[69] and by the re-classification of pathogenic variants as harmless when more data becomes available[49].

Furthermore, the production of genomic data is far outpacing the rate at which geneticists can interpret it, a circumstance referred to as the 'NGS data deluge'[303]. Big data analytics is thus a major challenge in healthcare[280, 26], as are related efforts to translate research data and research findings into healthcare improvements[9, 17]. This is especially true for the area of medical genetics[124, 359, 34]. Adding more layers of molecular information, such as transcriptomics or epigenetics, only makes sense when combined with infrastructure and analysis methods that use these data to make clinical decisions easier instead of more complicated.

Computers can help us integrate and analyze large and complex data, provided appropriate software is available to do so. The field of bioinformatics develops these tools, but it takes more than a few lines of code to improve patient care. This barriers for setting up infrastructure can be broken down into effective data integration, method development and implementation into practice. In this thesis we identify and address the following challenges:

1. We need data **models** to integrate life science data for genetic disease research. By systematically integrating and visualizing large amounts of data sets, we allow researchers to discover new disease genes. These genes can then be tested in patients, leading to higher diagnostic yield.
2. We need computational **methods** to translate research findings to medical genetics. Many research findings are of potential benefit to patient care, but they require tailoring, calibration and

validation into a clinical genomics context before they can be used. Using more advanced analysis methods will result in more accurate and efficient characterization of patient mutations.

3. We need software **systems** to implement methods into medical genetics practice. These systems are needed to test, validate and utilize new methods and must be flexible enough to allow quick adoption of future developments, including new methods and data modalities.

1.4 Bioinformatic opportunities

Empowering clinical geneticists with the tremendous amount and variety of new life science data is the huge challenge that forms the objective of this thesis. Basic science in biology and genetics includes studies on model organisms, human populations, creation of computational algorithms and the molecular characterization of cells and tissues, and all these types of research present possibilities to improve patient diagnoses. At the same time, medical practice offers invaluable insights about disease etiology, patient cases, and data gathered in a clinical setting that can be used to develop and validate new methods for medical application. All these new data offer major opportunities for finding, understanding and treating human disease. Figure 1.1 illustrates the efforts and collaborations in the translational research needed to realize this potential. In the paragraphs below and more detailed sections devoted to them that follow, we introduce the key research topics that are the focus of this thesis:

Reference genomes - Population studies can tell us what to expect in the average individual. Through phenotypic and molecular characterization of large groups of healthy individuals, we can establish a reference population. Strong deviation from this reference may point towards causal mechanisms of molecular disease for more severe disorders that are highly damaging or otherwise debilitating at a younger age.

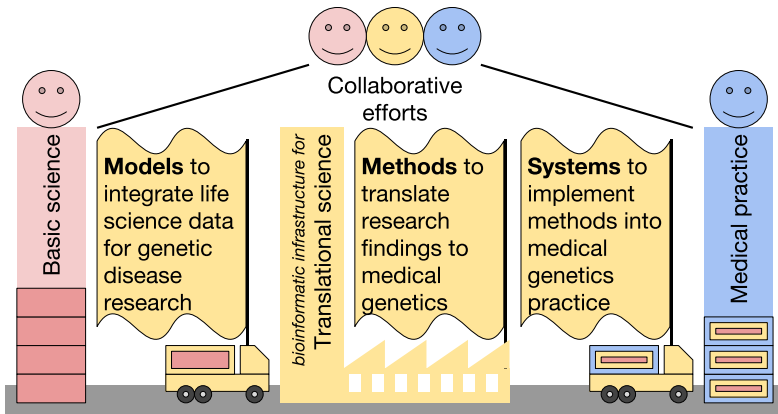


Figure 1.1: Overview of bioinformatic infrastructure for translational science. Fundamental knowledge originates in basic research (red). Translational research (yellow) bridges the gap between basic research and medical practice (blue) by collaborative efforts from all parties involved.

1

2

3

4

5

6

7

8

1 Association studies - Patient studies can tell us how disorders originate. While studying small numbers of individuals can still uncover new Mendelian disease genes[371], larger number of patients are required for statistical association of new disease candidate genes with less obvious effects[197]. By using extremely large sample sizes, we can also detect genetic associations for complex but more common afflictions such as celiac disease or obesity.

2
3 Additional molecular data - The genome is the prime information carrier within a living cell, but many more molecular levels stand between the DNA sequence and the eventual expression of a phenotype. By measuring these different levels we can attempt to reconstruct both the lateral interactions (protein-protein interactions or gene co-expression networks) and perpendicular interactions (protein binding to the genome to silence expression or metabolite accumulation causing neurodegeneration), which can help to understand the workings of disease in detail.

4
5
6
7 Computational and 'big data' approaches - The rich collection of current life science data provides great opportunities for the development of smart software programs, computational algorithms and statistical tools that can extract knowledge from these growing data resources. These must perform a multitude of roles and functions, including cleaning and quality control of raw data, imputing missing data points, finding statistical associations, modeling and running predictors, or constructing and pruning networks of detected relations. In the following paragraphs I will explore these opportunities in detail.

8 **1.4.1 Population reference genomes**

Genomes are relatively similar between individuals, therefore, instead of assembling the complete sequence for each person, we only determine points of DNA variation compared to a reference genome. Subsequently, we can aggregate the results by counting how often each point of variation was observed. This allows us to store the information of thousands of genomes in files that are still quite computationally manageable and

require smaller amounts of data storage capacity. There are a number of initiatives that have collected the DNA variation of healthy individuals, such as the Thousand Genomes project[63] (2,504 genomes), the Genome of the Netherlands[244] (750 genomes), the Exome Aggregation Consortium[196] (60,706 exomes), the NHLBI Exome Sequencing Project[95] (6,503 exomes) and the upcoming gnomAD from the ExAC authors[196] (126,216 exomes and 15,137 genomes). Here, the term “healthy” refers to individuals who do not suffer from a severe inborn disorder. They may still develop common late-onset diseases with genetic components such as type 2 diabetes, cardiovascular problems, obesity or common forms of cancer.

These large reference sets find eager uptake in all areas of genetics including research and genome diagnostics. Variants observed to have a high allele frequency in a population of individuals are called polymorphisms. Such polymorphisms are very unlikely to directly cause a disease, although they might still act as modifiers (or markers) for disease risk[82]. We may apply a filter based on Minor Allele Frequency (MAF): the alternative allele fraction compared to the most frequent reference allele. A typical setting may be to exclude any variant from further analysis of a patient’s genome when it occurs more than 1% in the general population. Depending on the rarity and severity of a disease, we may want to use thresholds as low as 0.01% (see chapter 6) and as high as 5%[326]. We can also use the genotype zygosity counts, which is the number of individuals heterozygous or homozygous for an allele. If only heterozygous genotypes are observed in the general healthy population, we may be dealing with a recessive-acting disease variant, which quickly becomes a candidate for being pathogenic when detected homozygously in a patient.

As other types of population reference data becomes available, e.g. from RNA-sequencing[78], we have the opportunity to also establish a baseline for healthy individuals for data other than DNA variation. We can use these references to investigate and manually predict potential pathogenic effects in patients and capture the outcomes. These results

1

2

3

4

5

6

7

8

are then used to develop tools to speed up the interpretation of new patient data and initiate a synergistic process leading to exponential tool development.

1 Furthermore, big population data provide insight into our genomic
2 architecture. The mention of Mendelian disease genes may give the
3 impression that our genomes are fragile, but there is also evidence that
4 shows they are surprising resilient. Each healthy human has about 100
5 Loss of Function (LoF) variants with 20 genes completely inactiva-
6 ted[215]. We now have enough reference data to calculate an accurate
7 LoF rate for every gene[196], and this rate may be compared to a null
8 distribution to determine which genes are LoF-tolerant and which are
not. Any LoF-intolerant genes found in patients with severe mutations
can then be prioritized as potentially disease-causing. By analyzing the
selection pressure on truncating variants we can then characterize genes,
and estimate whether one or two dysfunctional alleles are likely to be
disease causative[50].

Lastly, these large reference sets have put things in new perspec-
tive. Some variants that were previously thought to be surely disease-
causing were found to have low *penetrance*, meaning that not every
individual with that mutation actually becomes ill[234]. Other vari-
ants once thought to have pathogenic effects have turned out to be
far too common with respect to disease prevalence, revealing them as
false positives[354]. Finally, on a more critical note, ethnicity biases in
these reference sets may result in misclassifications[220], indicating a
need for more diverse and representative data sets to be used in genome
diagnostic interpretation.

1.4.2 Genomic association studies

In Mendelian or *monogenic* genetic disorders, a single dysfunctional gene can cause severe problems. There are, however, numerous disease-related phenotypes that are not attributable to just one or a few genes. Instead, many locations (or *loci*) on the genome seem to each contribute

a small amount to the risk of the disease[224, 219, 35]. Finding these weak associations requires large Genome-Wide Association Studies or GWAS which may include more than 250,000 participants[374]. These large samples sizes can be achieved by genotyping arrays which can cheaply ascertain alleles of a predetermined set of variants.

In human, we have currently discovered about 30,000 trait-genome associations[363]. While these include general traits like word reading ability, alcohol consumption, hair color, height, and freckling, most traits are of medical relevance and include susceptibility to common diseases such as hypertension, arthritis, celiac disease, cancer subtypes, diabetes, cardiovascular disease, ulcerative colitis, obesity, allergies, psoriasis and asthma.

Establishing these associations is important for several reasons. Most notably, the locations where they are found implicate nearby genes that may be involved, making these genes the best candidates for further study. However, genes must be carefully considered because the closest gene is often not relevant and statistical approaches have been developed[389] to identify the strongest candidate in the region.

Another application of GWAS associations is modeling of genetic risk scores. The effect size of the risk-associated alleles that an individual is harboring can be summed to a genetic risk score[84]. This risk, by definition, correlates to either the chance of developing a certain disease or the occurrence of a clinical event[217], but genetic risk scores can also predict the quantitative severity of a clinical phenotype[27].

Based on a higher risk score, individuals may choose to undergo a specific medical check regularly, or adjust their lifestyles to improve their odds of not developing a certain disease. Conversely, individuals with strong protective alleles might need fewer periodic examinations than usual, allowing physicians to spend more time on people with a higher risk.

1

2

3

4

5

6

7

8

1.4.3 Additional molecular data

1 Beyond the DNA sequence, much additional molecular data can now be
2 gathered that can be used to identify which DNA variations are relevant
3 for health and disease, and which are not. Regulation of gene transcrip-
4 tion, translation, protein activity and degradation constantly takes place
5 at between different molecular levels. For instance, the genes on the
6 genome itself can be made harder to transcribe through methylation of
7 the cytosine and adenine nucleotides[31]. In addition, the chromosomal
8 structure of DNA can be decondensated by histone acetylation (trans-
9 fer of acetyl groups to DNA organizational elements), making it more
10 accessible for transcription[87]. The transcriptional expression of genes
11 is further regulated by genetic variants themselves[7]. Finally, proteins
12 form a complex network of interactions[265] that, in turn, also regulate
13 gene expression[331].

14 We study the complex patterns of this regulation to understand how
15 genes act in concert, and how a disease phenotype presents in cells, tis-
16 sues and organisms. Large initiatives that pursue this goal include stud-
17 ies into expression quantitative trait loci (eQTL)[364] and allele-specific
18 expression[78], characterization of functional genomic elements includ-
19 ing methylation and acetylation patterns[85], comprehensive expression
20 studies across different tissues[213] and cell types[105].

21 These same kinds of studies can also be performed on model or-
22 ganisms, which can be bred and measured in highly controlled environ-
23 ments for pin-point phenotypic and molecular characterization. Studies
24 on mice have been an essential tool for biological research for more than
25 a century and continue their important role today[264]. Mice are evolu-
26 tionarily relatively close to humans, and their size and short generation
27 time allows experiments to be set up and run with large enough num-
28 bers for statistical significance. However, other types of model organ-
29 isms such as zebrafish[206] and worm[176] can offer unique advantages
30 over using rodents. While these organisms have a larger evolutionary
31 distance to humans, they are cheaper, faster and easier to breed and

have transparent bodies that are easy to dissect. The tiny *C. elegans* worm has by far the fastest life cycle, simplest anatomy and the unique property of strains that can be frozen and revived.

In addition to transcriptomics and epigenetics, we can also measure the levels of metabolites and proteins present in cells. These technologies, known as metabolomics and proteomics, can be integrated with genomics data[132] to obtain a more complete understanding of the complex processes in the cell that interplay with all these layers. Finally, we can also investigate the genomic variation that prevents disease or even increase our health instead of looking for genes that make people ill. The search for so-called 'protective alleles' is an up and coming area of study that will also result in healthcare advancements[145].

1.4.4 Computational and 'big data' approaches

Measuring and interpreting the large, complex and diverse life-science datasets has driven the development of a plethora of new computational methods and tools to analyze these data. These include methods to clean and prepare data for analysis, advanced statistical methods, relational databases, web applications, data integration and visualization tools.

A few notable examples include the Variant Quality Score Recalibration (VQSR), a module of the Genome Analysis Toolkit (GATK)[344]. This tool performs comparative machine learning on identified (*called*) NGS variants versus a reference truth set to find the optimal variables for determining which variants are true positives and which are false.

Variants can also be determined using genotyping platforms, but when multiple platforms are used, data are not comparable. However, they can be harmonized by inferring missing variants using genotype imputation[77], which also uses reference knowledge.

After variants are determined, there are many tools that estimate variant pathogenicity to assist genome diagnostics or research into genetic diseases[90]. A powerful method to prioritize variants for further

1 interpretation are CADD scores[185]. These scores are a measure of
2 evolutionary pressure on genetic variants that builds upon 60+ existing
3 tools and sources. Variants with a higher score are more likely to be
4 deleterious and are therefore the best candidates in disease research.

5 Using CADD scores, variants are discovered in genes of which the
6 function is not yet known. Knowledge networks such as GeneMA-
7 NIA[360] may help to infer a putative function by linking unknown
8 genes to genes known from previous studies to show a similar expres-
sion pattern. We can also characterize unknown genes by their evolu-
tionary, loss-of-function and network interaction properties to prioritize
candidate variants[184] and even predict disease inheritance mode to a
certain degree[153].

Taking this approach a step further, GeneNetwork[99] is constructed
from co-regulation patterns found within tens of thousands of samples
for which gene expression was measured. GeneNetwork provides un-
precedented resolution and predictive power across multiple cell types
and tissues. Analogous to discovering patterns in expression data, the
network of protein-protein interactions can also be computationally pre-
dicted using various methods[381].

The combined current knowledge of how cells control functions
such as growth, movement, differentiation, metabolism, communica-
tion, and response to stress or pathogens is captured in high-level path-
way databases such as WikiPathways[188], Reactome[97] or KEGG[180].

Taken together, these tools provide important clues for wet-lab stud-
ies, which then in turn provide better and more meaningful biological
measurements that can help to develop new and improved methods.

1.5 Thesis outline

In this thesis I show how, by addressing data challenges and bioinfor-
matics opportunities in translational infrastructure, we can advance our
genetic knowledge and its application in medical genetics. The focus

of the first two chapters is on models that integrate life science data as a basis for finding new gene-disease associations. I then develop methods to discover leads for human disease and utilize pathogenicity estimates for clinical application. Finally, I implement software systems that translate what we have learned to medical genetics practice. An overview of the chapter progression in this thesis is shown in Figure 1.2.

1.5.1 New models to integrate life science data for genetic disease research (chapters 2 and 3)

There are many approaches for gathering, structuring, integrating and analyzing life science data, each best suited to test a specific hypothesis[290]. To help domain experts test new ideas and quickly interpret interesting findings, they should be able run the necessary queries, tools and visualizations themselves. To achieve this, the underlying data has to be both properly modeled ('computer-readable') and fortified with enough metadata to describe what the data means[366] so that it can be automatically addressed by applicable tools.

As data volumes grow ever larger, these tools have to be executed on external high-throughput computational environments such as multi-node computer clusters. To facilitate storage of these huge datasets and parallelized computation, we investigated how to store complex data using the flexible XGAP model in chapter 2, and used this as a basis to develop xQTL workbench in chapter 3. xQTL workbench is a flexible database system designed to store any genotype and phenotype information with basic visualization and computational capabilities.

1.5.2 New methods to translate research findings to medical genetics (chapters 4 and 5)

Translational medicine investigates how relevant new findings can be used to improve patient diagnosis and care. To demonstrate how new findings can be generated, we loaded almost 100 data sets of *C. elegans*

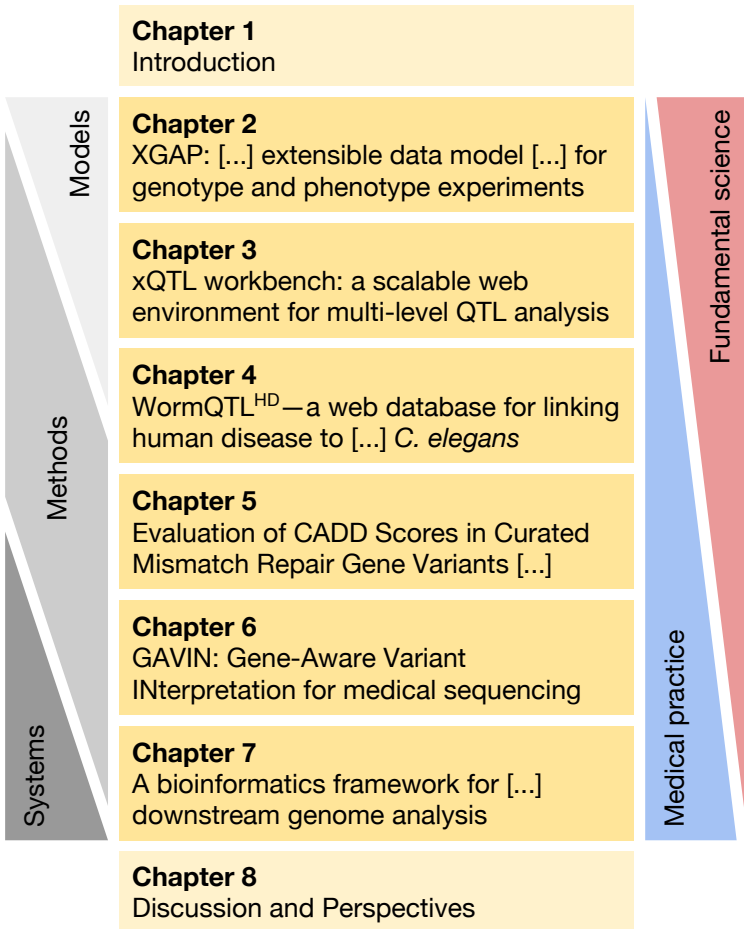


Figure 1.2: Overview of thesis chapter progression in terms of type of output and area of application. We can define an overall gradient from fundamental science to medical practice, as well as transitions from models to integrate life science data towards methods to translate discovered knowledge and systems to implement new methods into patient care.

into an xQTL database, containing around 300 million measurements. To show value for human health applications, we connected worm phenotypes to human disease at a molecular level using protein orthology. Chapter 4 shows how these data can now be used to find models and leads for human disease research. Furthermore, a biologist-friendly online environment enables the research community to join in and dig through the data.

Interesting findings need to be explored further and placed into clinical context before medical genetics can benefit from them. The previously mentioned CADD scores[185] are an example of an innovation with great potential. Doctors and clinical geneticists have an interest in such developments, but cannot use it in practice without guidelines about how to interpret these scores in patient cases. To explore how such a guideline is created and used, we translated CADD scores to the clinical classification of variants in mismatch repair genes in chapter 5. These genes may harbor variants that cause hereditary colorectal cancer. By characterizing these scores in this context, we learned both their pitfalls and how they can be used to prioritize new mutations or double-check existing classifications.

1.5.3 New systems to implement methods into medical genetics practice (chapters 6 and 7)

Large reference datasets and computational resources, when guided by translational research, should allow us to transform patient care. To facilitate this, we need to design, build and maintain reliable software systems[274] running on a stable server and database infrastructure[329]. These systems must handle rapidly increasing quantities of whole-genome data as sequencing costs dropped from a billion dollars to just a thousand dollars per patient. The data produced needs to be contrasted against large population reference sets and other patient genomes for research, interpretation or diagnosis using computational methods. The storage, processing and filtering solutions for these massive datasets

need the capabilities to be scaled up, fine-tuned and clinically validated accordingly.

1 Encouraged by results of chapter 5, we generalized the CADD score
2 calibration approach and applied it to >3,000 disease genes. We em-
3 phasize practical use by excluding variants that would also be excluded by
4 existing methods. On the variants that remain that are hard to inter-
5 pret, we find out if CADD scores can be of further help. The resulting
6 predictor tool, GAVIN, is described in chapter 6 and works remarkably
7 well for clinically characterized genes. It serves as a first-lead causal
8 variant screening tool with broad application in clinical genomics.

This work then feeds into chapter 7, where we define a framework
to automate the interpretation of genomic data, and to fast-track in-
novations in this process. We implement the GAVIN+ interpretation
tool, which combines GAVIN with additional knowledge and criteria
from clinical genetics to quickly identify variants and genotypes that
are potentially disease-causing. This tool outputs its result in the new
rVCF (Report VCF) format, which captures any relevant analysis re-
sults along with detailed provenance information and the reason why a
variant is of interest.

Using this format, we can run fast validation on known pathogenic
variants and estimation of false discovery rate on healthy control sam-
ples. The final result can be visualized in a customizable doctor-friendly
report, analyzed further as the format is fully compliant with existing
tools, and shared with peers. The modular framework design separates
the enrichment, interpretation and visualization of the data.

Our proposed solution is flexible and maintainable and its standard-
ized formats allows the community to develop focused software tools
that produce and utilize these files. As a result, newly developed meth-
ods can be quickly adapted and validated within local installation of the
framework. This high-throughput infrastructure will speed up molecu-
lar diagnostic practice, and prepare it for seamless future integration of
new analysis methods and powerful new omics techniques.

Chapter 2

XGAP: A uniform and extensible data model and software platform for genotype and phenotype experiments

Genome Biol. 2010;11(3):R27.
DOI: 10.1186/gb-2010-11-3-r27
PubMed ID: 20214801

CHAPTER 2. XGAP MODEL FOR GENOTYPE AND PHENOTYPE

Morris A. Swertz^{1,2,3,*}, K. Joeri van der Velde^{1,2}, Bruno M. Tesson², Richard A Scheltema², Danny Arends^{1,2}, Gonzalo Vera², Rudi Alberts⁴, Martijn Dijkstra⁵, Paul Schofield⁶, Klaus Schughart⁴, John M. Hancock⁷, Damian Smedley³, Katy Wolstencroft⁸, Carole Goble⁸, Engbert O. de Brock⁹, Andrew R. Jones¹⁰, Helen E. Parkinson³, members of the Coordination of Mouse Informatics Resources (CASIMIR)⁶, Genotype-To-Phenotype (GEN2PHEN) Consortiums¹, Ritsert C. Jansen^{1,2}

1. Genomics Coordination Center, Department of Genetics, University Medical Center Groningen and University of Groningen, 9700 RB Groningen, The Netherlands

2. Groningen Bioinformatics Center, University of Groningen, 9750 AA Haren, The Netherlands

3. EMBL - European Bioinformatics Institute, Hinxton, Wellcome Trust Genome Campus Hinxton, Cambridge CB10 1SD, UK

4. Experimental Mouse Genetics, Helmholtz Center for Infection Research, Inhoffenstraße 7, D-38124 Braunschweig, Germany

5. Center for Medical Biomics, University of Groningen, Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands

6. Physiological Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3DY, UK 7. Bioinformatics Group, MRC Harwell, Harwell, Oxfordshire OX11 0RD, UK

8. Information Management Group, School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, UK

9. Department of Business and ICT, Faculty of Economics and Business, University of Groningen, 9700 AV Groningen, The Netherlands

10. Department of Pre-Clinical Veterinary Science and Veterinary Pathology, Faculty of Veterinary Science, University of Liverpool, Liverpool L69 7ZJ, UK

Received 2009 Jul 14; Revised 2009 Dec 17; Accepted 2010 Mar 9.

* Corresponding author.

Abstract

We present an extensible software model for the genotype and phenotype community, XGAP. Readers can download a standard XGAP (<http://www.xgap.org>) or auto-generate a custom version using MOLGENIS with programming interfaces to R-software and web-services or user interfaces for biologists. XGAP has simple load formats for any type of genotype, epigenotype, transcript, protein, metabolite or other phenotype data. Current functionality includes tools ranging from eQTL analysis in mouse to genome-wide association studies in humans.

2.1 Background

Modern genetic and genomic technologies provide researchers with unprecedented amounts of raw and processed data. For example, recent genetical genomics[204, 167, 200] studies have mapped gene expression (eQTL), protein abundance (pQTL) and metabolite abundance (mQTL) to genetic variation using genome-wide linkage and genome-wide association experiments on various microarray, mass spectrometry and proton nuclear magnetic resonance (NMR) platforms and in a wide range of organisms, including human[88, 141, 80, 325, 148], yeast[37, 106], mouse[45], rat[156], *Caenorhabditis elegans*[205] and *Arabidopsis thaliana*[182, 183, 110].

Understanding these and other high-tech genotype-to-phenotype data is challenging and depends on suitable 'cyber infrastructure' to integrate and analyze data[322, 98]: data infrastructures to store and query the data from different organisms, biomolecular profiling technologies, analysis protocols and experimental designs; graphical user interfaces (GUIs) to submit, trace and retrieve these particular data; communicating infrastructure in, for example, R[158], Java and web

1 services to connect to different processing infrastructures for statistical
2 analysis[48, 8, 111, 30, 38] and/or integration of background informa-
3 tion from public databases[311]; and a simple file format to load and
4 exchange data within and between projects.

5 Many elements of the required cyber infrastructure are available:
6 The Generic Model Organism Database (GMOD) community developed
7 the Chado schema for sequence, expression and phenotype data[237]
8 and delivered reusable software components like gbrowse[321]; the Bio-
Conductor community has produced many analysis packages that in-
clude data structures for particular profiling technologies and experi-
mental protocols[121]; and numerous bespoke databases, data models,
schemas and formats have been produced, such as the public and pri-
vate microarray expression databases and exchange formats[36, 299,
115]. Some integrated cyber infrastructures are also available: the Na-
tional Center for Biotechnology Information (NCBI) has launched db-
GaP (database of genotypes and phenotypes)[216], a public database
to archive genotype and clinical phenotype data from human studies;
and the Complex Trait Consortium has launched GeneNetwork[57], a
database for mouse genotype, classical phenotype and gene expression
phenotype data with tools for ‘per-trait’ quantitative trait loci (QTL)
analysis.

However, a suitable and customizable integration of these elements
to support high throughput genotype-to-phenotype experiments is still
needed[340]: dbGaP, GeneNetwork and the model organism databases
are designed as international repositories and not to serve as general
data infrastructure for individual projects; many of the existing bespoke
data models are too complicated and specialized, hard to integrate be-
tween profiling technologies, or lack software support to easily connect
to new analysis tools; and customization of the existing infrastructures
dbGaP, GeneNetwork or other international repositories[384, 154] or
assembly of Bioconductor and generic model organism database com-
ponents to suit particular experimental designs, organisms and biotech-
nologies still requires many minor and sometimes major manual changes

in the software code that go beyond what individual lab bioinformaticians can or should do, and result in duplicated efforts between labs if attempted.

To fill this gap we here report development of an extensible data infrastructure for genotype and phenotype experiments (XGAP) that is designed as a platform to exchange data and tools and to be easily customized into variants to suit local experimental models. We therefore adopted an alternative software engineering strategy, as outlined in our recent review[329], that enables generation of such software efficiently using three components: a compact and extensible ‘standard’ model of data and software; a high-level domain-specific language (DSL) to simply describe biology-specific customizations to this software; and a software code generator to automatically translate models and extensions into all low-level program files of the complete working software, building on reusable elements such as listed above as well as general informatics elements and some new/optimized elements that were missing.

Below we detail XGAPs extensible ‘standard’ software model (XGAP-OM) and evaluate the auto-generated text file exchange format (XGAP-TAB) and customizable database software (XGAP-DB) that should help researchers to quickly use and adapt XGAP as a platform for their genetics and/or *omics experiments (Table 2.1). Harmonized data representations and programmatic interfaces aim to reduce the need for multiple format converters and easy sharing of downstream analysis tools via a hub-and-spoke architecture. Use of software auto-generation, implemented using MOLGENIS, aims to ease and speed up customization/variation into new XGAP versions for new biotechnologies and alternative experimental designs while ensuring consistent programming interfaces for the integration and sharing of existing analysis tools. Standardized extension mechanisms should balance between format/interface stability for existing data types and tools, and flexibility to adopt new ones.

CHAPTER 2. XGAP MODEL FOR GENOTYPE AND PHENOTYPE

1	Store	Store genotype and phenotype experimental data using only four 'core' data types: <i>Trait</i> , <i>Subject</i> , <i>Data</i> , and <i>DataElement</i> . For example: a single-channel microarray reports raw gene expression <i>Data</i> for each microarray probe <i>Trait</i> and each individual <i>Subject</i> . Add information on data provenance by giving details in <i>Investigation</i> , <i>Protocols</i> and <i>ProtocolApplications</i>
2	Customize	Customize 'my' XGAP database with extended variants of <i>Trait</i> and <i>Subject</i> . In the online XGAP demonstrator, <i>Probe</i> traits have a sequence and genome location and <i>Strain</i> subjects have parent strains and (in)breeding method. Describe extensions using MOLGENIS language and the generator automatically changes XGAP database software to your research
3		
4		
5	Upload	Upload data from measurement devices, public databases, collaborating XGAP databases, or a public XGAP repository with community data. Simply download trait information as tab-delimited files from one XGAP and upload it into another; this works because of the uniformity of the core data types (and extensions thereof)
6		
7	Search	Search genetical genomics data using the graphical user interface with advanced query tools. The uniformity of the 'code generated' interfaces make it easy to learn and use interfaces for both 'core' data types as well as customized extensions
8		
	Analyze	Analyze data by connecting tools using simple methods in Java, R, Web Services or Internet hyperlinks. For example, map and plot quantitative trait loci in R using XGAP data retrieved via the R interface
	Plug-in	Plug-in the best analysis tools into the user interface so biologists can use them. Bioinformaticians are provided with simple mechanisms to seamlessly add such tools to XGAP, building on the automatically generated GUI and API building blocks
	Share	Share data, customizations, connected analysis tools and user interface plug-ins with the genetical genomics community, using XGAP as exchange platform. For example, the MetaNetwork R package can talk to data in XGAP. This makes it easy for other XGAP owners to also use it
		API: application programming interface; GUI: graphical user interface; MOLGENIS: biosoftware generator for MOLEcular GENetics Information Systems.

Table 2.1: Features of XGAP database for genotype and phenotype experiments.

2.2 Minimal and extensible object model

We developed the XGAP object model to uniformly capture the wide variety of (future) genotype and phenotype data, building on generic standard model FuGE (Functional Genomics Experiment)[171] for describing the experimental ‘metadata’ on samples, protocols and experimental variables of functional genomics experiments, the OBO model (of the Open Biological and Biomedical Ontologies foundry for use of standard and controlled vocabularies and ontologies that ease integration[314], and lessons learned from previous, profiling technology-specific modeling efforts[36].

Figure 2.1b shows the core components of a genotype-to-phenotype investigation: the biological subjects studied (for example, human individuals, mouse strains, plant tissue samples), the biomolecular protocols used (for example, Affymetrix, Illumina, Qiagen, liquid chromatography-mass spectrometry (LC/MS), Orbitrap, NMR), the trait data generated (usually data matrices with, for example, phenotype or transcript abundance data), the additional information on these traits (for example, genome location of a transcript, masses of LC/MS peaks), the wet-lab or computational protocols used (for example, MetaNetwork[111] in the case of QTL and network analysis) and the derived data (for example, QTL likelihood curves).

We describe these biological components using FuGE data types and XGAP extensions thereof. *Investigation* binds all details of an investigation. Each investigation may apply a series of biomolecular[41] and computational[48, 8, 111, 30] *Protocols*. The applications of such *Protocols* are termed *ProtocolApplications*, which in the case of computational *Protocols* may require input *Data* and will deliver output *Data*. These *Data* have the form of matrices, the *DataElements* of which have a row and a column index. Each row and column refers to a *DimensionElement*, being a particular *Subject* or a particular *Trait*. Table 2.2 illustrates the usage of these core data types.

Figure 2.1a, c shows how the XGAP model can be extended to ac-

CHAPTER 2. XGAP MODEL FOR GENOTYPE AND PHENOTYPE

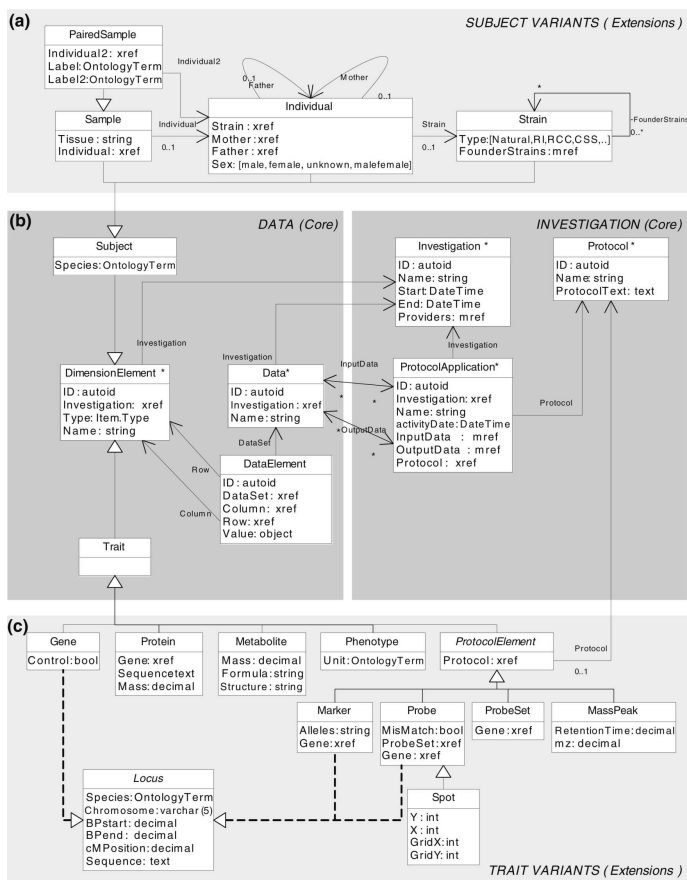


Figure 2.1: Extensible genotype and phenotype object model. Experimental genotype and (molecular) phenotype data can be described using *Subject*, *Trait*, *Data* and *DataElement*; the experimental procedures can be described using *Investigation*, *Protocol* and *ProtocolApplication* (b). Specific attributes and relationships can be added by extending core data types, for example, *Sample* and *Gene* (a, c). See Table 2.2, 2.3 and 2.4 for uses of this model. The model is visualized in the Unified Modeling Language (UML): arrows denote relationships (*Data* has a field *Investigation* that refers to *Investigation* ID); triangle terminated lines denote inheritance (*Metabolite* inherits all properties *ID*, *Name*, *Type* from *Trait*, next to its own attributes *Mass*, *Formula* and *Structure*); triangle terminated dotted lines denote use of interfaces (*Probe* 'implements' properties of *Locus*); relationships are shown both as arrows and as properties ('xref' for one-to-many, 'mref' for many-to-many relationships). Asterisks mark FuGE-derived types (for example, *Protocol**).

A growth measurement (*Data*) reports the time (*DataElement*) it took to flower (*Trait*) for an *Arabidopsis* plant (*Subject*)

A two-color microarray result (*Data*) describes raw intensities measured (*DataElement*) for gene transcript probe hybridization (*Trait*) for each pair of *Arabidopsis* individuals (*Subject*)

A marker measurement (*ProtocolApplication*) resulted in a genetic profile (*Data*) with genotype values (*DataElement*) for each SNP/microsatellite marker (*Trait*) for each human individual (*Subject*)

A genetical genomics stem cell *Investigation* was carried out on 30 recombinant mouse inbred strains (*Subject*). It involved a *ProtocolApplication* of the 'Affymetrix MG-U74Av2' *Protocol* to produce expression profiles (*Data*) for 12,422*16 microarray probes (*Traits*). These profiles consisted of a matrix of signals (*DataElement*) for each Probe (*Traits*) and each InbredStrain (*Subject*). Subsequently, these *Data* were taken as *inputData* in a normalization procedure (*ProtocolApplication*) using RMA normalization *Protocol*, which resulted in *outputData* of normalized profiles (*Data*) of Probe*InbredStrain (Trait*Subject)

RMA: robust multi-array average.

Table 2.2: Use cases of core data types.

1 commodate details on particular types of subjects and traits in a uniform
2 way. A *Trait* can be a classical phenotype (for example, flowering - the
3 flowering time is stored in the *DataElement*) or a biomolecular pheno-
4 type (for example, *Gene X* - its transcript abundance is stored in the
5 *DataElement*). A *Trait* can also be a genotype (for example, *Marker*
6 *Y* is a genomic feature observation that is stored in the *DataElement*).
7 Genomic traits such as *Gene*, *Marker* and *Probe* all need additional
8 information about their genome *Locus* to be provided. Similarly, a
Subject can be a single *Sample* (for example, a labeled biomaterial as
put on a microarray) and such a sample may originate from one partic-
ular *Individual*. It may also be a *PairedSample* when biomaterials come
from two individuals - for example, if biomaterial has been pooled as
in two-color microarrays. An individual belongs to a particular *Strain*.
When new experiments are added new variants of *Trait* and *Subject*
can be added in a similar way. Table 2.3 illustrates the generic usage
of these extended data types.

Several standard data types were also inherited from FuGE to enable
researchers to provide 'Minimum Information' for QTLs and Associa-
tion Studies such as defined in the MIQAS checklist[104] - a member
of the Minimum Information for Biological and Biomedical Investiga-
tions (MIBBI) guideline effort[335]. Data types *Action(Application)*,
Software(Application), *Equipment (Application)* and *Parameter(Value)*
can be used to describe *Protocol(Application)s* in more detail. For
example, a normalization *Protocol* may involve a 'robust multiarray av-
erage (RMA) normalization' *Action* that uses Bioconductor 'affy' *Soft-*
ware[161] with certain *ParameterValues*. Data types *Description*, *Bib-*
liographicReferences, *DatabaseEntry*, *URI*, and *FileAttachment* enable
researchers to freely add additional annotations to certain data types
- *DimensionElement*, *Investigation*, *Protocol*, *ProtocolApplication*, and
Data. For example, researchers can annotate a *Gene* with one or more
DatabaseEntries, referring to unique database accession numbers for
automated data integration.

A unique feature of XGAP is the uniform treatment of the various

Sample is a *Subject* with the additional property that 'Tissue' can be specified

Individual is a *Subject* with the additional property that relationships with Mother and Father individuals, as well as *Strain*, can be specified

PairedSample is a *Sample* with the additional property that 'Dye' has to be specified and which two Subjects (or subclasses such as *Individual*) are labeled with 'Cy3' and 'Cy5'

An *InbredStrain* is a *Strain* with the additional property that the 'Parents' (mother *Individual* and father *Individual*) are specified and the 'type' of inbreeding used

An amplified fragment length polymorphism, microsatellite or SNP *Marker* (is a *Trait*) may refer to genetic and possible genomics location (*Marker* also is a *Locus*)

A correlation computation (*Data*) reports associations (*DataElement*) between *Metabolite* (is a *Trait*); because *Trait* and *Subject* are both extensions of *DimensionElement*, they can be connected to a row and column of *DataElement* interchangeably

Table 2.3: Use cases of extended data types.

1 trait and subject annotations. The drawback of allowing users to freely
2 add additional annotations such as described above is that users and
3 tools using metabolite and gene traits, for example, would have to in-
4 spect each *Trait* instance to see whether it is actually a metabolite or
5 gene, and how it is annotated. That is why we instead use the object-
6 oriented method of 'inheritance' to explicitly add essential properties to
7 *Trait* and *Subject* variants to make sure that they are described in a
8 uniform way. For example, *Metabolite* extends *Trait*, which explicitly
adds properties ID, Name and Type (inherited from *DimensionElement*)
to metabolite specific properties Mass, Formula and Structure. See
Jones *et al.*[171] for the complete FuGE specifications and Jones and
Paton[172] for a discussion on the benefits and drawbacks of alterna-
tive mechanisms for supporting extension in object models. Table 2.4
illustrates the usage of these annotation data types.

Another feature of XGAP is the uniform treatment of all data on
these subjects and traits. To understand basic data in XGAP, newcom-
ers just have to learn that all data are stored as *Data* matrices with each
DataElement describing an observation on *Subjects* and/or *Traits* (rows
 \times columns). Unlike the proven matrix structures used in MAGE-TAB
(tabular format for microarray gene expression experiments)[282], in
XGAP these data can be on any *Trait* and/or *Subject* combination, that
is, we did not create many variants of *DataElement* to accommodate
each combination of *Trait* and *Subject* such as MAGE-TAB's *Expres-
sionDataElement* (Probe \times Sample), *MassSpecDataElement* (Mass-
Peak \times Sample), *eQTLMappingDataElement* (Marker \times Probe), and
so on. Instead, we store all these data using the generic type *DataEle-
ment* and limit extension to *Trait* and *Subject* only. This avoids the
(combinatorial) explosion of *DataElement* extensions so researchers can
provide basic data as common data matrices (of *DataElements*) and can
still add particular annotations flexibly to the matrix row and columns
to allow for (new) biotechnologies as demonstrated in the various *Trait*
extensions in Figure 2.1. Keeping this simple and uniform data structure
greatly enhances data and software (re)usability and hence productiv-

A *Gene* in an *Arabidopsis Investigation* can be connected to a *DatabaseEntry* describing a reference to related information in the TAIR database[286] and another *DatabaseEntry* describing a reference to the MIPS database[252]

Each *Individual* in a *C. elegans Investigation* is annotated with an *OntologyTerm* to indicate that it was grown in an environment of either 16°C or 24°C

The *Arabidopsis Investigation* was annotated with the *BibliographicReferences* pointing to the paper describing the investigation and expected results

A *Protocol* describes the 'MapTwoPart' method for QTL mapping and was annotated with the *URI* linking to the 'MetaNetwork R-package', which contains this method, and a *BibliographicReference* pointing to the paper[111, 250] that describes the MapTwoPart protocol

A file with a Venn diagram describing the number of masses detected in each population was added as *FileAttachement* to the *Arabidopsis* metabolite *Investigation*

Table 2.4: Use cases of annotation data types.

ity, in line with the findings by Brazma *et al.*[36] and Rayner *et al.*[282] that the simple tabular structures underlying biological data should be exploited instead of making it overly complicated.

1 After structural homogenization, such as provided by FuGE and
2 XGAP, semantic queries are the remaining major barrier for integra-
3 tion of experimental metadata. This requires ontologies that describe
4 the properties of the materials and also descriptions of experimental
5 processes, data and instruments. The former are provided by species-
6 specific ontologies that are available from various sources. The On-
7 tology for BioMedical investigation[275] may provide a solution for the
8 experimental descriptors and is being used in this context by, for ex-
ample, the Immune Epitope Database[260]. To enable researchers to
use these well understood descriptors, XGAP inherits from FuGE the
mechanism of ‘annotations’, a special field to link any data object to
one or more ontology terms. For example, researchers can annotate a
Gene with one or more *OntologyTerms* if required, referring to standard
ontology terms from OBO[314] or ontology terms defined locally.

2.3 Simple text-file format for data exchange

To enable data exchange using the XGAP model, we produced a simple text-file format (XGAP-TAB) based on the experience that for data formats to be used, data files should be easily created using simple Excel and text editor tools and closely resemble existing practices. This format is automatically derived from the model by requiring that all annotations on *Investigations*, *Protocols*, *Traits*, *Subjects*, and extensions thereof, are described as delimited text files (one file per data type) with columns matching the properties described in the object model and each row describing one data instance. Optionally, sets of *DataElements* can also be formatted as separate text matrices with row and column names matching these in the *Trait* and *Subject* annotation files, and with each matrix value matching one *DataElement*. The dimensions of each data

matrix are then listed by a row in the annotations on *Data*.

Figure 2.2 shows one investigation in the XGAP tabular data format with one delimited text file per data type - that is, there are files named 'probe.txt' and 'individual.txt', with each row describing a microarray probe or individual, respectively - and one text matrix file per set of *DataElements* - that is, there are files named 'data/expressions.txt' and 'data/genotypes.txt'. The properties of each data matrix is then described in 'data.txt'; that is, for the 'data/expressions.txt' there is a row in 'data.txt' that says that its columns refer to 'individual.txt', that its rows refer to 'probe.txt' and that its values are 'decimal'. Raw data sets and data sets in other formats can be retained in a directory labeled 'original'.

After proving its value in several proprietary projects, a growing array of public data sets are now available at[75] demonstrating the use of XGAP-TAB[148, 45, 205, 182, 324, 238].

2.4 Easy to customize software infrastructure

A pilot software infrastructure is available at[96] to help genotype-to-phenotype researchers to adopt XGAP as a backbone for their data and tool integration. We chose to use the MOLGENIS toolkit (bioinformatics generator for MOlecular GENetics Information Systems; see Materials and methods) to auto-generate from the XGAP model: 1, an SQL (Structured Query Language for relational databases) file with all necessary statements for setting up your own, customized variant of the XGAP database; 2, application programming interfaces (APIs) in R, Java and Web Services that allow bioinformaticians to plug-in their R processing scripts, Taverna workflows[311, 375, 157] and other tools; 3, a bespoke web-based graphical user interface (GUI) by which researchers can submit and retrieve data and run plugged-in tools; and 4, import/export wizards to (un)load and validate data sets exchanged

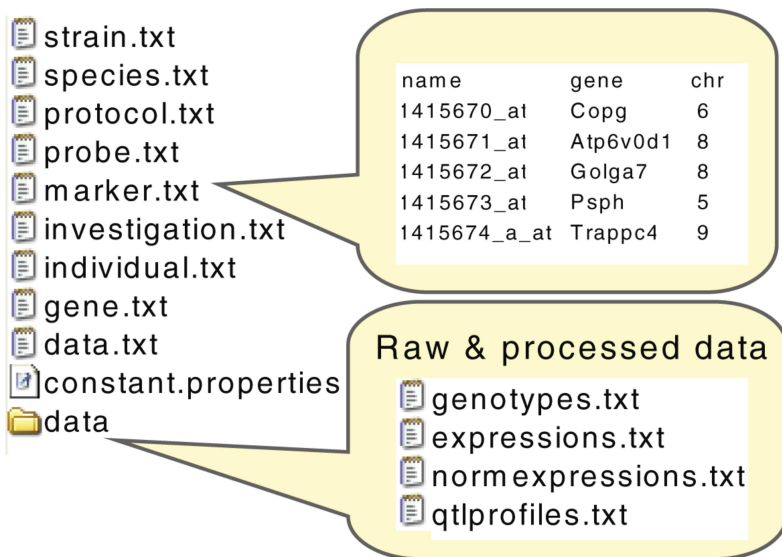


Figure 2.2: Simple text file format. A whole investigation can be stored by using easy-to-create tabular text files for annotations or matrix-shaped text files for raw and processed data. Each ‘annotation’ file relates to one data type in the object model shown in Figure 2.1 - for example, the rows in the file ‘probe.txt’ will have the columns named in data type ‘Probe’. Each ‘data’ file contains data elements and has row names and column names referring to annotation files - for example, ‘genotypes.txt’ may refer to ‘marker.txt’ names as row names and ‘individual.txt’ names as column names. If convenient, constant values can be described in the constant.properties file such as ‘species_name’.

in XGAP-TAB format. The auto-generation process can be repeated to quickly customize XGAP from an extended model, for example, to accommodate a particular new type of measurement technology or experimental design.

2.4.1 Graphical user interface

Figure 2.3 shows the GUI to upload, manage, find and download genotype and phenotype data to the database. The GUI is generated with a uniform 'look-and-feel', thereby lowering the barrier for novice users. Investigations can be described with all subjects, traits, data and protocol applications involved (1). (The numbers refer to steps in the figure.) Data can be entered using either the edit boxes or using menu-option 'file|upload' (2). This option enables upload of whole lists of traits and subjects from a simple tab-delimited format (3), which can easily be produced with Excel or R; MOLGENIS automatically generates online documentation describing the expected format (4). Subsequently, the protocol applications involved can be added with the resulting raw data (for example, genetic fingerprints, expression profiles) and processed data (for example, normalized profiles, QTL profiles, metabolic networks). These data can be uploaded, again using the common tab-delimited format or custom parsers (5) that bioinformaticians can 'plug-in' for specific file formats (for example, Affymetrix CEL files). The software behind the GUI checks the relationships between subjects, traits, and data elements so no 'orphaned' data are loaded into the database - for example, genetic fingerprint data cannot be added before all information is uploaded on the markers and subjects involved. Standard paths through the data upload process are employed to ensure that only complete and valid data are uploaded and to provide a consistent user experience.

Biologists can use the graphical user interface to navigate and retrieve available data for analysis. They can use the advanced search options (6) to find certain traits, subjects, or data. Using menu option

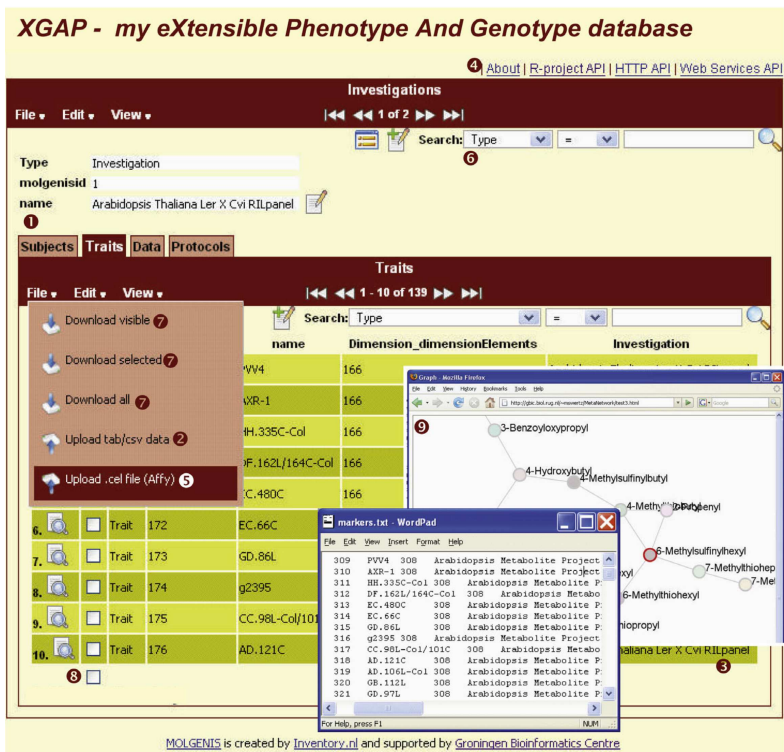


Figure 2.3: Graphical User Interfaces. A user interface enables biologists to add and retrieve data and run integrated tools. Genotype and phenotype information can be explored by investigation, subjects, traits or data. Hyperlinks following cross-references of the object model point to related information. Items indicated by 1-9 are described in the main text. See Table 2.5 for uses of this GUI. See also our online demonstrator at[96].

'file—download' (7) they can download visible/selected (8) data as tab-delimited files to analyze them in third party software. Bioinformaticians can 'plug-in' a custom-built screen (see 'customization' section) that allows processing of selected data inside the GUI, for example, visualizing a correlation matrix as a graph (9) without the additional steps of downloading data and uploading it into another tool. Biologists can create link-outs to related information, for example, to probes in GeneNetwork.org (not shown). Table 2.5 summarizes use cases of the graphical user interface.

2.4.2 Application programming interfaces

De facto standard analysis tools are emerging, for example, tools for transcript data[48, 8, 38] or metabolite abundance data[111] to mention just a few. These tools are typically implemented using the open source software for statistical analysis and graphics named R[158]. Bioinformaticians can connect their particular R or Java programs to the XGAP database using an API with similar functionality to the GUI, that is, using simple commands like 'find', 'add' and 'update' (R/API, Java/API). Scripts in other programming languages and workflow tools like Taverna[157] can use web services (SOAP/API) or a simple hyperlink-based interface (HTTP/API), for example, <http://my-xgap/api/find/Data?investigation=1> returns all data in investigation '1'. On top of this, conversion tools have been added to the R interface to read and write XGAP data to the widely used R/qlt package[38].

Figure 2.4 demonstrates how researchers can use the R/API to download (or upload) all trait/subject/data involved in their investigation from (or to) their XGAP database for (after) analysis in R. When XGAP is customized with additional data type variants, the APIs are automatically extended in the XGAP database instances by re-running the MOLGENIS generator, thus also allowing interaction with new data types in a uniform way. These new types can then be used as standard parameters for new analysis software written in R and Java. Table 2.6

1 Navigate all *Investigations*, and for each *Investigation*, see the
2 *Assays* and available *Data*

3 Select a *Gene* and find all *Investigations* in which this *Gene* is
4 regulated as suggested by significant eQTL *Data* (P -value < 0.001)

5 For a given *Locus*, select all *Genes* that have QTL *Data* mapping
6 'in *trans*'; and this may be regulated by this *Locus*, for example,
7 absolute(QTL locus - gene locus) > 10 Mb and QTL P -value $<$
8 0.001

Download a selection of raw gene expression *Data* as a
tab-delimited file (to import into other software)

Upload *Investigation* information from tab-delimited files

Upload Affymetrix *Assays* using custom *.CEL/*.CDF file readers

Plot highly correlated metabolic network *Data* in a network
visualization graph

Define security levels for *Assays/Investigations* to ensure that
appropriate data can be viewed only by collaborators, and not by
other people

A *MassPeak* has been identified to be 'proline' and we can follow
the link-out *URI* to Pubchem[275], because it was annotated to
have 'cid' 614, to find information on structure, activity, toxicology,
and more

Table 2.5: Use cases of the graphical user interface for biologists.

2.4. EASY TO CUSTOMIZE SOFTWARE INFRASTRUCTURE

In R, parse a set of tab-delimited *Marker*, *Genotype* and *Trait* files and load them into the database (R/API)

In R, retrieve all *Trait*, *Markers*, expression *Data*, and genotype *Data* from an investigation as data matrices, before QTL mapping with MetaNetwork (R/API)

In Java, retrieve a list of QTL profile correlation *Data* to show them as a regulatory network graph (J/API)

In Java, customize generated file readers to load specific file formats (J/API)

In Taverna, retrieve *Genes* from XGAP to find pathway information in KEGG (WS/API)

In Python, retrieve a list of QTL mapping *Data* using a hyperlink to XGAP (HTTP/API)

KEGG: Kyoto Encyclopedia of Genes and Genomes.

Table 2.6: Use cases of the application programming interface for bioinformaticians

summarizes use of the application programming interface.

2.4.3 Import/export wizards

A generated import tool takes care of checking the consistency of all traits, subjects and data that are provided in XGAP-TAB text files and loads them into the database. The entries in all files should be correctly linked, the data must be imported in the right order and the names and IDs need to be resolved between all the annotation files to check and link genes, microarray probes and gene expression to the

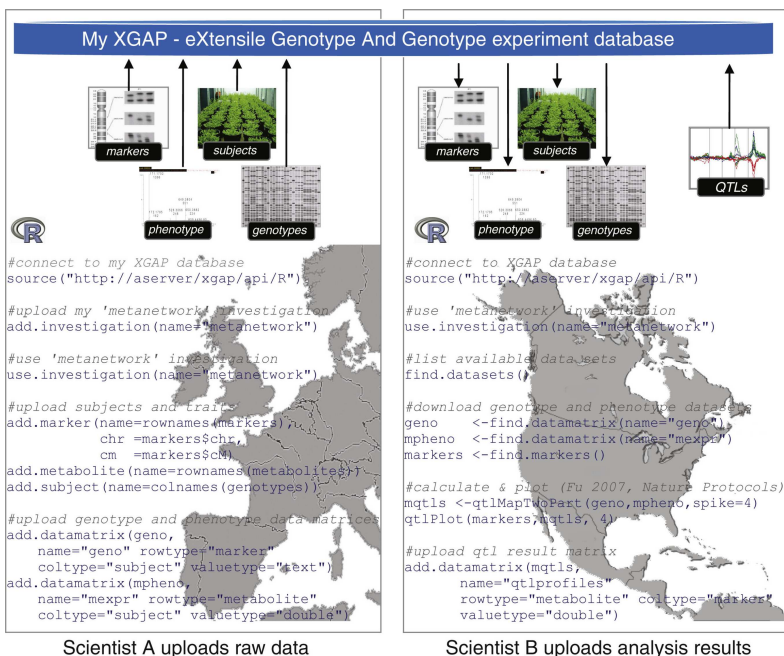


Figure 2.4: Application programming interfaces. APIs enable bioinformaticians to integrate data and tools with XGAP using web services, R-project language, Java, or simple HTTP hyperlinks. The figure shows how scientists can use the R/API to upload raw investigation data (Scientist A) so another researcher can download these data and immediately use it for the calculation of QTL profiles and upload the results thereof back to the XGAP database for use by another collaborator (Scientist B). Note how 'add.datamatrix' enables flexible upload of matrices for any *Subject* or *Trait* combination; this function adds one row to *Data* for each matrix, and as many rows to *DataElement* as the matrix has cells. See Table 2.6 for uses of these APIs.

data. The import program takes care of all these issues (conversion, relationship checks, dependency ordering, and so on). Moreover, the import program supports ‘transactions’, which ensures that all data inserts are rolled back if an import fails halfway, preventing incomplete or incorrect investigation data to be stored in the database. In a similar way, an export wizard is provided to download investigation data as a zipped directory of XGAP-TAB files.

When XGAP is customized with additional data type variants, the import/export program is automatically extended by the MOLGENIS generator, ‘future-proofing’ the data format for new biotechnological profiling platforms. Moreover, the auto-generated import program can also be used as a template for parsers of proprietary data formats, such as implemented in parsers for the PED/MAP, HapMap, and GeneNetwork data. Collaborations are underway within EBI and GEN2PHEN to also enable import/export of MAGE-TAB[282] files, the standard format for microarray experiments, of PAGE-OM[263] files, a specialized format for genome-variation oriented genotype-to-phenotype experiments, and of ISA-TAB[65] files, a generalized evolution of MAGE-TAB to represent all experimental metadata on any investigation, study and assay designed to be FuGE compatible. Also, converters to ease retrieval and submission to public repositories like dbGaP are under development. It is envisaged that integration of all these formats will enable integrated analysis of experimental data from, for example, mouse and human experiments using various biotechnology platforms, which was previously near impossible for biological labs to implement.

2.4.4 Customizing XGAP

Customizations and extensions of the XGAP object model can be described in a single text file using MOLGENIS[329, 327] DSL. On the push of a button, the MOLGENIS generator instantly produces an extended version of the XGAP database software from this DSL file. A regression test procedure assists XGAP developers to ensure their ex-

1 tensions do not break the XGAP exchange format. Figure 2.5a shows
2 how the addition of a *Metabolite* data entity as a new variant of *Trait*
3 takes only a few lines in this DSL. Figure 2.5b shows how the GUI can
4 be customized to suit a particular experimental process. Figure 2.5c
5 shows how programmers can add a ‘plug-in’ program that is not gener-
6 ated by MOLGENIS but written by hand in Java (for example, a viewer
7 that plots QTL profiles interactively). Moreover, use of Cascading Style
8 Sheets (CSS) enables research projects to completely customize the look
and feel of their XGAP.

All XGAP and MOLGENIS software can be downloaded for free
under the terms of the open source license LGPL. Extended documen-
tation on XGAP and MOLGENIS customization is available online at
the XGAP and MOLGENIS wikis[96, 103].

2.5 Conclusions

In this paper we report a minimal and extensible data infrastructure for
the management and exchange of genotype-to-phenotype experiments,
including an object model for genotype and phenotype data (XGAP-
OM), a simple file format to exchange data using this model (XGAP-
TAB) and easy-to-customize database software (XGAP-DB) that will
help groups to directly use and adapt XGAP as a platform for their
particular experimental data and analysis protocols.

We successfully evaluated the XGAP model and software in a broad
range of experiments: array data (gene expression, including tiling arrays
for detection of alternative splicing, ChIP-on-chip for methylation, and
genotyping arrays for SNP detection); proteomics and metabolomics
data (liquid chromatography time of flight mass spectrometry (LC-
QTOF MS), NMR); classical phenotype assays[148, 45, 205, 183, 324,
238, 21, 25]; other assays for detection of genetic markers; and an-
notation information for panel, gene, sample and clone. Nontechnical
partners successfully evaluated the practical utility by independently

(a) Create new data types using 'extends' (Figure 1)

```
<entity name="Metabolite" extends="Trait">
  <field name="Formula" nillable="true"
    description="The chemical formula." />
  <field name="Mass" type="decimal"
    nillable="true"
    description="The mass." />
  <field name="Structure" type="text"
    nillable="true"
    description="The chemical structure." />
</entity>
```

(b) Customize the GUI using 'form' and 'menu'

```
<form name="Investigations"
  entity="Investigation">
  <menu name="InvestigationMenu">
    <form name="Subjects"
      entity="Subject" />
    <form name="Traits"
      entity="Trait" />
    ...
  </menu>
</form>
```

(c) Plug-in hand-written components

```
<form ...
  <plugin name="myplugin"
    type="plugins.NetworkViewer"/>
</form>
```

Figure 2.5: Customizing XGAP. A file in MOLGENIS domain-specific language is used to describe and customize the XGAP database infrastructure in a few lines. **(a)** Shows how the addition of a *Metabolite* data entity as a new variant of *Trait* takes only a few lines in this DSL. **(b)** Shows how the GUI can be customized to suit a particular experimental process. **(c)** Shows how programmers can add a 'plug-in' program that is not generated by MOLGENIS but written by hand in Java.

1 formatting and loading parts of their consortium data: EU-CASIMIR
2 (for mouse; Table 2.7), EU-GEN2PHEN (for human; Table 2.7), EU-
3 PANACEA (for *C. elegans*) and IOP-Brassica (for plants). A public
4 subset of these data sets is available for download at[96]. When needed
5 we could quickly add customizations to the model, building on the gen-
6 eral schema, and then use MOLGENIS to generate a new version of
7 the software at the push of a button, for example, to support *NMR*
8 methods as an extended type of *Trait*[110]. Furthermore we success-
fully integrated processing tools, such as a two-way communication
with R/QTL[38] enabling QTL mapping on XGAP stored genotypes
and phenotypes with QTL results stored back into XGAP.

Based on these experiences, we expect use of XGAP to help the
community of genome-to-phenome researchers to share data and tools,
notwithstanding large variations in their research aims. The XGAP data
format can be used to represent and exchange all raw, intermediate and
result data associated with an investigation, and an XGAP database,
for instance, can be used as a platform to share both data and compu-
tational protocols (for example, written in the R statistical language)
associated with a research publication in an open format. We envision
a directory service to which XGAP users can publish metadata on their
investigations either manually or automatically by configuring this op-
tion in the XGAP administration user interface. This directory service
can then be used as an entry point for federated querying between the
community of XGAPs to share data and tools.

Groups that already have an infrastructure can assimilate XGAP to
ease evolution of their existing software. Next to their existing user
tools, they can 'rewire' algorithms and visual tools to also use the
MOLGENIS APIs as data backend. Thus, researchers still have the
same features as before, plus the features provided by the generated
infrastructure (for example, data management GUIs, R/API) and con-
nected tools (for example, R packages developed elsewhere). Moreover,
much less software code needs to be maintained by hand when replacing
hand-written parts by MOLGENIS-generated parts, allowing software

Consortium	Remit
CASIMIR	<p>The collection and distribution of large volumes of complex data typical of functional genomics is carried out by an increasing number of disseminated databases of hugely variable scale and scope. Combined analysis of highly distributed datasets provides much of the power of the approach of functional genomics, but depends on databases' ability to exchange data with each other and on analytical tools with semantic and structural integrity. Agreement on the standards adopted by databases will inevitably be a matter of community consensus and to that end a recent coordination action funded by the European Commission, CASIMIR[64], is engaged in a community consultation on the nature of the technical and semantic standards needed. What has already become clear in use-case studies conducted so far is that whatever standards are adopted, they will inevitably remain dynamic and continue to develop, particularly as new data types are collected. Crucially, they should allow the open-ended development of analytical and datamining software, while integration of efforts to agree such standards and develop new software is essential.</p>
GEN2PHEN	<p>Currently available genotype-to-phenotype (G2P) databases are few and far between, have great diversity of design, and limited or no interoperability between them. This arrangement provides no convenient way to populate the databases, no easy way to exchange, compare or integrate their content, and absolutely no way to search the totality of gathered information. In this context, the European Commission has recently funded the GEN2PHEN project[65], which intends to significantly improve the database infrastructure available within Europe for the collation, storage, and analysis of human and model-organism G2P data. This will be achieved by first developing various cutting-edge solutions, and then deploying these in conjunction with proven concepts, so as to transform the current elementary G2P database reality into a powerful networked hierarchy of interlinked databases, tools and standards.</p>

1
2
3
4
5
6
7
8

Table 2.7: XGAP participating consortia.

engineers to add new features for researchers much more rapidly.

We invite the broader community to join our efforts at the public XGAP.org wiki, mailing list and source code versioning system to evolve and share the best XGAP customizations and GUI/API ‘plug-in’ enhancements, to support the growing range of profiling technologies, create data pipelines between repositories, and to push developments in the directions that will most benefit research.

2.6 Materials and methods

Software modeling, auto-generation/configuration and component toolboxes are increasingly used in bioinformatics to speed up (bespoke) biological software development; see our recent review[329]. For XGAP we required a software toolbox providing query interfaces, data management interfaces, programming interfaces to R and web services, simple data exchange formats and a minimal requirement of programming knowledge. The MOLGENIS modeling language and software generator toolbox[329, 103] was chosen as it combines all these features.

Several alternative toolboxes were evaluated: BioMart[103, 312] and InterMine[214] generate powerful query interfaces for existing data but are not suited for data management; Omixed[246] generates programmatic interfaces onto databases, including a security layer, but lacks user interfaces; PEDRO/Pierre[166] generates data entry and retrieval user interfaces but lacks programmatic interfaces; and general generators such as AndroMDA[12] and Ruby-on-Rails[247] require much more programming/configuration efforts compared to tools specific to the biological domain. Turnkey[250] seemed to be closest to our needs: it emerged from the GMOD community having GUI and SOAP interfaces but lacks auto-generation of R interfaces and a file exchange format.

Figure 2.6 summarizes how MOLGENIS generates the XGAP database software in three layers: database, API and GUI. MOLGENIS either generates a high-performance ‘server’ edition, which requires installa-

tion on server software, or a limited 'standalone' edition that runs on a desktop computer without any additional configuration. The database layer is generated as SQL files with 'database CREATE statements' that are loaded into either MySQL (server), PostgreSQL (server) or HSQLDB (standalone). Each data type in the XGAP object model (Figure 2.1) is mapped to its own table - for example, there is a 'Trait' table. Each inheritance adds another table, for example, each *Gene* has an entry in the 'Gene' table and also in the 'Trait' table. One-to-many crossreferences between data types are mapped as foreign keys - for example, *Data* has a numeric field called 'Investigation' that must refer to the *foreign key* 'molgenisid' of *Investigation*. Many-to-many cross-references are mapped via a 'link-table' - for example, an additional table '*mref_import_data*' is generated for two foreign keys to *Data* and *ProtocolApplication*, respectively, to model the *importData* relationship between them. The API layer is generated as Java files either served via Tomcat (server) or Jetty (standalone). A Java class is generated for each data type - for example, there is a class *Gene*. All data can be queried programmatically via a central *Database* class, that is, command *db.find(Gene.class)* returns all *Gene* objects in the database. To enhance performance, the API uses the 'batched' update methods of Java's DataBase Connectivity (JDBC) package and the 'multi-row-syntax' of MySQL to allow inserts of 10,000s of data entries in a single command, an optimization that is 5 to 15 times quicker than standard one-by-one updates. The Java/API is exposed with a SOAP/API, HTTP/API and R/API, so XGAP can also be accessed via web service tools like Taverna, HTTP or R, respectively (accessible via hyperlinks in the GUI). The GUI layer is also generated as Java files. The GUI includes classes for each Menu and Form - for example, the *InvestigationForm* class generates a view- and editform for investigations in the GUI. The generation is steered from one XML file written in MOLGENIS DSL (partially shown in Figure 2.5). To enable FuGE extension, the FuGE model was automatically translated into MOLGENIS DSL. We therefore first downloaded the FuGE v1 MagicDraw file

1 from[235], exported from MagicDraw to XMI 2.1, parsed the XMI using
2 the EMF parser from Eclipse[267] and then automatically translated it
3 into MOLGENIS DSL using a newly built XmiToMolgenis tool. Com-
4 patibility with the FuGE standard is ensured via inheritance; that is,
5 *Investigation*, *Protocol*, *ProtocolApplication*, *Data* and *DimensionEle-*
6 *ment* in XGAP all extend FuGE data types of the same name. Further
7 implementation details can be found at [96, 103].

Abbreviations

4 API: application programming interface; dbGaP: database of genotypes
5 and phenotypes; DSL: domain-specific computer language; FuGE: Func-
6 tional Genomics Experiment model; GMOD: Generic Model Organism
7 Database; GUI: graphical user interface; LC/MS: liquid chromatography-
8 mass spectrometry; MAGE-TAB: tabular format for microarray gene
expression experiments; MOLGENIS: biosoftware generator for MOlecular
GENetics Information Systems; NMR: proton nuclear magnetic res-
onance; QTL: quantitative trait locus; SOAP: web services using simple
object access protocol; SQL: Structured Query Language for relational
databases; XGAP: eXtensible Genotype And Phenotype platform.

Acknowledgements

The authors thank CASIMIR (funded by the European Commission under contract number LSHG-CT-2006-037811,[64]; Table 2.7), and GEN2PHEN, a FP7 project funded by the European Commission (FP7-HEALTH contract 200754,[65]; Table 2.7). The authors also thank NWO (Rubicon Grant 825.09.008) for financial support.

Authors' contributions

MAS, ARJ, PS, KS, JMH, DS, EOB, HEP and RCJ compiled the functional requirements for the XGAP community platform and drafted the

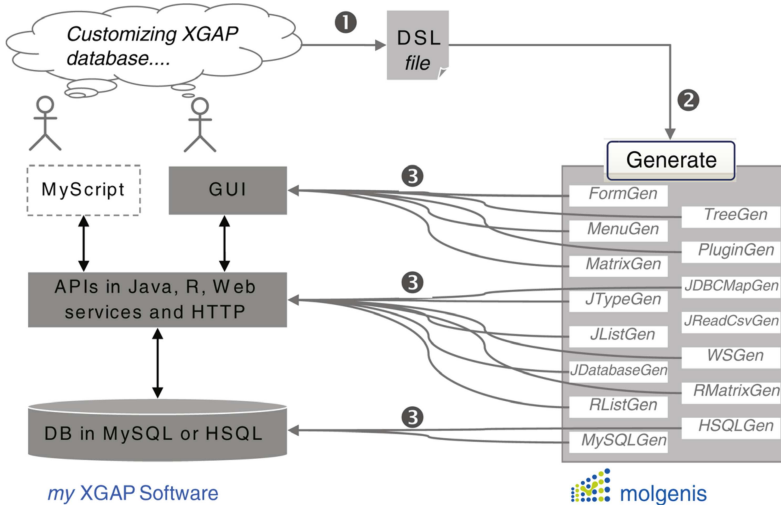


Figure 2.6: Auto-generation of XGAP software. Open source generator tools are used to produce a customized XGAP software infrastructure. 1, The XGAP object model is described using the MOLGENIS' little modeling language (Figure 2.4). 2, Central software termed MolgenisGenerate runs several generators, building on the MOLGENIS catalogue of reusable assets. 3, At the push of the button, the software code for a working XGAP implementation is automatically generated from the DSL file. GUI and APIs provide simple tools to add and retrieve data, while the reusable assets of MOLGENIS hide the complexity normally needed to implement such tools. For customization, only simple changes to the XGAP model file are required; the MOLGENIS generator takes care of rewriting all the necessary files of SQL and Java software code, saving time and ensuring a consistent quality.

CHAPTER 2. XGAP MODEL FOR GENOTYPE AND PHENOTYPE

1 extensible data model. MAS, KJV, BMT, RAS, and MD refined and
2 implemented the model using MOLGENIS, and added all parsers, and
3 user interfaces. MAS and KW implemented Taverna compatible web
4 services and GV, DA, KJV and MS implemented R-services. MAS,
5 HEP and RCJ drafted the manuscript. All authors evaluated XGAP
6 components in various settings. All authors read and approved the final
7 manuscript.
8

Chapter 3

xQTL workbench: A scalable web environment for multi-level QTL analysis

Bioinformatics. 2012 Apr 1;28(7):1042-4.

DOI: 10.1093/bioinformatics/bts049

PubMed ID: 22308096

Danny Arends^{1,†}, K. Joeri van der Velde^{1,†}, Pjotr Prins^{1,2}, Karl W. Broman³, Steffen Möller⁴, Ritsert C. Jansen¹ and Morris A. Swertz^{1,5,6,*}

1. Groningen Bioinformatics Centre, University of Groningen, Groningen.
2. Laboratory of Nematology, Wageningen University, Wageningen, The Netherlands.
3. Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA.
4. Institut für Neuro- und Bioinformatik, Universität zu Lübeck.
5. Genomics Coordination Centre, University Medical Centre Groningen, University of Groningen, The Netherlands.
6. EMBL-EBI, the European Bioinformatics Institute, Hinxton, UK.

Associate Editor: Jeffrey Barrett

Received on September 30, 2011; revised on December 19, 2011; accepted on January 20, 2012

* To whom correspondence should be addressed.

† The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Abstract

Summary: xQTL workbench is a scalable web platform for the mapping of quantitative trait loci (QTLs) at multiple levels: for example gene expression (eQTL), protein abundance (pQTL), metabolite abundance (mQTL) and phenotype (phQTL) data. Popular QTL mapping methods for model organism and human populations are accessible via the web user interface. Large calculations scale easily on to multi-core computers, clusters and Cloud. All data involved can be uploaded and

queried online: markers, genotypes, microarrays, NGS, LC-MS, GC-MS, NMR, etc. When new data types come available, xQTL workbench is quickly customized using the Molgenis software generator.

Availability: xQTL workbench runs on all common platforms, including Linux, Mac OS X and Windows. An online demo system, installation guide, tutorials, software and source code are available under the LGPL3 license from <http://www.molgenis.org/xqtl>¹.

Contact: m.a.swertz@rug.nl

3.1 Introduction

Modern high-throughput technologies generate large amounts of genomic, transcriptomic, proteomic and metabolomic data. However, existing open source web-based tools for QTL analysis, such as webQTL [358] and QTLNetwork [377], are not easily extendable to different settings and computationally scalable for whole genome analyses. xQTL workbench makes it easy to analyse large and complex datasets using state-of-the-art QTL mapping tools and to apply these methods to millions of phenotypes using parallelized 'Big Data' solutions [342]. xQTL workbench also supports storing of raw, intermediate and final result data, and analysis protocols and history for reproducibility and data provenance. Use of Molgenis [328] helps to customize the software. All is conveniently accessible via standard Internet browsers on Windows, Linux or Mac (and using Java, R for the server).

3.2 Features

xQTL workbench provides visualization of QTL profiles, single and multiple QTL mapping methods, easy addition of new QTL analyses, scalable data management and analysis tracking.

¹The URL in the original paper is no longer active and was updated here.

3.2.1 Explore QTL profiles

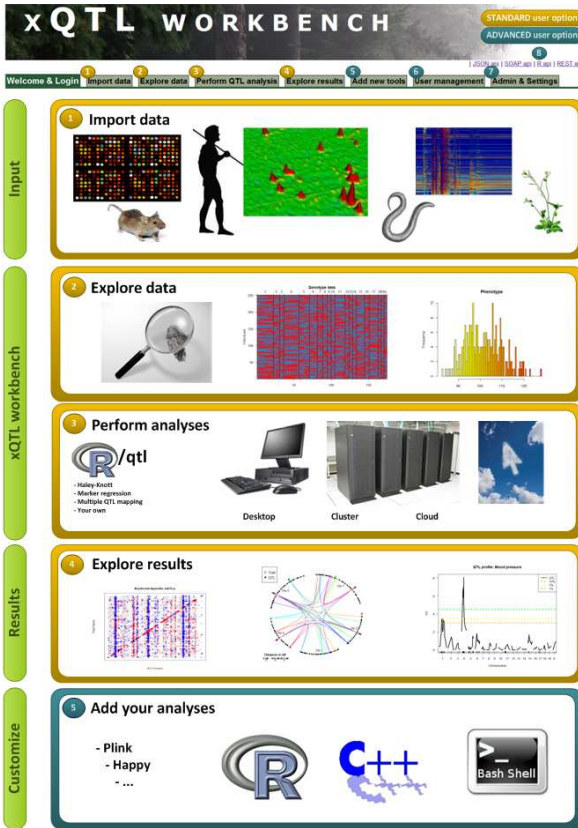
1 Through the web interface, users can explore mapped QTLs, and under-
2 lying information by viewing QTL plots in an interactive scrollable and
3 zoomable window. xQTL workbench has support for other common im-
4 age formats, such as PNG, JPG, SVG and embedded postscript; useful
5 for publishing scientific results online, and on paper. From the output,
6 main database identifiers, such as gene, protein and/or metabolite iden-
7 tifiers are automatically linked-out to matching external web pages of
8 public database such as NCBI, KEGG and Wormbase.

3.2.2 Single and multiple QTL mapping

xQTL workbench wraps R/qtl [15, 38] in a web-based analysis frame-
work offering all important QTL mapping routines, such as the EM algo-
rithm, imputation, Haley-Knott regression, the extended Haley-Knott
method, marker regression and Multiple QTL mapping. In addition,
xQTL workbench includes R/qtlbim, a library that provides a Bayesian
model selection approach for mapping multiple interacting QTL [376]
and Plink, a library for association QTL mapping on single nucleotide
polymorphisms (SNP) in natural populations [277].

3.2.3 Add new analysis tools

xQTL workbench supports flexible adding of more QTL analysis soft-
ware: any R-based, or command-line tool, can be plugged in. All anal-
ysis results are uploaded, stored and tracked in the xQTL workbench
database through an R-API. When new tools are added, they can build
on the high-level multi-core computer, cluster and Cloud management
functions, based on TORQUE/OpenPBS and BioNode [273]. xQTL
workbench can be made part of a larger analysis pipeline using inter-
faces to R, Excel, REST and SOAP web services and Galaxy [128].



1
2
3
4
5
6
7
8

Figure 3.1: Screenshot of xQTL workbench with all features enabled; (1) import phenotype, genotype and genetic map data, examples are given per import type; (2) search through the whole database, explore and browse your data using molgenis generated web-interfaces; (3) run R/qtl QTL mapping, the general plugin allows users to perform not only QTL mapping but also other analyze; (4) use default (or custom) plugins to explore results (e.g. Heatmaps, QTL profiles); (5) add new tools to the workbench (for Bio informaticians); (6) user management and access control of the system (Only for admins); (7) expert settings can be altered in the admin tab (Only for admins); (8) connect/share data using generated API's to R statistics, REST/JSON, SOAP. 71

3.2.4 Track analysis and monitor performance

1 When a new analysis protocol or R script is defined, this protocol can
2 easily be applied to new data. Also, xQTL workbench keeps track of his-
3 tory. Re-use of analysis protocols can be done in an automated fashion.
4 Previous analyses can be rerun without resetting parameters. xQTL
5 workbench provides an online overview of past analyses e.g. which
6 analyses were performed, by who, when and display settings applied.

3.2.5 Scalable data management

4 xQTL workbench has a consistency checking database based on XGAP
5 specification [330], user interfaces to manage and query genotype and
6 phenotype datasets and support for various database back-ends includ-
7 ing HSQL (standalone) and MySQL. Phenotype, genotype and genetic
8 map data can be imported as text (TXT), comma separated (CSV)
and Excel files. xQTL workbench handles and stores large data in a
new and efficient binary edition of the XGAP format, named XGAPbin
(extension .xbin), documented online. Such binary formats are essential
when handling, storing and transporting multi-Gigabyte datasets.

3.2.6 Customizable to research needs

Additional modules for new data modalities can be added using Mol-
genis software generator [330]. The 'look and feel' of xQTL work-
bench is adaptable to institute or consortium style by changing a simple
template, which is described in the xQTL workbench documentation
enabling seamless integration into an existing website or intranet site,
such as recently for EU-PANACEA model organism project and Life-
Lines biobank.

3.3 Implementation

We built xQTL workbench on top of Molgenis [327], a Java-based software to generate tailored research infrastructure on-demand [329]. From a single ‘blueprint’ describing the whole system, Molgenis auto-generates a full application including user interface, database infrastructure, application programming interfaces in R, REST and SOAP (APIs). Molgenis’ flexibility and robustness is proven by the wide range of research projects, e.g. the Nordic GWAS Control database [198], EB mutation database [343] and the Animal observation database [328].

For data storage, the eXtensible Genotype and Phenotype (XGAP) data model was adopted [330] and extended for big data. To support the increased demand for computational resources for included mapping routines, we added high-level cluster and cloud management functions for computation. The scalable QTL mapping routines of xQTL workbench are written in R and C. The choice of R ties in with the general practice of using R for QTL mapping. The user interface includes direct access to the R interpreter. Both xQTL workbench and Molgenis are open-source software, and source code is transparently stored and tracked in online source control repositories.

3.4 Conclusion

xQTL workbench provides a total solution for web-based analysis: major QTL mapping routines are integrated for use by experienced and inexperienced users. Researchers can upload raw data, run analyses, explore mapped QTL and underlying information, and link-out to important databases. New algorithms can be flexibly added, immediately available to all users. Large analyses can be easily executed on a cluster or in the Cloud. Future work include visualizations and search options to explore the results. We also had an EU-SYSGENET workshop that envisioned further integration of xQTL with analysis tools like HAPPY, databases like GeneNetwork, and the workflow manager TIQS [86].

Acknowledgements

We thank Konrad Zych for Figure 3.1.

1
2
3
4
5
6
7
8

Funding: National Institutes of Health (GM074244 to KB); Netherlands Organisation for Scientific Research (NWO)/TTI Green Genetics (1CC029RP to P.P.); NWO (Rubicon 825.09.008 to M.A.S), Centre for BioSystems Genomics (CBSG), Netherlands Consortium of Systems Biology (NCSB) (to D.A.), Netherlands Bioinformatics Center (NBIC) (to M.A.S.), all part of Netherlands Genomics Initiative/NWO; Target/LifeLines co-funded by the European Regional Development Fund and NWO (to M.A.S.); and EU-FP7 Projects PANACEA (222936 to K.J.v.d.v.) and EURATRANS (241504 to R.C.J.).

Conflict of Interest: none declared.

Chapter 4

**WormQTL^{HD}: A web
database for linking
human disease to natural
variation data in
*C. elegans***

Nucleic Acids Res. 2014 Jan;42(Database issue):D794-801.
DOI: 10.1093/nar/gkt1044
PubMed ID: 24217915

K. Joeri van der Velde^{1,2,3}, Mark de Haan^{1,2,3,4}, Konrad Zych², Danny Arends², L. Basten Snoek⁵, Jan E. Kammenga⁵, Ritsert C. Jansen², Morris A. Swertz^{1,2,3,*} and Yang Li^{2,3,*}

1. Genomics Coordination Center, University of Groningen, University Medical Center Groningen, P.O. Box 30001, 9700 RB Groningen, The Netherlands.

2. Groningen Bioinformatics Center, University of Groningen, P.O. Box 11103, 9700 CC Groningen, The Netherlands.

3. Department of Genetics, University of Groningen, University Medical Center Groningen, P.O. Box 30001, 9700 RB Groningen, The Netherlands.

4. Department of Bioinformatics, Hanze University of Applied Sciences, Groningen, Zernikeplein 11, 9747 AS, The Netherlands.

5. Laboratory of Nematology, Wageningen University, 6708 PB Wageningen, The Netherlands.

Received August 14, 2013; Revised October 9, 2013; Accepted October 10, 2013

*To whom correspondence should be addressed. Tel: +31 50 367100; Fax: +31 50 361 7230; Email: m.a.swertz@rug.nl

Correspondence may also be addressed to Yang Li. Tel: +31 50 367100; Fax: +31 50 361 7230; Email: yang.li@rug.nl

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Interactions between proteins are highly conserved across species. As a result, the molecular basis of multiple diseases affecting humans can

be studied in model organisms that offer many alternative experimental opportunities. One such organism —*Caenorhabditis elegans*— has been used to produce much molecular quantitative genetics and systems biology data over the past decade. We present WormQTL^{HD} (Human Disease), a database that quantitatively and systematically links expression Quantitative Trait Loci (eQTL) findings in *C. elegans* to gene-disease associations in man. WormQTL^{HD}, available online at <http://www.wormqtl-hd.org>, is a user-friendly set of tools to reveal functionally coherent, evolutionary conserved gene networks. These can be used to predict novel gene-to-gene associations and the functions of genes underlying the disease of interest. We created a new database that links *C. elegans* eQTL data sets to human diseases (34,337 gene-disease associations from OMIM, DGA, GWAS Central and NHGRI GWAS Catalogue) based on overlapping sets of orthologous genes associated to phenotypes in these two species. We utilized QTL results, high-throughput molecular phenotypes, classical phenotypes and genotype data covering different developmental stages and environments from WormQTL database. All software is available as open source, built on MOLGENIS and xQTL workbench.

4.1 Introduction

Many exciting data sets have been collected in recent years for *Caenorhabditis elegans*, a free-living, non-parasitic soil-related nematode that feeds on the bacteria of decaying organic matter. This worm has many useful features that have made it one of the most studied model organisms: it is small and easy to house, has a short generation time and is transparent. As a consequence, its genomic information is now available [91], and the developmental path and function of almost every cell in its body has been described [122]. In addition, recent genetical genomics studies in *C. elegans* have revealed thousands of genomic regions (loci) that are associated to the quantitative variation in a diverse range of

phenotypes, such as gene expression [expression Quantitative Trait Loci (eQTLs)] [114, 178, 254, 203, 350, 351, 293], lifespan [83], development [139, 140, 177], stress resistance [295, 147], behaviour [227, 283], dauer formation [147, 133] and sensitivity to RNAi treatments [92].

Genes having eQTLs mapping to the same genomic region (i.e. hotspot) are possibly involved in the same biological pathway/process. Palopoli *et al.* [254] have shown that biochemical processes and molecular functions of genes are generally highly conserved. Lee *et al.* [194] have shown that using the OMIM database [142] (<http://omim.org/>) and orthologue mapping data from INPARANOID [248], it is possible to infer new gene-gene interactions that are responsible for a certain disease in man from model organism data. McGary *et al.* [226] have shown that the conservation level between *C. elegans* and man is sufficient to infer gene-gene interactions in man from worm data. Even though the global disease phenotypes may not be at all comparable, the molecular basis may be common (e.g. breast cancer and high male incidence of progeny). For example, research on stress response in *C. elegans* has provided detailed insight into the genetic and molecular mechanisms underlying complex human diseases [294]. In addition, Shaye and Greenwald [307] have generated a compendium of *C. elegans* genes with human orthologues using four orthology prediction programmes for identifying *C. elegans* orthologues of human disease genes for potential functional analysis. As a result, linking *C. elegans* and human data could help to understand the mechanisms underlying many human diseases.

To facilitate the exploitation of the worm eQTL data for human disease research we developed a new database, WormQTL^{HD}, which quantitatively and systematically links many eQTLs findings in *C. elegans* to gene-disease associations in human. The database is based on the detection of the overlapping sets of orthologous genes associated with different phenotypes, or ‘phenologs’ [226] between these two species. The data, mainly eQTL results, were taken from different platforms (e.g. Agilent) and experiments (e.g. developmental stages). We

provide a set of web-based analysis tools to search the database and explore phenotypes based on gene orthologues between worm and man. The result can be downloaded and visualized in a comprehensive yet clear way. All data and tools can be accessed via a public web user interface, as well as basic programming interfaces, which were built using the MOLGENIS biosoftware toolkit [328].

To our knowledge, this is the first online database for the systematic investigation of *C. elegans* phenotype equivalents of human diseases by integrating known disease-gene associations, gene orthologue data, molecular phenotypes and QTL results. WormQTL^{HD} allows researchers to explore these complex data in a user-friendly way, finding new genes, interactions and loci for human disease models.

WormQTL^{HD} is freely accessible without registration and is hosted at <http://www.wormqtl-hd.org>. All underlying software is open source and can be downloaded and freely used, for example, as a local mirror of the database and/or to host new studies, which can be uploaded using XGAP format [330]. Below we describe the results, methods used to implement the system and future plans.

4.2 Implementation

WormQTL^{HD} was compiled using data from six sources that are described below: (I) WormQTL [315]¹, (II) WormBase Phenotypes ([379], (III) Online Mendelian Inheritance in Man (OMIM) [142], (IV) The Disease and Gene Annotations (DGA) [259], (V) NHGRI GWAS Catalogue [152] (<http://www.genome.gov/gwastudies>) and (VI) GWAS Central [339, 39] (Figure 4.1). (I) WormQTL (<http://www.wormqtl.org>) contains many published ‘genetical genomics’ experiments and consists of 47 public data sets with eQTL data on 500 panels (Recombinant Inbred Lines or natural strains), 68,452 microarray probes, 1,630 samples and 1,579 markers. The tools that were present in Wor-

¹The original paper erroneously cited [294, 352] here.

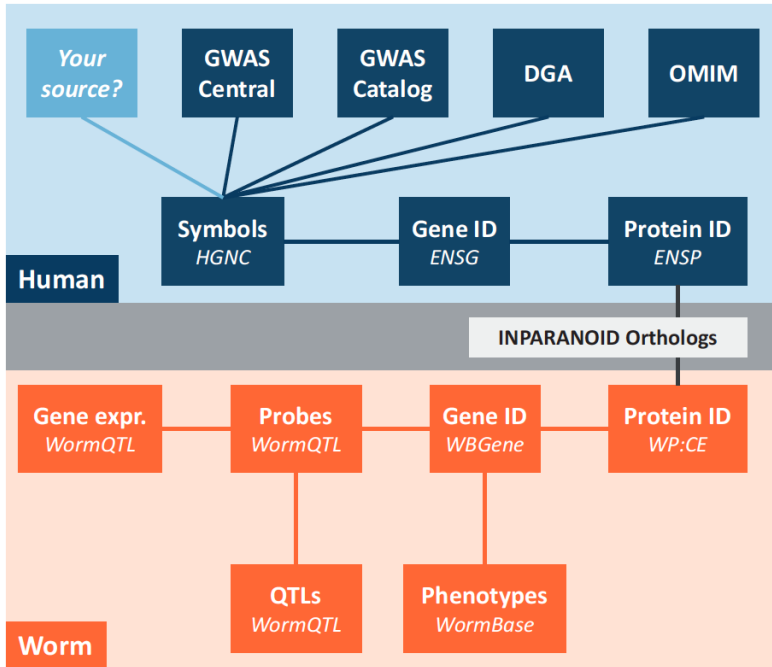


Figure 4.1: Human and worm data integration. WormQTL^{HD} was compiled using data derived from WormQTL, WormBase, OMIM, DGA, GWAS Catalogue and GWAS Central.

mQTL, such as the QTL Finder and the Genome Browser, are also available in WormQTL^{HD}. (II) WormBase is ‘an international consortium of biologists and computer scientists dedicated to providing the research community with accurate, current, accessible information concerning the genetics, genomics and biology of *C. elegans* and related nematodes’ (WormBase Mission statement, Todd Harris, 26 November 2012). From WormBase, we downloaded all the gene-phenotype associations (total 227,216) via WormMart. (III) OMIM is one of the most popular databases containing 14,164 human gene-disease associations. (IV) The DGA database (2,961 associations) was started in 2013 and claims to be more comprehensive than OMIM. (V) The NHGRI GWAS Catalogue is a collection of 12,925 SNP-to-disease associations published in GWAS studies with at least 100,000 assayed SNPs and a *P*-value of $<1.0 \times 10^{-5}$. The SNPs were linked to human genes by the authors of the original papers that have been included in the catalogue. (VI) GWAS Central [339, 39, 108] is a database of summary level findings from genetic association studies. The authors of GWAS Central gathered and curated many datasets from public domain projects, and supplied us with a list of 4,487 gene-disease associations having a *P*-value of $<1.0 \times 10^{-10}$. Because of the non-overlapping information in these four sources of human genes linked to disease, they are all provided and can be selected by the user. Human and worm data are connected based on the detection of orthologous genes in these two species. We downloaded all INPARANOID orthologues between *C. elegans* and *Homo sapiens* with a 100% bootstrap value. The bootstrap value indicates how often the pair is found as reciprocally best matched in a sampling with a replacement procedure that was applied to the original Blast alignment.

To explore this database, WormQTL^{HD} features four major searching tools for different purposes. The starting points are summarized in Figure 4.2 and described in detail below, followed by a short summary of the software used.

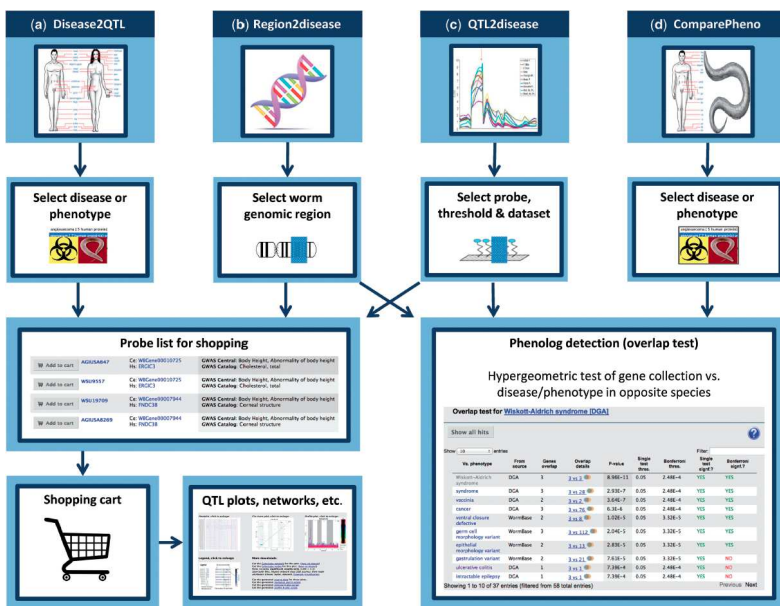


Figure 4.2: Cross-experiment search. WormQTL^{HD} provides four tools to explore the database: mapping human diseases to worm QTLs (Disease2QTL); mapping a worm genomic region to human diseases (Region2disease); mapping worm QTLs to human diseases (QTL2disease); and linking worm phenotypes to human diseases (ComparePheno).

4.2.1 Tool 1: ‘Disease2QTL’, mapping human diseases to worm eQTLs

Exploring the genetic variation data for human gene orthologues in worm can provide useful insight into the function and regulation of human diseases. WormQTL^{HD} provides a tool for human geneticists to explore novel causal genes for a specific human disease by using worm QTL findings. Using a selection of one or multiple human diseases (from OMIM, DGA, NHGRI GWAS Catalogue or GWAS Central), a ‘shopping’ page is presented with worm gene expression probes and their human disease association. More information about the gene orthology mapping and association studies can be browsed. Users can put individual probes, or all probes at once, into the ‘shopping cart’. Subsequently, they can explore the genetic variation of those genes across the different experiments and studies that are stored in the WormQTL^{HD} database. The shopping cart is a central place in WormQTL^{HD} where users can see the various worm gene probes that they have selected, and create QTL/eQTL visualizations from the items in the shopping cart using ‘Plot QTLs’.

Using the ‘Plot QTLs’ function, researchers can test if genes associated with the selected diseases have any QTLs and if they map to a common genomic region. Shared QTLs suggest that those genes are regulated by the same genetic variation and are possibly involved in the same biological pathways. The genes with *cis*-QTLs in that genomic region are used as candidate genes in several types of studies [182, 336, 316]. The same approach can be used for causal genes of human diseases. Alternatively, users can also select worm phenotypes (1,504 total) instead of human diseases as a starting point. The shopping window is presented in exactly the same way as before, so users can browse human diseases from a worm phenotype perspective instead, or simply shop for probes of choice for a given worm phenotype and plot their QTLs, without considering any human disease relation.

4.2.2 Tool 2: ‘Region2disease’, mapping worm genomic regions to human diseases

1 Researchers can link worm genomic regions to human diseases. This
2 approach starts by selecting a region in the worm genome, e.g. a known
3 ‘eQTL hotspot’, where a number of eQTLs are located. The region is
4 selected by providing the chromosome name, start and end base pair
5 positions. Users can quickly define a region of interest by using the
6 location of any *C. elegans* gene. The database then returns all worm
7 gene expression probes that are annotated in this region. From the
8 probes, the corresponding worm genes are gathered, plus their human
orthologues. The user is presented with a table containing the human-
worm orthology and disease/phenotype associations in man and worm.
After shopping for some or all of the relevant probes, users can choose
to visualize eQTL results for them (similar to Tool 1), or perform a
disease enrichment test.

The hypergeometric gene overlap test [291] to discover phenologs
(phenotype orthologues) can be performed by clicking on ‘Disease en-
richment’. All probes in the region are linked to their corresponding
genes in worm, and a test is performed whether this entire group of
genes is significantly ‘enriched’ for one or more human diseases by over-
lapping orthologous groups and worm and human genes. The statistical
significance of phenologs (P -value) is listed in an output table. A signif-
icant result means that the input genomic region shares a significantly
larger set of orthologous genes with a human disease than would be
expected at random, even if the expressed phenotype in worm appears
very different from the human disease phenotype (e.g. breast cancer
and fertility). This tool can provide novel interpretation of genomic
regions of interest.

4.2.3 Tool 3: ‘QTL2disease’, mapping worm QTLs to human diseases

Researchers can start by selecting a QTL/eQTL in worm to find potential relationships with human diseases. We can select QTLs of interest based on three criteria: a selected experiment, a certain threshold for significance (LOD score) and a specific gene expression probe with a suspected QTL. If there is a QTL with a LOD score above the threshold, we automatically select the closest 50 probes on both sides of the highest peak marker. These probes are presented and available for browsing, shopping and plotting of QTLs, or can be the input for the disease enrichment test to find phenologs.

4.2.4 Tool 4: ‘ComparePheno’, linking worm phenotypes to human diseases

WormQTL^{HD} also provides a tool that links human diseases to classical worm phenotypes (and *vice versa*) to discover phenologs in a systematic way. Users begin by selecting one or more human diseases and clicking on ‘Compare’. The genes associated with the selected disease are tested for enrichment against all sets of known associated genes for worm phenotypes. The result reveals functionally coherent, evolutionarily conserved gene networks.

Alternatively, users can also start by selecting worm phenotypes, which are tested against human diseases. In addition to cross-species testing, results of within-species disease enrichment are also available (e.g. to find the closest related human disease for another input human disease).

4.2.5 Software used

All the software has been implemented using the open source ‘MOlecular GENetics Information Systems’—MOLGENIS—toolkit [328]. The

MOLGENIS toolkit is Java-based software to generate tailored research infrastructure on-demand [329]. In particular, we built on an existing MOLGENIS application, the extensible xQTL workbench [14] and the R/qtl QTL mapping and visualization package for the R language [38, 15]. All software is available as open source on <http://github.com/molgenis> for others to reuse locally. Related technical documentation is available at [http://www.molgenis.org/xqtl²](http://www.molgenis.org/xqtl<sup>2</sup), <http://www.rqtl.org> and <http://www.molgenis.org>.

4.3 Results

To demonstrate the added value of WormQTL^{HD}, we have reproduced findings from known studies and have shown that novel insights and hypotheses can be achieved with little time and effort. Subsequently, we performed a broad-sweep disease-enrichment test to find all non-evident phenologs and to explore which new putative candidate genes for human diseases could be elucidated for future research.

4.3.1 Case 1: Linking disease to worm phenotype from McGary et al. [226]

McGary *et al.* performed a phenolog mapping between the high incidence of male *C. elegans* progeny to human breast/ovarian cancers. Of 4,649 total orthologues, McGary *et al.* reported 3 overlapping genes of 12 human disease-associated genes and 16 worm phenotype-associated genes—which is a significant enrichment (hypergeometric test P -value of $\leq 7.2 \times 10^{-6}$). From the 13 worm phenotype-associated genes that were not overlapping, 9 had orthologues that had already been linked to breast cancer in the primary literature. They implicated the remaining four genes as new breast cancer candidates. We replicated these findings using the ComparePheno tool of WormQTL^{HD}, searching for the

²The URL in the original paper is no longer active and was updated here.

WormBase phenotype 'high incidence male progeny'. The first human disease among the results is '{Breast cancer, susceptibility to}, 114480 (3)' from OMIM. Our tool reported 2 overlapping genes of 4 human disease-associated genes and 63 genes from the worm phenotype. This resulted in a P -value $\leq 1.4 \times 10^{-3}$ (uncorrected). The second best human hit in the results is 'malignant neoplasm of ovary' from DGA. We found two overlapping genes of six ovarian cancer associated genes, resulting in a P -value $\leq 3.41 \times 10^{-3}$ (uncorrected). ComparePheno also indicated enrichment of these categories. The P -values are 'less significant' than McGary et al. because (i) their definition of 'high incidence male progeny' included only 16 rather than 63 genes and (ii) they used an older INPARANOID version, so the overlap test was performed on a different orthologue mapping. Together, these results from our database do indeed replicate their findings. See Online Use Case 1 on the Help page to repeat this case.

4.3.2 Case 2: Worm eQTL hotspot from two temperature expression data from Li et al. [205]

Li et al. [205] found an eQTL hotspot (77.56Mb on chromosome V) on the worm genome in which genetic variation is associated with the expression of 66 genes, while these genes are located elsewhere on the genome. This indicates that these genes are possibly involved in the same biological process/pathway and potentially share a regulatory element. They may be physically located on the eQTL hotspot, which controls gene expression responding to different ambient temperatures. First, we used the Region2disease tool and input positions ChrV:15430739-16430739 (a non-cumulative 1 Mb region around the hotspot). We put all 931 probes located in this region in the shopping cart, and selected 'Disease enrichment'. The best hit was 'Response to antineoplastic agents' (agents used in chemotherapeutic treatment of cancer) from GWAS Catalogue (P -value $\leq 4.92 \times 10^{-3}$, uncorrected). For this hit, the associated human gene, *PPP2R5E*, is orthologous to

1 WBGene00012348 (*pptr-1*) present in this region. The best WormBase
2 hit is 'thermotolerance increased' (P -value $\leq 1.5 \times 10^{-2}$, uncorrected),
3 also via association with *pptr-1*. Padmanabhan *et al.* [251] showed that
4 *pptr-1* is involved in regulating subcellular localization and transcrip-
5 tional activity of the forkhead transcription factor *daf-16*. Rodriguez
6 *et al.* [294] reviewed the role of heat stress response experiments in *C.*
7 *elegans* for detecting human disease genes. They reported that *daf-16*
8 in worms controls lifespan and stress response. In humans, the *daf-16*
orthologue *FOXO3A* is associated with aging and prevalence of cancer
[369]. Using the Disease2QTL tool, a search for 'Response to antineoplastic agents' results in six probes for orthologues of *PPP2R5E* (WBGene00012348) and *ACOX3* (WBGene00019060). We selected them all and plotted the QTLs. This revealed a highly significant (LOD > 50) *cis*-eQTL for *pptr-1* in the Rockman *et al.* [293] dataset. Given all the evidence, we believe *pptr-1* might be an interesting candidate in the further development of a temperature-based *C. elegans* model for understanding human cancer and developing potential therapeutic drugs. Moreover, it shows that combining the 'Region2disease' and 'Disease2QTL' tools can lead to an interesting hypothesis ready for experimental validation. See Online Use Case 2 on the Help page to reproduce this case.

4.3.3 Case 3: Osmotic stress as a model for Bardet-Biedl syndrome from Rodriguez *et al.* [294]

Rodriguez *et al.* proposed hypertonic or osmotic stress in *C. elegans* as a model to study human diseases related to protein aggregation, such as Alzheimer's and Parkinson's. Hypertonic stress due to loss of water causes an intracellular ionic imbalance, which leads to rapid accumulation of organic osmotic glycerol and accumulation of damaged proteins. Shaye and Greenwald [307] showed that *osm-12* (associated with osmotic stress response) is orthologous to *BBS7* in man, which is associated to Bardet-Biedl syndrome [20]. We used the Disease2QTL

tool to look for QTLs associated with Bardet-Biedl syndrome by selecting all ‘Bardet-Biedl syndrome’ entries (seven in total) from OMIM. When we plotted the QTLs in worm for these entries, three significant eQTLs (LOD > 5) were found for *osm-12* (in *cis*), *bbs-5* (also in *cis*) and *bbs-2* (in *trans*). The strongest QTL (LOD > 6) was found for *bbs-5*, reported by probe AGIUSA3442 in the Rockman *et al.* dataset. We used the QTL2disease tool to investigate this QTL further. It revealed a nearby, very significant eQTL (LOD > 10) for a gene named *T07C4.10*, which can be investigated further as a potential candidate for this disease model. See Online Use Case 3 on the Help page to replicate this example.

4.3.4 Novel disease-gene associations by ‘broad-sweep’ disease-enrichment test

We performed hypergeometric gene overlap tests to find phenologs between all worm phenotypes versus all human diseases. Table 4.1 lists the 15 most significant hits for human diseases that have significant gene overlap with worm phenotypes (see Tables 4.2 and 4.3 for the top 100). New candidate genes for human diseases can be discovered from phenologs by investigating human orthologues of worm genes that did not overlap with known human genes of the disease of interest.

McGary *et al.* [226] reported ‘Zellweger syndrome’ in man to be a phenolog with ‘Reduced number of peroxisomes’ in yeast (P -value < 1.0×10^{-9}). Our best hit was ‘Zellweger syndrome’ with ‘peroxisome physiology variant’ in worm (P -value < 3.6×10^{-10}). Encouragingly, certain top hits such as ‘coenzyme Q depleted’ in worm versus ‘Coenzyme Q10 deficiency’ in man, and ‘spontaneous mutation rate increased’ in worm versus ‘Mismatch repair cancer syndrome’ in man make sense, thereby validating this approach and adding credibility to potentially non-evident human disease models.

Phenotype ₁ (Ce)	Phenotype ₂ (Hs)	n_1	n_2	k	P -value
Peroxisome physiology variant	Zellweger syndrome, 214100 (3) [OMIM]	3	4	3	3.58E-10
Coenzyme Q depleted	Coenzyme Q10 deficiency, 607426 (3) [OMIM]	9	3	3	7.53E-09
Spontaneous mutation rate increased	Mismatch repair cancer syndrome, 276300 (3) [OMIM]	42	4	4	9.88E-09
Mitochondrial metabolism variant	Coenzyme Q10 deficiency, 607426 (3) [OMIM]	17	3	3	6.09E-08
AWA odorant chemotaxis defective	Cardiofaciocutaneous syndrome, 115150 (3) [OMIM]	3	2	2	3.64E-07
Peroxisome physiology variant	Adrenoleukodystrophy, neonatal, 202370 (3) [OMIM]	3	3	2	1.09E-06
AWC odorant chemotaxis defective	Cardiofaciocutaneous syndrome, 115150 (3) [OMIM]	5	2	2	1.21E-06
Germ nuclei rachis	Cardiofaciocutaneous syndrome, 115150 (3) [OMIM]	6	2	2	1.82E-06
Excretory cell development variant	Rheumatoid arthritis [GWAS Cataloge]	3	5	2	3.64E-06
Bacterially unswollen	Cardiofaciocutaneous syndrome, 115150 (3) [OMIM]	11	2	2	6.67E-06
Organism starvation response variant	Ovarian cancer, somatic, 604370 (3) [OMIM]	12	2	2	8.00E-06
Neuron development variant	Diastolic blood pressure [GWAS Catalog]	17	11	3	9.85E-06
Ventral closure defective	Wiskott-Aldrich syndrome [DGA]	8	3	2	1.02E-05
Egg laying imipramine resistant	Bone mineral density [GWAS Catalog]	26	23	4	1.08E-05
mRNA export variant	disease by infectious agent [DGA]	4	6	2	1.09E-05

Table 4.1: Top 15 results for the ‘broad-sweep’ disease enrichment. n_1 indicates the number of orthologues in *C. elegans* (Ce) with phenotype₁, n_2 the number in *H. sapiens* (Hs) with phenotype₂ and k the number in both sets. The significance of each phenolog is assessed by the hypergeometric probability (P -value).

Phenotype ₁ (Ce)	Phenotype ₂ (Hs)	n_1	n_2	k	P -value
constipated	Cardiofaciocutaneous syndrome, 115150 (3) [OMIM]	14	2	2	1.10E-05
neuron development variant	Blood Pressure [GWAS Central]	17	12	3	1.31E-05
reproductive system development variant	Palmitoleic acid (16:1n-7) plasma levels [GWAS Catalog]	6	5	2	1.82E-05
germ cell morphology variant	Wiskott-Aldrich syndrome [DGA]	112	3	3	2.04E-05
life span variant	Coenzyme Q10 deficiency, 607426 (3) [OMIM]	114	3	3	2.15E-05
organism starvation response variant	Colorectal cancer, somatic, 114500 (3) [OMIM]	12	3	2	2.40E-05
gastrulation variant	vaccinia [DGA]	21	2	2	2.55E-05
neuron development variant	Blood pressure [GWAS Catalog]	17	15	3	2.69E-05
epithelial morphology variant	Wiskott-Aldrich syndrome [DGA]	13	3	2	2.83E-05
osmotic stress response variant	Chronic obstructive pulmonary disease [GWAS Catalog]	13	3	2	2.83E-05
radiation induced reproductive cell death variant	Mismatch repair cancer syndrome, 276300 (3) [OMIM]	10	4	2	3.26E-05
distal tip cell development variant	Pulmonary function (interaction) [GWAS Catalog]	6	7	2	3.81E-05
dauer arrest variant	Ovarian cancer, somatic, 604370 (3) [OMIM]	26	2	2	3.94E-05
bag of worms	Heart Rate [GWAS Central]	29	2	2	4.92E-05
halothane hypersensitive	Leigh syndrome, 256000 (3) [OMIM]	6	8	2	5.07E-05
vulvalless	Cardiofaciocutaneous syndrome, 115150 (3) [OMIM]	30	2	2	5.27E-05
transgene expression variant	Pulmonary function (interaction) [GWAS Catalog]	147	7	4	5.30E-05
sodium chloride chemotaxis defective	Cardiofaciocutaneous syndrome, 115150 (3) [OMIM]	31	2	2	5.64E-05
synaptogenesis variant	Hippocampal atrophy [GWAS Catalog]	13	4	2	5.65E-05
body size variant	Inflammatory bowel disease [GWAS Catalog]	11	30	3	5.77E-05
gastrulation variant	Wiskott-Aldrich syndrome [DGA]	21	3	2	7.61E-05
pathogen susceptibility increased	Cardiofaciocutaneous syndrome, 115150 (3) [OMIM]	36	2	2	7.64E-05
level of protein expression variant	hepatitis B [DGA]	39	2	2	8.98E-05
male gonad morphology variant	Bone mineral density [GWAS Catalog]	3	23	2	9.17E-05
flaccid	Electrocardiographic traits [GWAS Catalog]	13	5	2	9.41E-05
paraquat resistant	Ovarian cancer, somatic, 604370 (3) [OMIM]	41	2	2	9.94E-05
reproductive system development variant	Metabolic traits [GWAS Catalog]	6	11	2	9.94E-05
spontaneous mutation rate increased	Muir-Torre syndrome, 158320 (3) [OMIM]	42	2	2	1.04E-04
dauer formation variant	Ovarian cancer, somatic, 604370 (3) [OMIM]	43	2	2	1.09E-04
dauer arrest variant	Colorectal cancer, somatic, 114500 (3) [OMIM]	26	3	2	1.18E-04
loss of asymmetry ASE	Cardiofaciocutaneous syndrome, 115150 (3) [OMIM]	47	2	2	1.31E-04
large cytoplasmic granules early emb	Eye Color [GWAS Central]	28	3	2	1.37E-04
gonad sheath contraction rate reduced	Systolic blood pressure [GWAS Catalog]	7	11	2	1.39E-04
osmotic stress response variant	Sudden cardiac arrest [GWAS Catalog]	13	6	2	1.41E-04
male tail morphology variant	substance dependence [DGA]	29	3	2	1.47E-04
dauer constitutive	Ovarian cancer, somatic, 604370 (3) [OMIM]	50	2	2	1.49E-04
excessive blebbing early emb	Eye Color [GWAS Central]	30	3	2	1.57E-04
nucleus reforms cell division remnant early emb	lymphoma [DGA]	14	6	2	1.64E-04
egg laying variant	Body mass index (interaction) [GWAS Catalog]	228	3	3	1.75E-04
cell polarity reversed	Bone mineral density [GWAS Catalog]	4	23	2	1.83E-04
nuclear positioning variant	hepatitis [DGA]	18	5	2	1.84E-04
alae variant	Immune response to smallpox (secreted IFN-alpha) [GWAS Catalog]	11	8	2	1.85E-04
habituation variant	Asthma [GWAS Catalog]	15	6	2	1.89E-04

Table 4.2: Result hits 16-58 for the ‘broad-sweep’ disease enrichment. n_1 indicates the number of orthologues in *C. elegans* (Ce) with phenotype₁, n_2 the number in *H. sapiens* (Hs) with phenotype₂ and k the number in both sets. The significance of each phenolog is assessed by the hypergeometric probability (P -value).

CHAPTER 4. LINKING HUMAN DISEASE TO *C. ELEGANS*

Phenotype ₁ (Ce)	Phenotype ₂ (Hs)	n_1	n_2	k	P -value
age associated fluorescence increased	narcolepsy [DGA]	57	2	2	1.94E-04
pharyngeal development variant	PR interval [GWAS Catalog]	24	4	2	1.99E-04
movement variant	Acute lymphoblastic leukemia (childhood) [GWAS Catalog]	19	5	2	2.06E-04
miRNA expression variant	Type 2 diabetes [GWAS Catalog]	50	25	4	2.12E-04
hypoxia hypersensitive	Mismatch repair cancer syndrome, 276300 (3) [OMIM]	25	4	2	2.17E-04
P0 spindle rotation failure early emb	Colitis, Ulcerative [GWAS Central]	9	11	2	2.38E-04
paralyzed body	Dyssegmental dysplasia, Silverman-Handmaker type, 224410 (3) [OMIM]	1	1	1	2.46E-04
protein phosphorylation increased	glioma [DGA]	21	5	2	2.52E-04
transgene expression variant	Red blood cell traits [GWAS Catalog]	147	17	5	2.52E-04
anaphase bridging	lymphoma [DGA]	18	6	2	2.75E-04
body wall muscle morphology variant	Response to antidepressant treatment [GWAS Catalog]	22	5	2	2.77E-04
coelomocyte uptake defective	Peters anomaly, 604229 (3) [OMIM]	70	2	2	2.93E-04
paraquat resistant	Colorectal cancer, somatic, 114500 (3) [OMIM]	41	3	2	2.96E-04
HSN migration variant	Bone mineral density [GWAS Catalog]	24	23	3	2.97E-04
small	Leukoencephalopathy with vanishing white matter, 603896 (3) [OMIM]	174	4	3	2.99E-04
embryo osmotic integrity defective early emb	Cardiofaciocutaneous syndrome, 115150 (3) [OMIM]	71	2	2	3.01E-04
endomitotic oocytes	glioma [DGA]	23	5	2	3.04E-04
egg laying defective	Bone mineral density [GWAS Catalog]	174	23	6	3.11E-04
dauer formation variant	Colorectal cancer, somatic, 114500 (3) [OMIM]	43	3	2	3.26E-04
P0 spindle rotation delayed early emb	Platelet counts [GWAS Catalog]	6	20	2	3.41E-04
drug induced gene expression variant	Ventricular conduction [GWAS Catalog]	10	12	2	3.55E-04
multiple nuclei oocyte extended life span	Cholesterol [GWAS Central]	45	3	2	3.58E-04
short	Leigh syndrome, 256000 (3) [OMIM]	206	8	4	3.83E-04
lipid composition variant	Palmitoleic acid (16:1n-7) plasma levels [GWAS Catalog]	26	5	2	3.89E-04
fat content increased	Urate levels [GWAS Catalog]	9	14	2	3.92E-04
fewer germ cells	Ovarian cancer, somatic, 604370 (3) [OMIM]	81	2	2	3.93E-04
oogenesis variant	Mean corpuscular hemoglobin [GWAS Catalog]	34	4	2	4.04E-04
oocyte number decreased	rheumatoid arthritis [DGA]	83	2	2	4.13E-04
germ cell compartment anucleate	Cholesterol, LDL [GWAS Central]	194	4	3	4.14E-04
early larval lethal	Cardiofaciocutaneous syndrome, 115150 (3) [OMIM]	84	2	2	4.23E-04
dauer constitutive	kidney disease [DGA]	85	2	2	4.33E-04
transgene expression variant	Colorectal cancer, somatic, 114500 (3) [OMIM]	50	3	2	4.42E-04
sex determination variant	Systolic blood pressure [GWAS Catalog]	147	11	4	4.46E-04
body wall muscle thick filament variant	Bone mineral density [GWAS Catalog]	6	23	2	4.54E-04
response to injury variant	Atrial septal defect 3 (3) [OMIM]	2	1	1	4.92E-04
diplotene progression during oogenesis variant	Ulcerative colitis [GWAS Catalog]	6	24	2	4.95E-04
osmotic stress response variant	Response to Vitamin E supplementation [GWAS Catalog]	38	4	2	5.05E-04
male tail morphology variant	Systolic blood pressure [GWAS Catalog]	13	11	2	5.12E-04
anchor cell invasion variant	Bone mineral density [GWAS Catalog]	29	23	3	5.27E-04
exploded through vulva	Attention deficit hyperactivity disorder [GWAS Catalog]	57	12	3	5.27E-04
antimicrobial gene expression variant	Bone mineral density [GWAS Catalog]	277	23	7	6.03E-04
	Ventricular conduction [GWAS Catalog]	13	12	2	6.13E-04

Table 4.3: Result hits 59-100 for the ‘broad-sweep’ disease enrichment. n_1 indicates the number of orthologues in *C. elegans* (Ce) with phenotype₁, n_2 the number in *H. sapiens* (Hs) with phenotype₂ and k the number in both sets. The significance of each phenolog is assessed by the hypergeometric probability (P -value).

4.4 Discussion

The current version of WormQTL^{HD} (August 2013) is a comprehensive and compendious database that enables molecular model organism data to be studied in the context of human diseases. Just as with WormQTL [315]³, we believe that WormQTL^{HD} will be continuously curated by the members of the *C. elegans* community. The results of the 'broad-sweep' disease-enrichment test in combination with the web tool will be of special interest to researchers in the human or worm domain. We believe these results could also be applied to prioritize the pathogenic variants increasingly being produced by next-generation sequencing in diagnostic labs. Genetic variants affecting human genes of unknown function may have worm orthologues that are part of human-worm phenologs and these may reveal or imply a role in a human disease. Thus, through functionally conserved networks, missing information can be inferred and candidate genes can be selected via model organisms.

The approach of WormQTL^{HD} is conceptually similar to that described by Smedley *et al.* [310]. They created an automated method called PhenoDigm to provide evidence about gene-disease associations by analysing phenotypic information. In their case, phenotypes consist of a collection of ontology terms, which are aligned and scored to derive an overall phenotype-similarity score. Using this method, known gene-phenotype associations in model organisms (mouse, zebrafish) can be transferred to other organisms such as man, and help us to understand the genetic cause of disease. This method works best when the model organism is physiologically close to man and has comparable classical phenotypes. It would therefore be less useful for *C. elegans*. However, combining the molecular (WormQTL^{HD}) and phenotypical (PhenoDigm) approaches may result in a very powerful tool to discover novel gene-disease associations in man, especially when using physiologically close model organisms.

We plan to further develop the WormQTL^{HD} data and toolset.

³The original paper erroneously cited [294] here.

1 There might be more ways in which researchers would like to search
2 through the large amounts of data, for example, based on custom lists
3 of gene identifiers, or by combining tools such as finding QTLs within
4 specific regions. The QTL plots could be improved or replaced with
5 interactive graphs that are more informative and would allow the users
6 to continue 'drilling down' in the data instead of returning to the home
7 page for a new analysis with a different tool. Furthermore, we envisage
8 close integration with other data sources and tools such as WormNet,
R/qtl and GO Enrichment to provide even more biological context and
analytical tools for the user.

Our new database makes this data attractive and easy-to-use for
an even wider community of quantitative geneticists working on worms
and man. We are committed to maintaining the data and software in
the future and invite the community to add and share their new data
and ideas.

Supplementary Data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank Prof. Anthony J. Brookes and Dr Robert Hastings at the Department of Genetics, University of Leicester, UK, for providing us with the data from GWAS Central via a custom batch query, WormBase, the McKusick-Nathans Institute of Genetic Medicine (OMIM), Northwestern University, Chicago, Illinois (DGA) and The National Human Genome Research Institute (GWAS CATALOG) for openly providing their data in a structured way. We also thank Bart Charbon, Erwin Winder and Dennis Hendriksen for proofreading the manuscript and for their helpful suggestions, Jackie Senior for editing, Patrick Deelen, Roan Kanninga and Pieter Dopheide for testing the WormQTL^{HD} software,

and many members of the *C. elegans* community for their comments and ideas.

Funding

European Union Seventh Framework Programme (FP7/ 2007-2013) research projects BioSHaRE-EU [261433 to K.J.V. and M.A.S.]; Research Project PANACEA [222936 to J.E.K. and R.C.J.]; TI Food and Nutrition [TIFN GH001 to M.A.S.]; Dutch Carbohydrate Competence Center [CCC WP23 to K.Z.]; Centre for BioSystems Genomics (CBSG) and the Netherlands Consortium of Systems Biology (NCSB), both of which are part of the Netherlands Genomics Initiative/ Netherlands Organisation for Scientific Research [to D.A.]; ERASysbio-plus ZonMW project GRAPPLE— Iterative modelling of gene regulatory interactions underlying stress, disease and ageing in *C. elegans* [90201066 to L.B.S.]; Netherlands Organisation for Scientific Research (NWO) VENI grant [863.13.011 to Y.L.]. Funding for open access charge: EU FP7.

Conflict of interest statement. None declared.

1
2
3
4
5
6
7
8

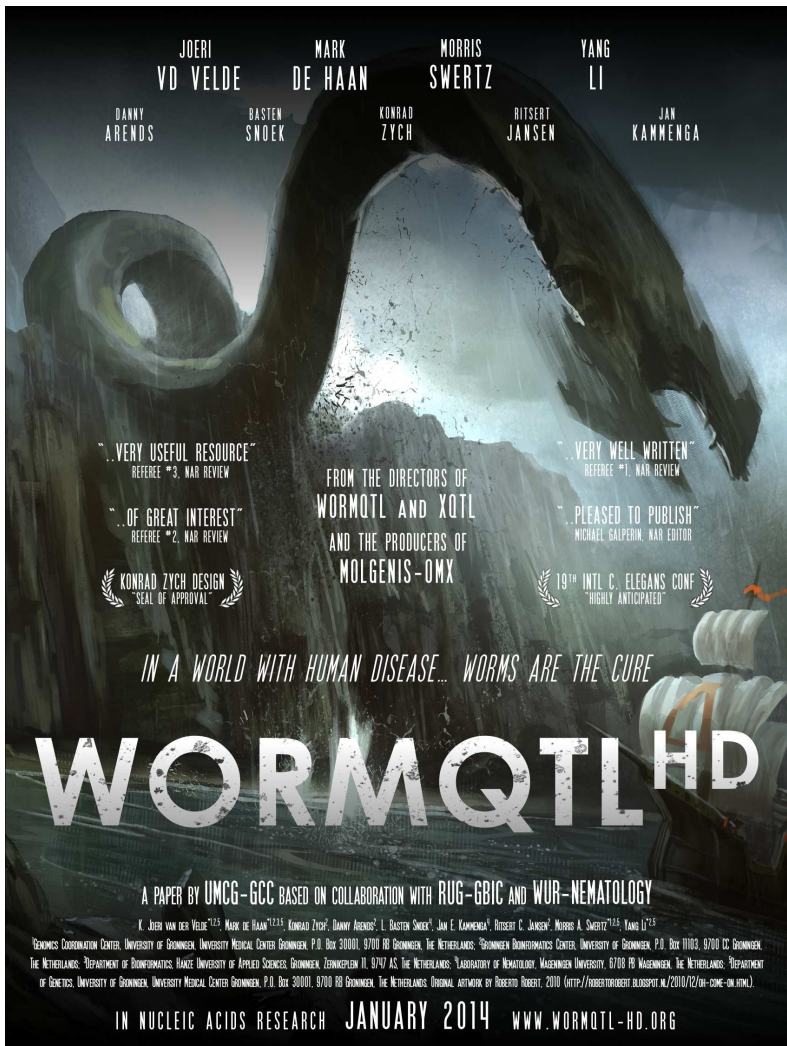


Figure 4.3: WormQTL^{HD} poster. Special thanks to Roberto Robert for kind permission to use his incredible artwork. This figure was not part of the published article but has eye-catching properties (anecdotal evidence only). 96

Chapter 5

**Evaluation of CADD
Scores in Curated
Mismatch Repair Gene
Variants Yields a Model
for Clinical Validation and
Prioritization**

1
2
3
4
5
6
7
8

Hum Mutat. 2015 Jul;36(7):712-9.

DOI: 10.1002/humu.22798

PubMed ID: 25871441

1 K. Joeri van der Velde^{1,2,†}, Joël Kuiper^{2,3,†}, Bryony A. Thompson⁴,
2 John-Paul Plazzer⁵, Gert van Valkenhoef³, Mark de Haan^{1,2}, Jan D.H.
3 Jongbloed², Cisca Wijmenga², Tom J. de Koning², Kristin M. Abbott²,
4 Richard Sinke², Amanda B. Spurdle⁴, Finlay Macrae^{5,6}, Maurizio Genuardi⁷,
5 Rolf H. Sijmons², Morris A. Swertz^{1,2,*} and InSiGHT Group^{2,4,5,6,7}

1. Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands.

2. Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands.

3. Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands.

4. Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Australia.

5. Department of Colorectal Medicine and Genetics, Royal Melbourne Hospital, Melbourne, Australia.

6. Department of Medicine, The Royal Melbourne Hospital, University of Melbourne, Melbourne, Australia.

7. Institute of Medical Genetics, "A. Gemelli" School of Medicine, Catholic University of the Sacred Heart, Rome, Italy.

Communicated by Peter N. Robinson

Received 9 January 2015; accepted revised manuscript 30 March 2015.
Published online 13 April 2015 in Wiley Online Library.

† The authors wish it to be known that the first two authors should be regarded as joint first authors.

* Correspondence to: Morris A. Swertz. E-mail: m.a.swertz@rug.nl

Abstract

Next-generation sequencing in clinical diagnostics is providing valuable genomic variant data, which can be used to support healthcare decisions. In silico tools to predict pathogenicity are crucial to assess such variants and we have evaluated a new tool, Combined Annotation Dependent Depletion (CADD), and its classification of gene variants in Lynch syndrome by using a set of 2,210 DNA mismatch repair gene variants. These had already been classified by experts from InSiGHT's Variant Interpretation Committee. Overall, we found CADD scores do predict pathogenicity (Spearman's $\rho = 0.595$, $P < 0.001$). However, we discovered 31 major discrepancies between the InSiGHT classification and the CADD scores; these were explained in favor of the expert classification using population allele frequencies, cosegregation analyses, disease association studies, or a second-tier test. Of 751 variants that could not be clinically classified by InSiGHT, CADD indicated that 47 variants were worth further study to confirm their putative pathogenicity. We demonstrate CADD is valuable in prioritizing variants in clinically relevant genes for further assessment by expert classification teams.

Key words: Lynch syndrome; variant classification; pathogenicity prediction; cumulative link model

5.1 Introduction

Reliable estimation of gene variant pathogenicity, especially for missense variants and small in-frame insertions/deletions (indels), is a major challenge in clinical genetics. This challenge is now being exacerbated by the introduction of next-generation sequencing in clinical diagnostics, which is identifying large numbers of candidate disease-causative variants, ranging from about 250 [212], to 400–700 [378], up to a mean of 1,083 [302] variants per exome, depending on which filter steps and stringency are applied. Since it is not feasible to perform functional analysis of each variant, in silico tools have become an important tool

1 in assessing variant pathogenicity. Unfortunately, although there are
2 many potential methodologies and tools [68], they often lack clinical
3 validation. As the adaptation of high-throughput sequencing in clinical
4 practice increases, the need for standardized, validated, and easy-to-use
5 in silico classification tools is becoming even more pressing [302, 378].

6 The recently launched Combined Annotation Dependent Depletion
7 (CADD) [185] method offers a standardized, genome-wide, variant scor-
8 ing metric (C-score) that incorporates the weighted results of widely
used in silico pathogenicity prediction tools, such as SIFT [187] and
PolyPhen [5], and of genomic annotation sources like ENCODE [85].
The resulting CADD scores are expressed as a measure of deleterious-
ness (selection pressure bias) for single-nucleotide variants (SNVs) and
small indels. A high score represents variants that are not stabilized
by selection, which are more often disease-causing than expected by
random chance [185]. In contrast, a low score indicates that a vari-
ant resembles evolutionary stable, commonly occurring genetic varia-
tion that poses no apparent disadvantage for an organism. The scores
were shown to correlate strongly to known variant pathogenicity, such
as those causing a predisposition to autism spectrum disorders, intellectual
disability, thalassemia, and more broadly to pathogenic variants taken
from the NHGRI GWAS catalog [363] and ClinVar [190] database. To
make interpretation and comparison easier, C-scores are logarithmically
ranked to form scaled C-scores, similar to how PHRED scores are used
in the FASTQ format.

As an easy-to-use resource that brings out the predictive power of
many programs and data combined, CADD may replace the plethora of
tools currently being used. However, before considering implementation
of CADD in clinical work, it is important to evaluate and validate its
utility by comparing its outcome with that of existing, consistent, large-
scale expert assessments.

The Variant Interpretation Committee (VIC) is an expert panel of
the International Society for Gastrointestinal Hereditary Tumours (In-
SiGHT). They conducted a thorough clinical classification of 2,360

variants (as of February 2014) in the DNA mismatch repair (MMR) genes *MLH1* (MIM #120436), *MSH2* (MIM #609309), *MSH6* (MIM #600678), and *PMS2* (MIM #600259) that had been identified in patients suspected of having Lynch syndrome [337]. This cancer predisposition syndrome, previously known as hereditary nonpolyposis colorectal cancer, is caused by DNA MMR deficiency.

The InSiGHT variant classification method is based on a combination of clinical and experimental (molecular) evidence, such as family history and cosegregation with the disease, tumor findings, population allele frequencies, and mRNA/protein functional assays (in accordance with established guidelines, available at <http://www.insight-group.org/criteria>).

The variants were classified following a five-tier system [268], with class descriptions as follows:

- Class 1: not pathogenic/no clinical significance.
- Class 2: likely not pathogenic/little clinical significance.
- Class 3: uncertain clinical significance.
- Class 4: likely pathogenic.
- Class 5: pathogenic.

Variants that cannot be placed in classes 1, 2, 4, or 5 based on existing evidence are assigned to class 3 by default and are considered variants of uncertain clinical significance. It is recognized [337] that class 3 may include some cases with conflicting evidence.

Here, we investigate whether CADD scores are concordant with variant classifications assigned by the InSiGHT VIC. We show that, overall, CADD and InSiGHT yield similar results, but that there are also some important discordant cases. Our contributions in this paper are:

1. An extensive evaluation of agreement between the *in silico* CADD predictions and the InSiGHT expert classifications of variant pathogenicity.

2. Detection and assessment of conflicting classifications.
3. A CADD-based prioritization of variants of uncertain clinical significance.
4. Assessment of the reliability of CADD for use in a clinical setting.

These contributions shed light on an important question in clinical genetic diagnostics: are bioinformatics tools powerful enough to enable genome-wide variant interpretation without loss of quality when compared with classification by clinical expert panels that can also take into account a range of clinical and molecular data relevant for specific genetic diseases?

5.2 Materials & Methods

5.2.1 Data processing

We downloaded 2,744 variants (as of February 2014) from the InSiGHT LOVD database (at http://chromium.liacs.nl/LOVD2/colon_cancer/) for *MLH1*, *MSH2*, *MSH6* and *PMS2*. RefSeq identifiers NM_000249.3, NM_000251.2, NM_000179.2, NM_000535.5 were added to the cDNA position. This allowed the successful conversion of 2,582 variants to genomic DNA notation in VCF format by running Ensembl VEP5[228]. CADD (version 1.0) was able to score 2,580 of those (NM_000249.3:c.1254T>R and NM_000535.5:c.1875A>Y failed). Of these 2,580 variants, 370 were not assessed by InSiGHT, or in a few cases belonged to multiple classes. This means that 2,210 variants were classified and belong to one of the five classes of the International Agency for Research on Cancer (IARC) five-tiered classification system: 151 variants belong to class 1 (not pathogenic), 84 to class 2, 751 to class 3, 181 to class 4, and 1,043 to class 5 (pathogenic).

In addition, we ran SnpEff to obtain functional effect predictions using canonical transcript references and an upstream downstream interval

length of five bases. The output was curated to reduce the number of effects from two to one in the case of both INTRON and SPLICE SITE “effects,” by removing the INTRON effect. We used NM 000251.2 for MSH2 (whereas the LOVD was based on NM 000251.1) to enable ENSEMBL VEP to process the data, without issues (out of 920 MSH2 variants, 855 were successfully converted to VCF/gDNA notation).

5.2.2 Cumulative link model

To detect discrepancies between the CADD scores and the InSiGHT classification, we assumed that a partitioning of the scores would exist. In other words, the continuous scaled C-scores can be binned into the ordinal IARC classes. Working on this assumption, we were able to define a cumulative link model (ordinal regression) [6, 225]. In a cumulative link model, an ordinal response variable Y_i can fall in $j = 1, \dots, J$ ordered classes. This response variable Y_i then follows a distribution with parameter π_i where π_{ij} denotes the probability that the i th observation falls in the j th response class (such that $\sum_{j=1}^J \pi_{ij} = 1$). Since we are dealing with individual observations (instead of counts) the categorical distribution is used, which can be viewed as a special case of the multinomial distribution of n observations $Y_i \sim \text{Mult}(n, \pi_i)$ with $n = 1$:

$$Y_i \sim \text{Categorical}(\pi_i)$$

The cumulative probability is then defined as:

$$\gamma_{ij} = P(Y_i \leq j) = \pi_{i1} + \dots + \pi_{ij}$$

Here we considered a proportional odds model, using a logit link function: $\text{logit}(p) = \log[p/(1-p)]$. The cumulative logits for all but the last class, $j = 1, \dots, J-1$, are then defined as:

$$\begin{aligned} \text{logit}(\gamma_{ij}) &= \text{logit}(P(Y_i \leq j)) \\ &= \log \frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)} \end{aligned}$$

This gives a regression for the cumulative logits:

$$\text{logit}(\gamma_{ij}) = \theta_j - \mathbf{x}_i^\top \boldsymbol{\beta}$$

where θ_j represents the logit-scaled cut-off for class j , \mathbf{x}_i being the vector of explanatory variables for the i th observation and $\boldsymbol{\beta}$ is the corresponding set of regression parameters. Note that $\mathbf{x}_i^\top \boldsymbol{\beta}$ does not contain an intercept. The parameters θ_j act as a set of continuous "cut-off points" such that $-\infty < \theta_1 < \dots < \theta_{J-1} < \infty$. To assess the probability that the i th observation falls within one of ordinal response classes j , we can write:

$$P(Y_i = j \mid \mathbf{x}_i^\top \boldsymbol{\beta}) = \begin{cases} \gamma_{ij}, & j = 1 \\ \gamma_{ij} - \gamma_{i(j-1)}, & j = 2, \dots, J-1 \\ 1 - \gamma_{i(J-1)}, & j = J \end{cases}$$

We used the CADD score as an explanatory variable for the ordinal response of InSiGHT. The parameters were estimated using JAGS, a program for analysis of Bayesian graphical models using Gibbs sampling [269]. Convergence of the Markov Chain Monte Carlo inference was assessed using the potential scale reduction factor [270, 120]. Figure 5.1 shows the probability that a given CADD score belongs to a certain InSiGHT class by using the posterior distributions for θ after convergence. Discrepancies were detected by analyzing the deviance of the observations. Deviance can be thought of as a measure of "surprise", how likely a certain observation is under the fitted parameters of the model. Formally:

$$D(Y_i, \hat{\theta}) = -2 \log[P(Y_i | \hat{\theta})]$$

with Y_i being the observation and $\hat{\theta}$ the parameters of the fitted model. Observations of θ corresponding to variants in the 95th percentile of the mean deviance—those with the highest deviance—were re-examined.

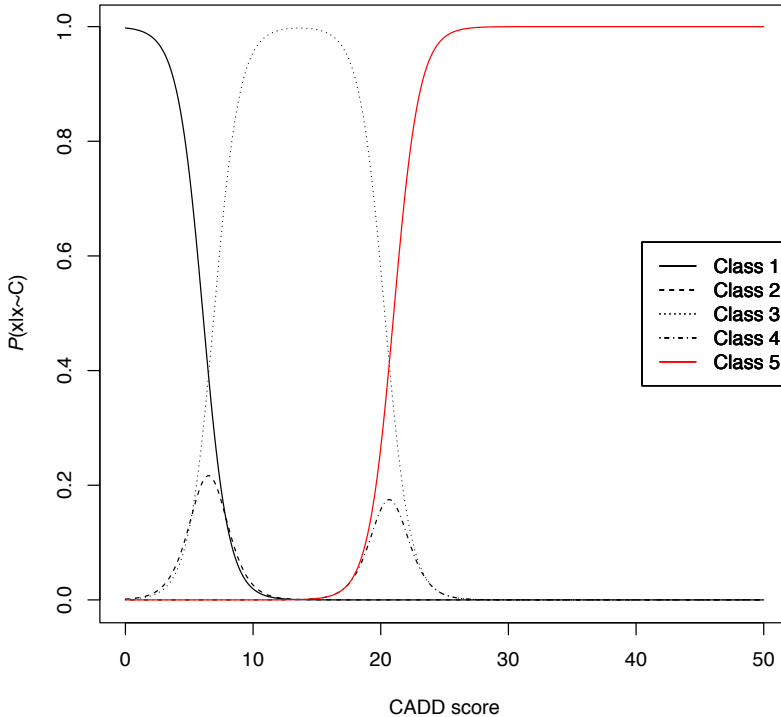


Figure 5.1: Probability that a CADD score will belong to a certain InSiGHT class. The inverse logit (logit^{-1}) was applied to each of the response variables. Classes 2 and 4 are dominated by class 3 under this model.

5.2.3 Data availability

The data and scripts used in this paper can be downloaded from: http://molgenis.org/downloads/vdVelde_Kuiper_etal_2015/

5.3 Results

5.3.1 Exploratory data analysis

We calculated the CADD scores for 2,744 MMR gene variants that were downloaded from the InSiGHT group LOVD (available at http://chromium.liacs.nl/LOVD2/colon_cancer/). A total of 534 variants had to be omitted, either because converting the complementary DNA HGVS nomenclature[79] based notation to genomic DNA VCF (Variant Call Format version 4.0 [73]) based notation failed (162 variants), or the CADD scores could not be unambiguously assigned (2 variants with $T>R$ and $A>Y$ substitutions), or because they had not yet been classified by the InSiGHT VIC (i.e., they were recent submissions, or not reported as germline variants [370 variants]). See Figure 5.2 and *Materials & Methods* for details. The 2,210 remaining variants fell within one of the five classes: class 1 ($n = 151$), class 2 ($n = 84$), class 3 ($n = 751$), class 4 ($n = 181$), or class 5 ($n = 1,043$).

Overall, the CADD scaled C-score distributions for each class correlate with the InSiGHT classification (Spearman's $\rho = 0.595$, $p < 0.001$). In Figure 5.3, the distribution of the scores per class is represented in a beanplot[179]. See also Figures 5.4, 5.5, 5.6 and 5.7 for CADD scores of the InSiGHT variants for each gene, using known variants identified in the Genome of the Netherlands[243, 244] and 1000 Genomes[63] projects as population background reference.

5.3.2 Discrepancy assessment

Using a Bayesian cumulative link model, we identified 108 (4.89% of 2,210) cases for which a different class would be assigned (see *Materials & Methods*). Further analysis focused on the cases for which the nonpathogenic (class 1) and pathogenic (class 5) classifications were reversed, as these suggested major disagreements between CADD and the InSiGHT VIC verdict (see Table 5.1). The explanations per variant for this analysis can be found in Table 5.2 and Table 5.3.

CADD model	InSiGHT classification				
	Class 1	Class 2	Class 3	Class 4	Class 5
Class 1	135				19
Class 2		71			
Class 3	4	1	704	3	3
Class 4				171	
Class 5	12	12	47	7	1021

Table 5.1: Number of InSiGHT variants reassigned to alternative classes according to the cumulative link model fitted on CADD scores.

5.3.3 False positives

We identified 12 variants (0.54% of 2,210) that were classified as non-pathogenic (class 1) by the InSiGHT VIC, but they were predicted to be pathogenic (class 5) according to the CADD-based cumulative link model (see *Materials & Methods*). Re-examination of the available data for these variants strongly supports the original InSiGHT classification based on the following evidence:

- Segregation data is inconsistent with the variant being a dominant, high-risk, pathogenic sequence variant in pedigrees (likelihood ratio ≤ 0.01).

CHAPTER 5. EVALUATION OF CADD SCORES IN MMR GENES

Gene	Variant	In-SiGHT class	CADD-based class	Explanation
1	MLH1 c.394G>C	1	5	Attenuated protein function, but does not cause Lynch syndrome. Multifactorial likelihood analysis posterior probability <0.001
2	MLH1 c.1852_1853delinsGC	1	5	Low risk, not associated with Lynch. Multifactorial likelihood analysis posterior probability <0.001
3	MLH1 c.803A>G	1	5	Multiple microsatellite stable tumours and does not segregate with disease. Multifactorial likelihood analysis posterior probability <0.001
4	MLH1 c.977T>C	1	5	Multiple microsatellite stable tumours and does not segregate with disease. Multifactorial likelihood analysis posterior probability <0.001
5	MLH1 c.1853A>C	1	5	Multiple microsatellite stable tumours and does not segregate with disease. Multifactorial likelihood analysis posterior probability <0.001
6	MLH1 c.1853A>C	1	5	Multiple microsatellite stable tumours and does not segregate with disease. Multifactorial likelihood analysis posterior probability <0.001
7	MLH1 c.2146G>A	1	5	Multiple microsatellite stable tumours and does not segregate with disease. Multifactorial likelihood analysis posterior probability <0.001
8	MLH1 c.1151T>A	1	5	Population minor allele frequency >1%
	MLH1 c.2152C>T	1	5	Population minor allele frequency >1%
	MSH2 c.1077-10T>C	1	5	Population minor allele frequency >1%
	MLH1 c.1799A>G	1	5	Does not segregate with disease. Multifactorial likelihood analysis posterior probability <0.001
	MLH1 c.790+10A>G	1	5	Does not cause splicing aberration and does not segregate with disease. Multifactorial likelihood analysis posterior probability <0.001
	MSH2 c.593A>G	1	5	May be low-moderate risk, but certainly not high-risk associated with Lynch
	MSH6 c.642C>A	5	1	Stop-gain variant causing protein truncation

Table 5.2: Overview of explanations according to InSiGHT why the cumulative link model based on CADD scores encountered certain false positives and false negatives, pt. 1/2.

Gene	Variant	In-SiGHT class	CADD-based class	Explanation	
MSH6	c.642C>G	5	1	Stop-gain variant causing protein truncation	1
MSH2	c.212-478T>G	5	1	Splicing aberration introduces premature termination codon (also missed by SnpEff)	2
MSH2	c.646-3T>G	5	1	Splicing aberration introduces premature termination codon	3
MSH2	c.367-480_645+644del	5	1	Deletion of Exon 3	4
MLH1	c.307-1420_380+624del	5	1	Deletion of Exon 4	5
MLH1	c.307-820_380+896del	5	1	Deletion of Exon 4	6
MLH1	c.381-415_453+733del	5	1	Deletion of Exon 5	7
MLH1	c.454-665_545+49del	5	1	Deletion of Exon 6 (raw score of 527)	8
MLH1	c.1039-675_1409+26del	5	1	Deletion of Exon 12 (raw score of 361)	
MLH1	c.1039-2329_1409+827del	5	1	Deletion of Exon 12 (raw score of 353)	
MLH1	c.1732-2243_1896+404del	5	1	Deletion of Exon 16	
MSH2	c.1077-135_1276+119dup	5	1	Duplication of Exon 7 (also missed by SnpEff)	
MSH2	c.1077-220_1276+6245del	5	1	Deletion of Exon 7	
MSH2	c.1277-572_1386+2326del	5	1	Deletion of Exon 8 (raw score of 464)	
PMS2	c.804-?_903+?del	5	1	Deletion of Exon 8	
PMS2	c.804-?_2006+?del	5	1	Deletion of Exons 8-11	
PMS2	c.989-296_1144+706del	5	1	Deletion of Exon 10 (raw score of 527)	
PMS2	c.2276-113_2445+1596del	5	1	Deletion of Exon 14	

Table 5.3: Overview of explanations according to InSiGHT why the cumulative link model based on CADD scores encountered certain false positives and false negatives, pt. 2/2.

- Variant with reported frequency $\geq 1\%$ in the general population (1000 Genomes Project), and no evidence that variant is a founder mutation.
- These are not high-risk variants that are uniquely associated with Lynch syndrome (they have also been seen in individuals who do not meet the international criteria for Lynch syndrome).
- Variant leads to a known attenuated protein function, but this does not cause Lynch syndrome (it has also been seen in healthy individuals and there is a lack of evidence for MMR deficiency as shown by MSI and immunohistochemical testing).

Although these explanations are specific to Lynch syndrome-related variants, they indicate that CADD might overestimate the general pathogenicity of some variants. Most overestimations could be easily resolved in a clinical Standard Operating Procedure (SOP) by using population allele frequency as a filter or incorporating the use of patient pedigree analysis data; these are already common practices in many clinical laboratories. The remainder could be resolved by incorporating more in-depth findings from validated protein functional assays or from risk estimates based on large, well-designed, case-control studies that consider cohort size, geography/ethnicity, and quality control measures[338]. An evaluation of likely not pathogenic (class 2) variants predicted to be pathogenic (class 5) can be found in Table 5.4.

5.3.4 False negatives

We identified 19 cases (0.86% of 2,210) for which the cumulative link model predicted the respective variants to be class 1, whereas InSiGHT scored them as class 5. This indicates that the model might also underestimate effects. Similar to the approach to the false-positives, outlined above, our re-examination of these variants supported the original InSiGHT classification.

Gene	Variant	AA change	Probability	VIC justification
MLH1	c.117-43_117-39del	intronic	0.99	Intronic substitution with no associated splicing aberration, tested with NMD inhibitors
MLH1	c.845C>G	A282G	0.92	Posterior probability 0.001-0.049
MLH1	c.885-24T>A	intronic	0.81	Intronic substitution with no effect on splicing and MAF 0.01-1%
MLH1	c.974G>A	R325Q	0.99	Posterior probability 0.001-0.049
MLH1	c.1742C>T	P581L	0.55	Posterior probability 0.001-0.049. No CMMRD phenotype with co-occurrence and MAF 0.01-1%
MLH1	c.1808C>G	P603R	0.99	Posterior probability 0.001-0.049
MLH1	c.1820T>A	L607H	0.99	Posterior probability 0.001-0.049
MSH2	c.991A>G	N331D	0.69	Posterior probability 0.001-0.049
MSH2	c.1730T>C	I577T	0.86	Posterior probability 0.001-0.049
MSH2	c.2500G>A	A834T	0.99	Posterior probability 0.001-0.049
MSH6	c.3488A>T	E1163V	0.92	MAF >1% in specific population
MSH6	c.4068_4071dup	Lys1358 Aspfs*2	0.99	MAF >1% in specific ethnic group

Table 5.4: Variants of class 2 (likely benign) for which class 5 (pathogenic) is the predicted class according to the CADD-based model. Posterior probabilities are derived from a multifactorial likelihood analysis.

1 CADD scores are developed for scoring any possible human SNVs or
2 small indels[185]. It was therefore expected that large structural variants
3 would be missed or inaccurately scored (for 5/15 structural variants) by
4 CADD. To simplify the interpretation, the scaled C-scores are based on
5 the rank of the C-score relative to all the C-scores for 8.6 billion possible
6 SNVs. Typical variant C-scores in this study ranged from -4 to 14, while
7 the five structural variants in question scored very highly (between 350
8 and 550), whereas was expected considering the likely pathogenicity of
exon deletions relative to missense variants or codon deletions, for ex-
ample. However, the scaling algorithm seems to fail for such extreme
C-scores, and this results in reverting the score for the respective vari-
ant into a very low scaled C-score instead. We applied SnpEff[60] as a
second-tier test. This tool has been developed to annotate and predict
the effects of variants in genes in a robust and qualitative way, thereby
complementing the quantitative nature of CADD scores. Using SnpEff,
we were able to correct 17 of the 19 false-negative cases. SnpEff recog-
nized 14 of the 15 structural variants, most as "EXON_DELETED", one
of two splice aberrations as "FRAME_SHIFT", and two of two truncat-
ing mutations as "STOP_GAINED". These effect types are annotated
as HIGH impact in SnpEff, in contrast to MODIFIER, LOW or MOD-
ERATE effect types. By using SnpEff information, we have shown that
CADD results should be complemented by this tool, or a comparable
tool, to compensate for sporadic underestimations. See Figure 5.8 for
an overview of SnpEff variant effect predictions in relation to CADD
scores and InSiGHT classifications.

5.3.5 Variants of unknown significance

Class 3 mainly contains variants for which insufficient clinical or molec-
ular data are available, but also a limited number of variants that have
discordant findings (i.e., are resistant to classification). Most of these
variants can easily be assigned to another class as soon as more data
become available. As expected, the distribution of the CADD scores

for class 3 variants, as visualized in Figure 5.3, is much flatter than the distributions for the other classes. Matching the CADD score of each class 3 variant to the distributions of the other classes (and thus, the likelihood of belonging to one of them) allows us to propose an endpoint classification that is, according to the model, more likely than belonging to class 3 for these variants. In other words, we can suggest prioritization of a variant for reclassification (using additionally obtained clinical and molecular evidence) when its CADD score deviates far enough from this mean, reaching a score that falls into the distributions of known nonpathogenic or pathogenic variant classes (see *Materials & Methods*).

We performed this analysis and 47 variants (2.13% of 2,210) that the InSiGHT VIC classified as class 3 (uncertain significance) had CADD scores ≥ 34 , which fell in the $>99\%$ probability range for known class 5 (pathogenic) variants (see Figure 5.1). Of these 47 variants, 43 were missense with a mean CADD score of 35.33 ($\sigma = 1.04$, 27 in *MLH1*, 10 in *MSH2*, four in *MSH6* and two in *PMS2*). The remaining four were truncating mutations: two stop-gain variants (c.2250C>A and c.2250C>G, both with a CADD score of 41), and two frameshift variants (c.2252_2253del and c.2262del) with CADD scores of 39 and 40. These four variants are all located in the *MLH1* gene; they were classified as class 3 by the InSiGHT VIC due to insufficient evidence, because the stop codons are introduced in the last exon (19) and are located outside any known functional domains.

We compared these findings with the previous use of a prediction model[337] on 481 substitutions[338] of uncertain effect. In this analysis, 173 InSiGHT missense variants of uncertain significance (class 3) with a $>80\%$ probability in favor of pathogenicity, were prioritized for further investigation using multifactorial likelihood analysis. The model calibrated a combination of in silico tools to predict probabilities of pathogenicity, which is conceptually somewhat similar to the way CADD scores are constructed, except here the model was specifically for MMR gene variants associated with Lynch syndrome.

1 By comparing the two sets of results, that is, the 173 previously
2 identified variants with our 43 prioritized variants, we found an overlap
3 of 24 variants(see Table 5.5). Since they were called by both mod-
4 els, we consider these 24 missense variants to be the most urgent
5 candidates for further research to determine their pathogenicity. Of
6 the remaining 19 variants prioritized uniquely by CADD, 17 had been
7 evaluated before with prior probabilities of pathogenicity ranging from
8 7% (MSH2:c.1418C>T) to 74-76% (MLH1: c.85G>T, c.187G>A,
c.299G>A, c.794G>A, c.955G>A, c.1976G>A, PMS2: 137G>A).

We also compared a CADD-based binary classifier for missense vari-
ants with the multifactorial likelihood model[337]. The multifactorial
model's combination of customized MAPP + PolyPhen2 was found
to perform best with an R^2 (the coefficient of determination) of 0.62
and an area under curve receiver operating characteristic (ROC-AUC)
of 93%, when distinguishing classes 1 + 2 collapsed as "likely not
pathogenic" versus classes 4 + 5 collapsed as "likely pathogenic". As a
comparison, and not related to the cumulative link model, we performed
a binary classification using CADD scores and obtained a ROC-AUC of
85%, showing that while a CADD-based binary classifier for MMR gene
missense variants performs reasonably well, it does not perform as well
as a disease-specific model.

5.4 Discussion

We investigated the use of CADD scores for the prediction of clinical classifications by comparing them with a high quality clinical data set developed by the InSiGHT VIC, which is based on quantitative and qualitative interpretation of both clinical and molecular data. Generally, the CADD model predictions fitted the InSiGHT classification. Out of the 2,210 variants we tested and classified by InSiGHT, we identified 12 (0.54%) nonpathogenic (class 1) variants that the CADD model predicted to be pathogenic (class 5), and 19 variants (0.86%) of class

5 that CADD predicted to be class 1. The difference could be explained by two considerations: the CADD model was not designed to classify large structural or splice-site variants (55% of all the discordant cases, 89% of the false-negatives), and the clinical observations, population allele frequencies, and experimental molecular data sometimes convincingly suggested an alternative interpretation (39% of all discordant cases, 100% of the false-positives). CADD's main underestimation of pathogenicity was due to its inability to accurately predict the effects of whole exon deletions or duplications. In five such cases, the C-score was in fact extremely high, but this was not translated into a high scaled C-score. The use of a second-tier test, in this case SnpEff, boosted the sensitivity of classifying via CADD by correcting 17 out of 19 of these underestimations.

We showed that estimating the deleteriousness of whole exon deletions/duplications is a weakness of CADD and this needs to be addressed. The InSiGHT data shows that such structural variation is often pathogenic, but this is not always recognized by CADD. To avoid incorrect results, and in line with the design limitations of CADD as acknowledged by its authors, we recommend CADD should not be used to judge the pathogenicity of large structural variation as part of an automated variant processing pipeline.

We also investigated the 12 cases of pathogenicity overestimation by CADD, which showed that these false-positives could be explained by data used for the InSiGHT classification that was not used for in silico prediction (such as the presence of the variant in the general population or lack of cosegregation of the variant with the disease). These results underscore the importance of using clinical data in the diagnostic interpretation of variants.

There are a few variants in the InSiGHT database with a known negative effect, such as attenuated protein function, that are classified as nonpathogenic. The InSiGHT VIC require both concordant functional and clinical evidence to assign pathogenicity; they do not accept that attenuated function would necessarily be associated with Lynch

Syndrome – or any phenotype for that matter. In our analysis, for example, CADD predicted a deleterious effect for MLH1:c.394G>C, which is indeed known to cause attenuated protein function[208], but is not considered to be pathogenic in the context of Lynch syndrome because it is not known to be associated with the causal phenotype. Variant classifications such as those currently provided by the InSiGHT VIC for MMR genes are specifically developed for a given phenotype, namely, Lynch syndrome. Therefore, as acknowledged by the VIC[338], they may not capture modest disease penetrance or other disease phenotypes associated with a given variant. This highlights the fact that some apparent discrepancies may simply be explained by the difference in application of "research tools" such as CADD and "clinical tools" such as the InSiGHT database; the latter focuses on results that are of practical value for a clinical geneticist instead of yielding a spectrum of variants with possible intermediate penetrance that then require further interpretation and individualized risk management protocols.

In general, there is limited added value in using CADD scores to assess truncating variants since they are already known to often be pathogenic for known disease genes. The field of in silico prediction benefits most from the power of CADD scores when they are applied to predict the pathogenicity of nonsynonymous SNVs. Here, we show that CADD performs well on this type of variant for Lynch syndrome, although a disease-specific model performs better.

We identified 47 variants that had been assigned by InSiGHT to class 3 (uncertain significance), which, according to the CADD model, had a high probability of being pathogenic. Of these, 24 missense variants were already strongly suspected of being pathogenic by a previous in silico study on MMR gene variant classification[337] and we consider them to be top candidates for further study to confirm their pathogenicity. This suggests that CADD, in a fashion similar to existing disease-specific pathogenicity prediction models, can help in prioritizing variants for the collection of missing clinical and molecular data.

Taken together, we have shown that CADD scores are in high agree-

ment with expert assessments of MMR gene variant pathogenicity that is based on multiple data sources for quantitative-multifactorial and qualitative analysis. As expected, CADD scores are not yet suitable to interpret large structural variants such as deletions and duplications of exons. Other underestimation effects are rare and often detectable with a second-tier test. Any overestimated variants could be excluded based on population frequency, cosegregation analyses, or evidence showing no association or causality.

Calibrated *in silico* pathogenicity prediction models are not intended to replace functional wet-laboratory studies, but are instead complementary methods to let clinics benefit from existing gold standard classifications, by accessing their expert knowledge and making it possible to assess and prioritize novel variants with reasonable confidence, without the need for often unfeasible amounts of laboratory work. We believe CADD fits this translatory role very well, particularly because of its generic and high-throughput nature. Although CADD cannot replace clinical and molecular validation, it can, in a practical sense, assist in prioritizing variants for functional testing when an affected patient carries multiple poorly understood candidate variants, reducing waiting time for results.

However, translating this knowledge into a clinical setting is not trivial. We constructed a model based on ordinal regression of known classifications to calibrate CADD scores as a predictor of pathogenicity for gene variants in the Lynch syndrome-associated MMR genes. Similar efforts are required to unlock the potential of CADD scores as predictors for other disorders, leading to gene- or disease-specific guidelines that can help clinicians translate CADD scores into clinical practice. The threshold for "what is pathogenic" is expected to be rather different to define depending on whether the disease is caused by dominantly or recessively acting mutations, whether the disease is Mendelian or complex/multigenic in origin, and so on. Although the fact that CADD scores are largely based on conservation indicates that it may not work as well for every gene, we believe that its overall usefulness is currently

unmatched by other quantitative pathogenicity estimates.

As a preliminary proof of principle, we compared the distributions of CADD scores of known pathogenic variants (from ClinVar[190]) with the distributions of variants found in the general population (from Genome of the Netherlands[243, 244] and 1000 Genomes[63]), for as many genes as data availability allowed. This approach can be used to estimate the predictive power of CADD scores and, thereby, provide valuable information to clinicians regarding how effective CADD scores are for predicting variant pathogenicity in the context of a specific gene. Encouragingly, out of 373 genes with sufficient data, we found 272 genes (73%) for which CADD has good predictive power (AUC of >90%).

However, this approach is currently still in development. For reliable automated calibration of CADD scores on many genes into a clinical setting, we need to consider many factors and sources of bias potentially influencing the informativity of CADD scores, such as mutation spectrum, penetrance, disorder heterogeneity, variant classification quality, classification semantics, and disorder inheritance patterns.

We conclude that *in silico* pathogenicity predictions are becoming powerful enough to facilitate accurate variant prioritization, at least for dominantly inherited disorders such as Lynch syndrome.

Acknowledgements

This work was supported by grants from BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007). We thank Jackie Senior for editing. Amanda Spurdle was supported by an NHMRC Senior Research Fellowship. K. Joeri van der Velde was supported by UMCG MGE Systems medicine (project nr. 671285).

Author contributions

RHS and MAS conceived and supervised the research. JPP, BAT and AS supplied the data from the InSiGHT database and helped in interpreting discrepancies, and critical review of the manuscript. KJvdV and JK performed the data analysis and wrote the manuscript with input from MdH, JDHJ, CW, TdK, KMA, RS, MG and FM. GvV and JK defined the cumulative-link model in R. All authors read and approved the manuscript.

1
2
3
4
5
6
7
8

CHAPTER 5. EVALUATION OF CADD SCORES IN MMR GENES

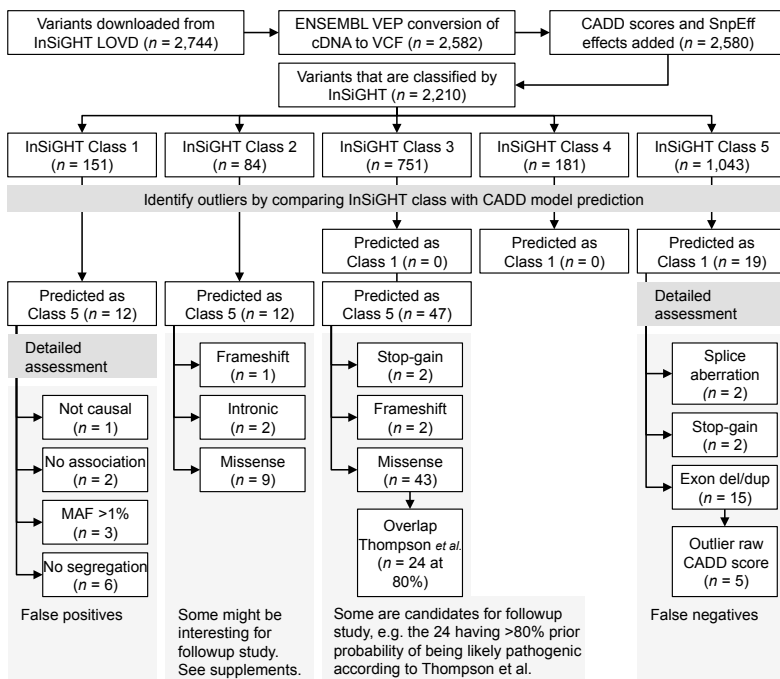


Figure 5.2: Flowchart describing the steps and results of the analysis. [Note: the original figure mentions 2,274 downloaded variants, this was corrected to 2,744 here.]

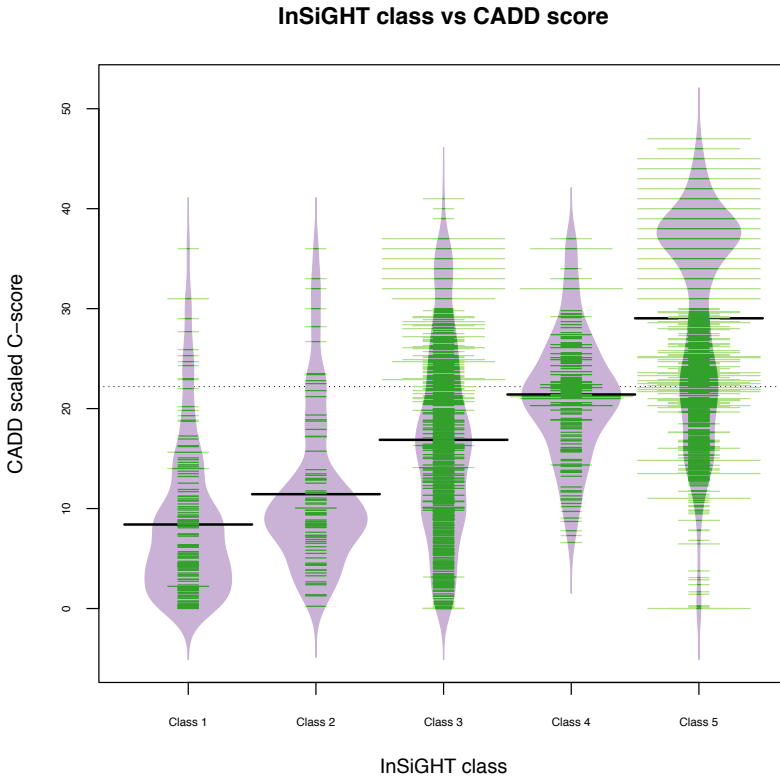


Figure 5.3: Beanplot[179] showing the data points (green) and density estimation (purple) of the scaled CADD C-score per InSiGHT class. The width of the green lines is relative to the number of data points at that score. Black horizontal lines indicate the mean per InSiGHT class; the dotted line shows the overall mean. The mean scores of classes 1-5 show a respective stepwise increase of 8.41 ($\sigma = 7.46$), 11.44 ($\sigma = 7.72$), 16.87 ($\sigma = 9.40$), 21.41 ($\sigma = 6.13$), and 29.04 ($\sigma = 10.28$). The unclassified group (class 3) shows a flatter distribution than the other classes.

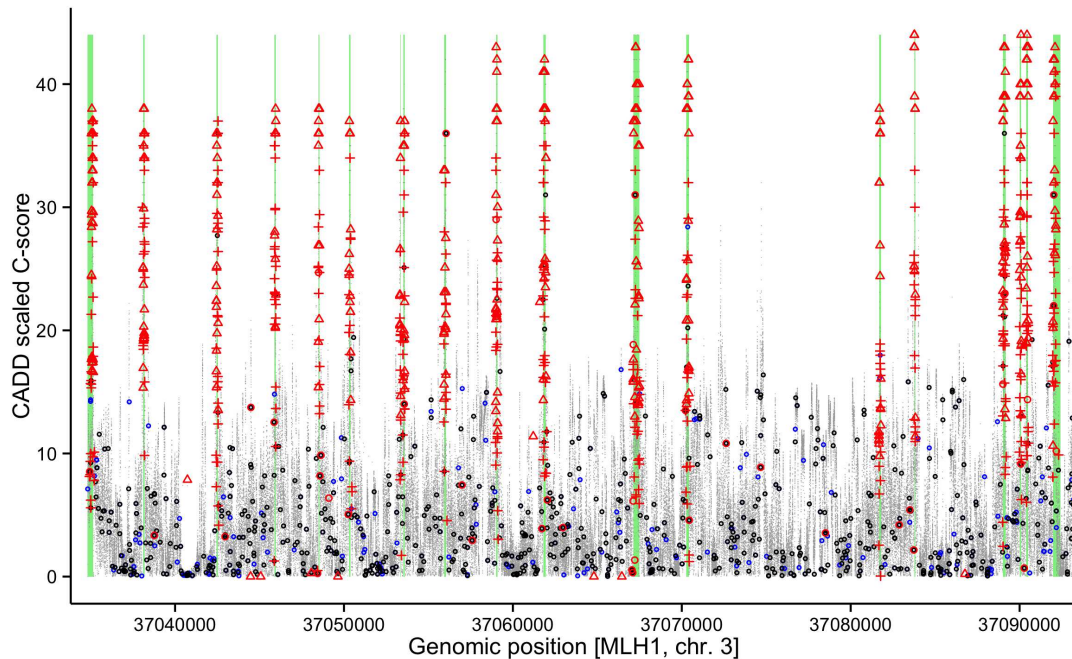


Figure 5.4: CADD scaled C-scores vs. genomic coordinates for MLH1 gene variants. The green bands are the exons. Red are InSiGHT variants, where triangles represent class 5, circles class 1, and pluses class 2-4. The black circles are variants seen in 1000 Genomes[63], blue circles are seen in the Genome of the Netherlands[243, 244]. The gray dots represent all potential SNVs.

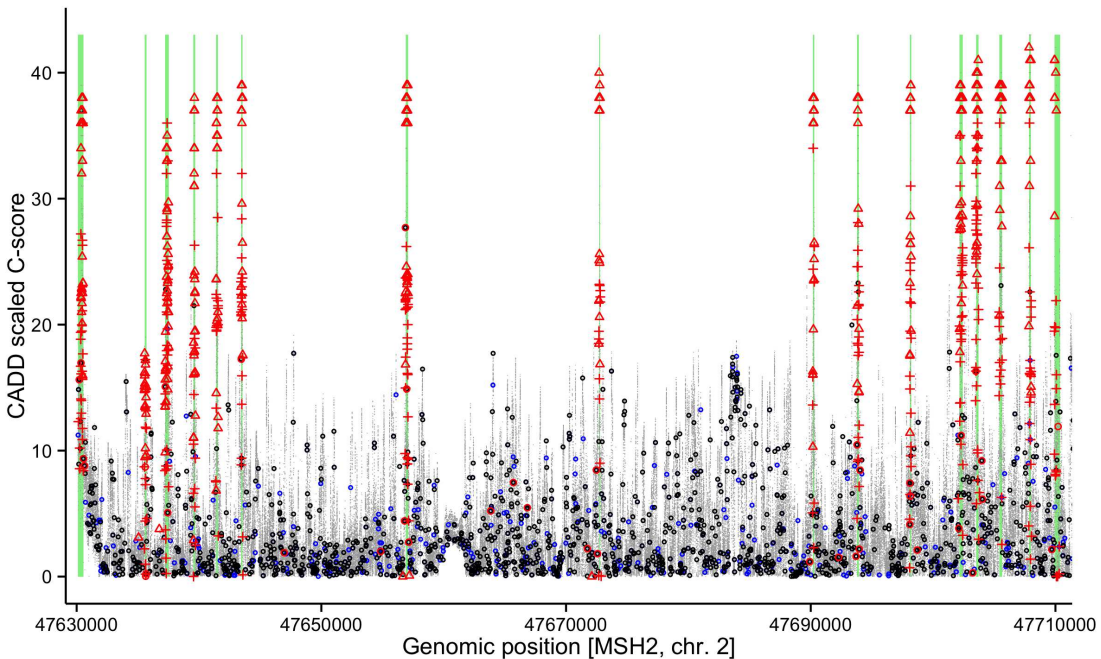


Figure 5.5: CADD scaled C-scores vs. genomic coordinates for MSH2 gene variants. The green bands are the exons. Red are InSiGHT variants, where triangles represent class 5, circles class 1, and pluses class 2-4. The black circles are variants seen in 1000 Genomes[63], blue circles are seen in the Genome of the Netherlands[243, 244]. The gray dots represent all potential SNVs.

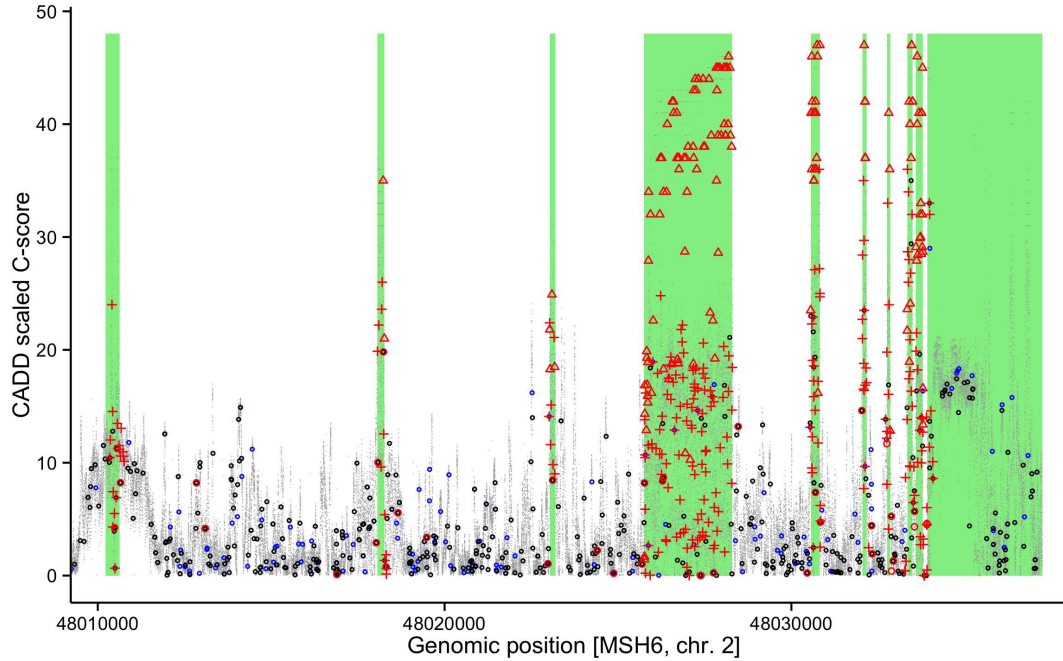


Figure 5.6: CADD scaled C-scores vs. genomic coordinates for MSH6 gene variants. The green bands are the exons. Red are InSiGHT variants, where triangles represent class 5, circles class 1, and pluses class 2-4. The black circles are variants seen in 1000 Genomes[63], blue circles are seen in the Genome of the Netherlands[243, 244]. The gray dots represent all potential SNVs.

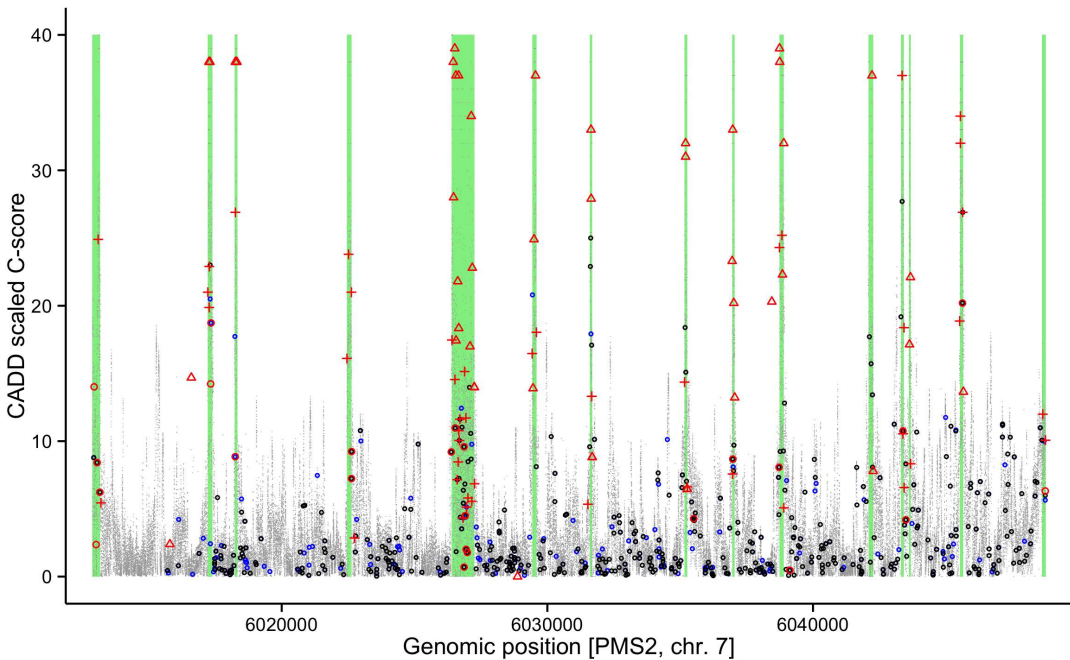


Figure 5.7: CADD scaled C-scores vs. genomic coordinates for PMS2 gene variants. The green bands are the exons. Red are InSiGHT variants, where triangles represent class 5, circles class 1, and pluses class 2-4. The black circles are variants seen in 1000 Genomes[63], blue circles are seen in the Genome of the Netherlands[243, 244]. The gray dots represent all potential SNVs.

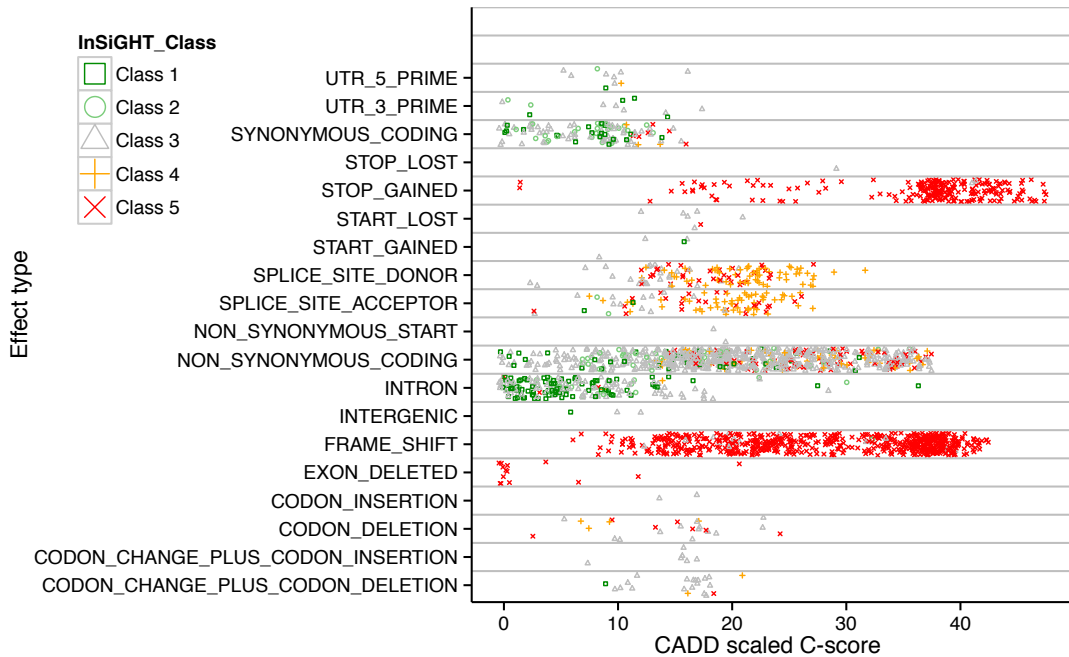


Figure 5.8: Primary SnpEff effect prediction vs. CADD scaled C-score, with InSiGHT classifications colored.

Gene	Variant	AA ch.	Previous[337]	Here	
MLH1	c.1037A>G	Q346R	0.95	0.99	
MLH1	c.109G>A	E37K	0.87	0.99	1
MLH1	c.112A>G	N38D	0.94	0.99	
MLH1	c.125C>T	A42V	0.96	0.99	2
MLH1	c.184C>A	Q62K	0.88	0.99	
MLH1	c.1918C>T	P640S	0.82	0.99	3
MLH1	c.1919C>T	P640L	0.93	0.99	4
MLH1	c.304G>A	E102K	0.87	0.99	
MLH1	c.307G>C	A103P	0.97	0.99	5
MLH1	c.331G>C	A111P	0.97	0.99	
MLH1	c.347C>A	T116K	0.93	0.99	6
MLH1	c.65G>C	G22A	0.89	0.99	
MLH1	c.67G>A	E23K	0.86	0.99	7
MLH1	c.74T>C	I25T	0.86	0.99	8
MLH1	c.80G>C	R27P	0.97	0.99	
MLH1	c.925C>T	P309S	0.83	0.99	
MSH2	c.1799C>T	A600V	0.96	0.99	
MSH2	c.1826C>T	A609V	0.96	0.99	
MSH2	c.2064G>A	M688I	0.89	0.99	
MSH2	c.2141C>T	A714V	0.87	0.99	
MSH2	c.2168C>T	S723F	0.88	0.99	
MSH2	c.2187G>T	M729I	0.88	0.99	
MSH2	c.529G>A	E177K	0.86	0.99	
MSH6	c.3682G>C	A1228P	0.97	0.99	

Table 5.5: The 24 variants that are still uncertain and predicted by bioinformatic tools to be likely pathogenic, according to the probabilities of the MAPP + PolyPhen2 calibrated model[337] and the CADD model.

CHAPTER 5. EVALUATION OF CADD SCORES IN MMR GENES

1

2

3

4

5

6

7

8

Chapter 6

**GAVIN: Gene-Aware
Variant INterpretation for
medical sequencing**

Genome Biol. 2017 Jan 16;18(1):6.

DOI: 10.1186/s13059-016-1141-7

PubMed ID: 28093075

1

2

3

4

5

6

7

8

1 K. Joeri van der Velde^{1,2}, Eddy N. de Boer², Cleo C. van Diemen²,
2 Birgit Sikkema-Raddatz², Kristin M. Abbott², Alain Knopperts², Lude
3 Franke², Rolf H. Sijmons², Tom J. de Koning², Cisca Wijmenga²,
4 Richard J. Sinke² and Morris A. Swertz^{1,2,*}

1. University of Groningen, University Medical Center Groningen, Ge-
2 nomics Coordination Center, Groningen, The Netherlands.

3 2. University of Groningen, University Medical Center Groningen, De-
4 partment of Genetics, Groningen, The Netherlands.

5 Received: 13 July 2016; accepted: 19 December 2016; published: 16
6 January 2017

* Correspondence: m.a.swertz@gmail.com

7 **Abstract**

8 We present Gene-Aware Variant INterpretation (GAVIN), a new method
that accurately classifies variants for clinical diagnostic purposes. Clas-
sifications are based on gene-specific calibrations of allele frequencies
from the ExAC database, likely variant impact using SnpEff, and esti-
mated deleteriousness based on CADD scores for >3,000 genes. In a
benchmark on 18 clinical gene sets, we achieve a sensitivity of 91.4%
and a specificity of 76.9%. This accuracy is unmatched by 12 other
tools. We provide GAVIN as an online MOLGENIS service to annotate
VCF files and as an open source executable for use in bioinformatic
pipelines. It can be found at <http://molgenis.org/gavin>.

Keywords: Clinical next-generation sequencing, Variant classification,
Automated protocol, Gene-specific calibration, Allele frequency, Protein
impact, Pathogenicity prediction

6.1 Background

Only a few years ago, the high costs and technological challenges of whole exome and whole genome sequencing were limiting their application. Today, the practice of human genome sequencing has become routine even within the healthcare sector. This is leading to new and daunting challenges for clinical and laboratory geneticists[29]. Interpreting the thousands of variations observed in DNA and determining which are pathogenic and which are benign is still difficult and time-consuming, even when variants are prioritized by state-of-the-art in silico prediction tools and heuristic filters[68]. Using the current, largely manual, variant classification protocols, it is not feasible to assess the thousands of genomes per year now produced in a single hospital. It is the challenge of variant assessment which now impedes the effective uptake of next-generation sequencing into routine medical practice.

The recently introduced CADD[185] scores are a promising alternative[347]. These are calculated on the output of multiple in silico tools in combination with other genomic features. They trained a computer model on variants that have either been under long-term selective evolutionary pressure or none at all. The result was an estimation of deleteriousness for variants in the human genome, whether already observed or not. It has been shown to be a strong and versatile predictor for pathogenicity[185] with applications and popular uptake in many areas of genome research. Variant interpretation in a diagnostic setting may also benefit from this method. However, succesful uptake requires a translational effort because CADD scores are intended to rank variants, whereas NGS diagnostics requires a discrete classification for each variant. For example, SIFT[187] probabilities are used to partition 'tolerated' (probability >0.05) from 'damaging' variants (probability ≤ 0.05). CADD scores may be used to define such a binary classifier, but using a single, arbitrary cut-off value is not recommended by the CADD authors[46]. Moreover, clinicians and laboratories cannot rely on a single threshold approach because it has been shown that

1 individual genes differ in their cut-off thresholds for what should be con-
2 sidered the optimal boundary between pathogenic or benign[347]. This
3 issue has been partly addressed by mutation significance cutoff (MSC)
4 [164], which provides gene-based CADD cut-off values to remove in-
5 consequential variants safely from sequencing data. While MSC aims
6 to quickly and reliably reduce the number of benign variants left to
7 interpret, it was not developed to detect/classify pathogenic variants.

8 The challenge is thus to find robust algorithms that classify both
pathogenic and benign variants accurately and that fit into existing best
practice, diagnostic filtering protocols[288]. Implementing such tools is
not trivial because genes have different levels of tolerance to various
classes of variants that may be considered harmful[196]. In addition,
the pathogenicity estimates for benign variants are intrinsically lower
because these are more common and of less severe consequence on
protein transcription. Comparing the prediction score distributions of
pathogenic variants with those of typical benign variants is therefore
biased and questionable. Using such an approach means it will be un-
clear how well a predictor truly performs if a benign variant shares the
same allele frequency and consequence with known pathogenic vari-
ants. Here, we present Gene-Aware Variant Interpretation (GAVIN), a
new method that addresses these issues by gene-specific calibrations on
closely matched sets of variants. GAVIN delivers accurate and reliable
automated classification of variants for clinical application.

6.2 Results

6.2.1 Development of GAVIN

GAVIN classifies variants as benign, pathogenic or a variant of uncer-
tain significance (VUS). It considers ExAC[196] minor allele frequency,
SnpEff[60] impact and CADD score using gene-specific thresholds. For
each gene, we ascertained ExAC allele frequencies and effect impact
distributions of variants described in ClinVar (November 2015 release)

[191] as pathogenic or likely pathogenic. From the same genes we selected ExAC variants that were not present in ClinVar as a benign reference set. We stratified this benign set to match the pathogenic set with respect to the effect impact distribution and minor allele frequencies (MAFs). Using these comparable variant sets we calculated gene-specific mean values for CADD scores (across all genes, the pathogenic mean of means was 28.44 and that of benign 23.08) and MAFs, as well as 95th percentile sensitivity/specificity CADD thresholds for both benign and pathogenic variants. Of 3,237 genes that underwent the calibration process, we found 2,525 informative gene calibrations, i.e. thresholds for CADD, effect impact, pathogenic 95th percentile MAFs or a combination thereof (see Additional file 1: Table S1). We used fixed genome-wide classification thresholds as a fall-back strategy based on CADD scores < 15 for benign, > 15 for pathogenic and on a MAF threshold of 0.00426, which was the mean of all gene-specific pathogenic 95th percentile MAFs. This allowed classification when insufficient variant training data were available to allow for gene-specific calibrations, or when the gene-specific rules failed to classify a variant. Based on the gene calibrations we then implemented GAVIN, which can be used online or via commandline (see <http://molgenis.org/gavin>) to perform variant classification.

6.2.2 Performance benchmark

To test the robustness of GAVIN, we evaluated its performance using six benchmark variant classification sets from VariBench[239], MutationTaster2[304], ClinVar (only recently added variants that were not used for calibrating GAVIN), and a high-quality variant classification list from the University Medical Center Groningen (UMCG) genome diagnostics laboratory. These sets and the origins of their variants and classifications are described in Table 6.1. The combined set comprises 25,765 variants (17,063 benign, 8,702 pathogenic). All variants were annotated by SnpEff, ExAC and CADD prior to classification

1 by GAVIN. To assess the clinical relevance of our method, we strat-
2 ified the combined set into clinically relevant variant subsets based
3 on organ-system specific genes. We formed 18 subset panels such
4 as Cardiovascular, Dermatologic, and Oncologic based on the gene-
5 associated physical manifestation categories from Clinical Genomics
6 Database[319]. A total of 11,679 out of 25,765 variants were not linked
7 to clinically characterized genes and formed a separate panel (see Table
8 6.2 for an overview, which includes the number of pathogenic variants
in each panel). In addition, we assessed the performance of GAVIN
in comparison to 12 common in silico tools for pathogenicity predic-
tion: MSC (using two different settings), CADD (using three differ-
ent thresholds), SIFT[187], PolyPhen2[5], PROVEAN[59], Condel[129],
PON-P2[241], PredictSNP2[28], FATHMM-MKL[308], GWAVA[289],
FunSeq[112] and DANN[278].

Across all test sets, GAVIN achieved a median sensitivity of 91.4%
and a median specificity of 76.9%. Other tools with >90% sensitivity
were CADD (93.6% at threshold 15, with specificity 57.1%, and 90.4%
at threshold 20, with specificity 68.8%) and MSC (97.1%, specificity
25.7%). The only tool with a higher specificity was CADD at threshold
25 (85.3%, sensitivity 71.5%). See Table 6.3 for an overview of tool
performance or Figure 6.1 for more detail. In all the clinical gene sets
GAVIN scored >89.7% sensitivity, including >92% for Cardiovascular,
Biochemical, Obstetric, Neurologic, Hematologic, Endocrine and Der-
matologic genes. The non-clinical genes scored 71.3%. The specificity
in clinical subsets ranged from 70.3% for Endocrine to 84.2% for Den-
tal. Non-clinical gene variants were predicted at 70.6% specificity. See
Additional file 2: Table S2 for detailed results.

6.2.3 Added value of gene-specific calibration

We then investigated the added value of using gene-specific thresholds
on classification performance relative to using genome-wide thresholds.
We bootstrapped the performance on 10,000 random samples of 100

Dataset	Ben-ign variants (n)	Patho-genic variants (n)	Origin
VariBench tolerance DS7, training set	11,347	6,143	PhenCode database, IDbases, and 18 individual LSDBs
VariBench tolerance DS7, test set	1,377	510	PhenCode database, IDbases, and 18 individual LSDBs
MutationTaster2 benchmark set	1,194	161	HGMD Professional and 1000 Genomes
ClinVar (additions of Nov 2015 to Feb 2016)	1,668	1,688	Submissions by clinical molecular geneticists, expert panels, diagnostic laboratories and companies
UMCG, variants exported from clinical diagnostic interpretation software	1,176	174	Clinical diagnostic classifications of variants in cardiology, dermatology, epilepsy, dystonia and preconception screening
UMCG, germline variants for familial cancer cases	301	26	Hereditary cancer variant classifications by an M.D. following ACMG guidelines
Total	17,063	8,702	25,765

Table 6.1: Variant and classification origins of the benchmark data sets used.

CHAPTER 6. VARIANT INTERPRETATION FOR MEDICAL SEQ.

	CGD manifestation panel	Genes (n)	Variants (n)	Likely pathogenic/pathogenic variants (n)
1	Allergy / Immunology / Infectious	253	1,952	1,324
2	Audiologic / Otolaryngologic	217	1,215	668
3	Biochemical	354	2,538	1,933
4	Cardiovascular	446	4,360	2,408
5	Craniofacial	387	1,861	1,106
6	Dental	80	783	518
7	Dermatologic	345	2,749	1,662
8	Endocrine	240	1,801	1,340
	Gastrointestinal	338	2,351	1,620
	Genitourinary	149	1,026	753
	Hematologic	267	2,571	1,914
	Musculoskeletal	676	4,935	2,864
	Neurologic	1,012	6,363	4,055
	Obstetric	34	223	140
	Oncologic	203	2,157	1,207
	Ophthalmologic	479	3,649	2,406
	Pulmonary	90	717	485
	Renal	302	2,143	1,459
	<i>NotInCGD</i>	<i>5,806</i>	<i>11,679</i>	<i>122</i>

Table 6.2: Stratification of the combined variant data set into manifestation categories. The categories are defined by Clinical Genomics Database and are associated to clinically relevant genes. Variants were allocated to the manifestation categories based on their gene and were placed in multiple categories if a gene was associated to multiple manifestations.

Tool	Median sensitivity (%)	Median specificity (%)
CADD (thr. 15)	93.6	57.1
CADD (thr. 20)	90.4	68.8
CADD (thr. 25)	71.5	85.3
Condel	70.3	39.5
DANN	63.8	66.7
FATHMM	69.5	61.9
FunSeq	61.7	50.2
GAVIN	91.4	76.9
GWAVA	47.6	26.2
MSC_ClinVar95CI	84.7	64.4
MSC_HGMD99CI	97.1	25.7
PolyPhen2	68.0	46.8
PONP2	47.5	26.9
PredictSNP2	66.8	70.6
PROVEAN	65.9	62.1
SIFT	67.9	57.9

Table 6.3: Performance overview of all tested tools.

CHAPTER 6. VARIANT INTERPRETATION FOR MEDICAL SEQ.

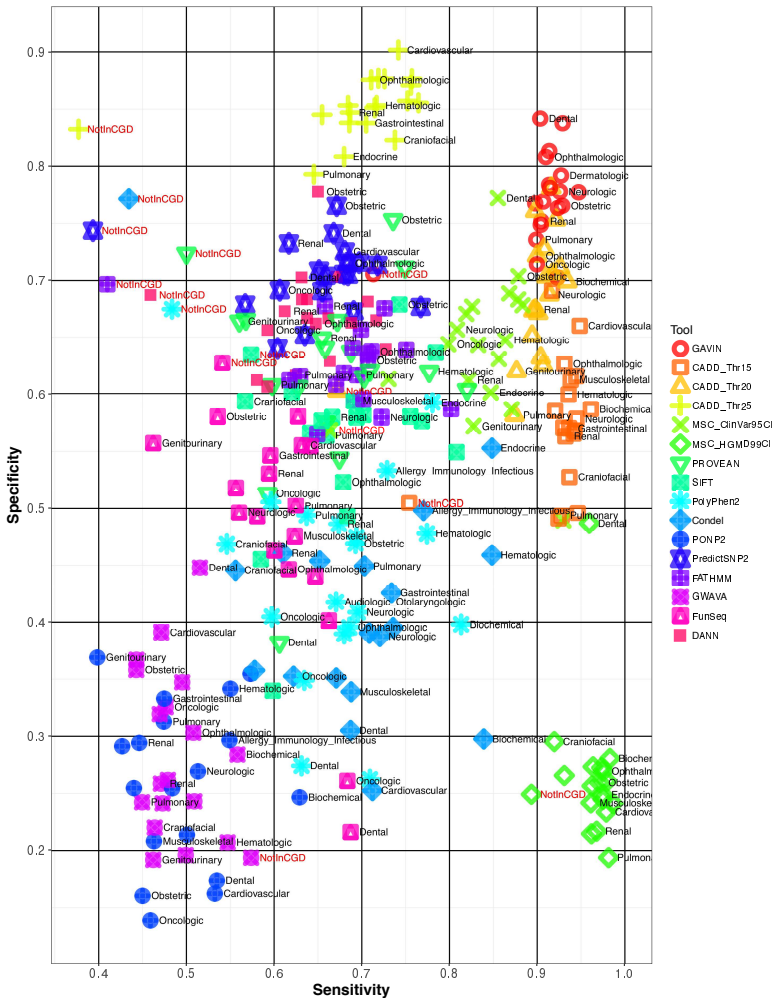


Figure 6.1: Performance of GAVIN and other tools across different clinical gene sets. Prediction quality is measured as sensitivity and specificity, i.e. the fraction of pathogenic variants correctly identified and the fraction of misclassifications/non-classifications while doing so.

benign and 100 pathogenic variants. These variants were drawn from the three groups of genes described in "Methods": (1) genes for which CADD was significantly predictive for pathogenicity ($n = 681$); (2) genes where CADD was not significantly predictive ($n = 732$); and (3) genes with scarce variant data available for calibration ($n = 774$). For each of these sets we compared the use of gene-specific CADD and MAF classification thresholds with that of genome-wide filtering rules.

We observed the highest accuracy on genes for which CADD had significant predictive value and for the gene-specific classification method (median accuracy = 87.5%); this was significantly higher than using the genome-wide method for these same genes (median accuracy = 84.5%, Mann-Whitney U test p value $< 2.2e-16$). For genes for which CADD had less predictive value we found a lower overall performance, but still reached a significantly better result using the gene-specific approach (median accuracy = 84.5% versus genome-wide 82.5%, p value $< 2.2e-16$). Lastly, the worst performance was seen for variants in genes with scarce training data available. The gene-specific performance, however, was still significantly better than using genome-wide thresholds (median accuracy = 82.5% and 80.5% respectively, p value = $2.2e-16$). See Figure 6.2.

6.3 Discussion

We have developed GAVIN, a method for automated variant classification using gene-specific calibration of classification thresholds for benign and pathogenic variants.

Our results show that GAVIN is a powerful classifier with consistently high performance in clinically relevant genes. The robustness of our method arises from a calibration strategy that first corrects for calibration bias between benign and pathogenic variants, in terms of consequence and rarity, before calculating the classification thresholds. A comprehensive benchmark demonstrates a unique combination of high

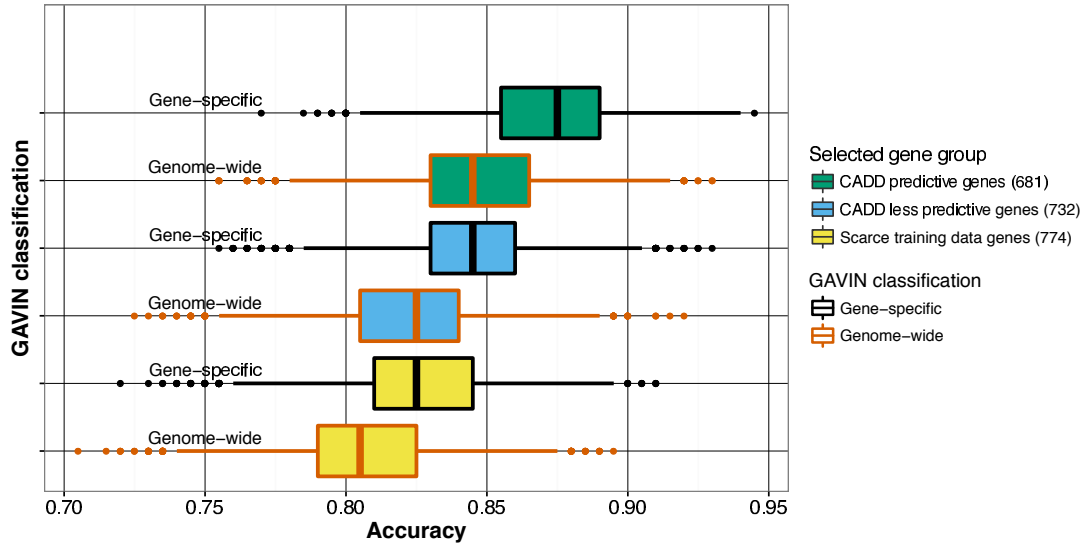


Figure 6.2: Comparison of gene-specific classification thresholds with genome-wide fixed thresholds in three groups of genes: 737 genes for which CADD is predictive, 684 genes for which CADD is less predictive, and 766 genes with scarce training data. For each group, 10,000 sets of 100 benign and 100 pathogenic variants were randomly sampled and tested from the full set of 25,765 variants and accuracy was calculated for gene-specific and genome-wide CADD and MAF thresholds.

sensitivity (>90%) and high specificity (>70%) for variants in genes related to different organ systems. This is a significant improvement over existing tools that tend to achieve either a high sensitivity (MSC, CADD at lower thresholds) or a high specificity (PredictSNP2, CADD at higher thresholds). A high sensitivity is crucial for clinical interpretation because pathogenic variants should not be falsely discarded. In addition, having a higher specificity means that the results will be far less "polluted" with false positives and thus less risk of patients being given a wrong molecular diagnosis. GAVIN decreases false positives by 10-20% compared to using CADD for the same purpose, thereby reducing interpretation time. The difference between using a high and low performance method can be dramatic in practice. In a hypothetical example, GAVIN would make downstream variant interpretation twice as effective as a low performance method, with more sensitive detection of pathogenic variants (see Table 6.4).

Even though an optimal combination of sensitivity and specificity may be favorable in general terms, there may still be a need for tools that perform differently. The MSC gene-specific thresholds based on HGMD[323] at 99% confidence interval show a very high sensitivity (97.1%), but at the expense of a very low specificity (25.7%). Such low specificity thresholds will pick up almost all the pathogenic variants with scores exceeding gene thresholds. This allows safe removal (<3% error) of benign variants that fall below these thresholds, which was their authors' aim. However, this tool cannot detect pathogenic variants due its low specificity. Other tools, such as PON-P2, may show a relatively low performance, but not necessarily because of true errors. Such tools may simply be very 'picky' and only return a classification when the verdict carries high confidence. If we ignore the variants that PON-P2 did not classify (52% of total benchmark variants) and only consider how many of the variants that it did classify were correct, we find a positive predictive value of 96% and a negative predictive value of 94%. Thus, while this tool might not be useful for exome screening because too many pathogenic variants would be lost, it can still be an

Hypothetical data set 100 benign variants 10 pathogenic variants	<i>90% sensitive method</i> 9 pathogenic found 1 pathogenic missed	<i>70% sensitive method</i> 7 pathogenic found 3 pathogenic missed
<i>80% specific method</i> 80 benign found, 20 benign missed	9+20 = 29 variants to interpret 9/29 = 31% positive predictive value	7+20 = 27 variants to interpret 7/27 = 26% positive predictive value
<i>60% specific method</i> 60 benign found, 40 benign missed	9+40 = 49 variants to interpret 9/49 = 18% positive predictive value	7 + 40 = 47 variants to interpret 7/47 = 15% positive predictive value

Table 6.4: Estimate of the practical impact in clinical diagnostics of using methods of different sensitivity and specificity on a data set with 100 benign and 10 pathogenic variants.

excellent choice for further investigation of interesting variants. We would therefore emphasize that appropriate tools should be selected depending on the question or analysis protocol used and by taking their strengths and weaknesses into account.

Not surprisingly, we could confirm that the use of gene-specific thresholds instead of genome-wide thresholds led to a consistent and significant improvement of classification performance. This shows the added value of our strategy. Overall performance was slightly lower in genes for which CADD has limited predictive value and even lower in genes with few "gold standard" pathogenicity data available. Evaluating variants in uncharacterized genes is rare in clinical diagnostics, although it may occur when exome sequencing is aimed at solving complex phenotypes or undiagnosed cases. Nevertheless, GAVIN is likely to improve continuously in an increasing number of genes, propelled by the speed at which pathogenic variants are now being reported. The results of this paper are based on the ClinVar release of November 2015 and comprise 2,525 informative gene calibrations, i.e. thresholds for CADD, impact, MAF or a combination thereof. When we calibrate on the September 2016 ClinVar release, we obtain more informative gene calibrations (2,770) with stable gene CADD thresholds (mean pathogenic difference of 0.1%, mean benign difference of 1.1%) and a slight drop in pathogenic MAF (0.00426 to 0.00346). Using these newer calibrations, the benchmark performance of GAVIN increases to 91.7% sensitivity (up from 91.4%) and 78.2% specificity (up from 76.9%). If this trend continues and $(2770-2525)/10 = 24.5$ genes per month are added, we estimate that calibrating all disease genes in CGD (3,316 per Sept. 2016) will take another $(3316-2770)/24.5/12 = 1.86 \approx 2$ years.

With GAVIN, we were also able to demonstrate the residual power of CADD scores as a predictor for pathogenicity on a gene-by-gene basis, revealing that the scores are informative for many genes (these results can be accessed at <http://molgenis.org/gavin>). There are several possible explanations for potential non-informativity of CADD scores. It may have bias towards the in silico tools and sources it was

1 trained on, limiting their predictiveness for certain genomic regions or
2 disease mechanisms[222]. Furthermore, calibration of pathogenic vari-
3 ants could be difficult in genes with high damage tolerance, i.e. having
4 many missense or loss-of-function mutations[165]. In addition, cali-
5 bration may be impaired by false input signals, such as an incorrect
6 pathogenic classification in ClinVar or inclusion of disease cohorts in
7 large databases such as ExAC could misrepresent allele frequencies[320].
8 Lastly, pathogenic variants could have a low penetrance or their effect
mitigated by genetic modifiers, causing high deleteriousness to be tol-
erated in the general population against expectations[66].

The field of clinical genomics is now moving towards interpretation
of non-coding disease variants (NCVs) identified by WGS [385]. A num-
ber of recently introduced metrics, including EIGEN[160], FATHMM-
MKL, DeepSEA[387], and GWAVA, specialize in predicting the func-
tional effects of non-coding sequence variation. When a pathogenic
NCV reference set of reasonable quantity becomes available, a calibra-
tion strategy as described here will be essential to be able to use these
metrics effectively in whole-genome diagnostics.

6.4 Conclusions

GAVIN provides an automated decision-support protocol for classifying
variants, which will continue to improve in scope and precision as more
data is publicly shared by genome diagnostic laboratories. Our approach
bridges the gap between estimates of genome-wide and population-wide
variant pathogenicity and contributes to their practical usefulness for
interpreting clinical variants in specific patient populations. Databases
such as ClinVar contain a wealth of implicit rules now used manually by
human experts to classify variants. Rules on minor allele frequencies,
estimated effect impact and CADD scores are deduced and employed
by GAVIN to classify variants that have not been seen before.

We envision GAVIN accelerating NGS diagnostics and becoming par-

ticularly beneficial as a powerful (clinical) exome screening tool. It can be used to quickly and effectively detect over 90% of pathogenic variants in a given data set and to present these results with an unprecedented small number of false positives. It may especially serve laboratories that lack the resources necessary to perform reliable and large-scale manual variant interpretation for their patients and spur the development of more advanced gene-specific classification methods. We provide GAVIN as an online MOLGENIS[328] web service to browse gene calibration results and annotate VCF files and as a commandline executable including open source code for use in bioinformatic pipelines. GAVIN can be found at <http://molgenis.org/gavin>.

6.5 Methods

6.5.1 Calibration of gene-specific thresholds

We downloaded ClinVar (variant_summary.txt.gz from ClinVar FTP, last modified date: 05/11/15) and selected GRCh37 variants that contained the word “pathogenic” in their clinical significance. These variants were matched against the ClinVar VCF release (clinvar.vcf.gz, last modified date: 01/10/15) using RS (Reference SNP) identifiers in order to resolve missing indel notations. On the resulting VCF, we ran SnpEff version 4.1 L with these settings: hg19 -noStats -noLog -lof -canon -ud 0. As a benign reference set, we selected variants from ExAC (release 0.3, all sites) from the same gene regions with +/- 100 bases of padding on each side to capture more variants residing on the same exon. We first determined the thresholds for gene-specific pathogenic allele frequency by taking the ExAC allele frequency of each pathogenic variant, or assigning zero if the variant was not present in ExAC, and calculating the 95th percentile value per gene using the R7 method from Apache Commons Math version 3.5. We filtered the set of benign variants with this threshold to retain only variants that were rare enough to fall into the pathogenic frequency range.

1 Following this step, the pathogenic impact distribution was calcu-
2 lated as the relative proportion of the generalized effect impact cate-
3 gories, as annotated by SnpEff on the pathogenic variants. The same
4 calculation was performed on the benign variants uniquely present in
5 ExAC. To facilitate this, we annotated ExAC with SnpEff (4.1 L, same
6 settings as above) to get the same impact, transcript and gene nomen-
7 clature as our ClinVar set. Overlapping genes were not an issue be-
8 cause SnpEff variant annotations include the gene symbol to which an
estimated impact is applicable and subsequently only those matching
impacts were considered. The benign variants were subsequently down-
sized to match the impact distribution of the pathogenic variants.

For instance, in the case of 407 pathogenic MYH7 variants, we
found a pathogenic allele frequency threshold of $4.942e-5$, and an im-
pact distribution of 5.41% HIGH, 77.4% MODERATE, 17.2% LOW
and 0% MODIFIER. We defined a matching set of benign variants by
retrieving 1,799 MYH7 variants from ExAC (impact distribution: 2%
HIGH, 23.59% MODERATE, 32.59% LOW, 41.82% MODIFIER), from
which we excluded known ClinVar pathogenic variants ($n = 99$), variants
above the AF threshold ($n = 246$), and removed interspersed variants
using a non-random 'step over' algorithm until the impact distribution
was equalized ($n = 960$). We thus reached an equalized benign set
of 494 variants, having an impact distribution of 5.47% HIGH, 77.33%
MODERATE, 17.21% LOW and 0% MODIFIER).

We then obtained the CADD scores for all variants and tested
whether there was a significant difference in scores between the sets of
pathogenic and benign variants for each gene, using a Mann-Whitney
U test. Per gene we determined the mean CADD score for each group
and also the 95th percentile sensitivity threshold (detection of most
pathogenic variants while accepting false positives) and 95th percentile
specificity threshold (detection of most benign variants while accepting
false negatives), using the Percentile R7 function. All statistics were
done with Apache Commons Math version 3.5. This calibration pro-
cess was repeated for 3,237 genes, resulting in 2,525 genes for which

we learned classification rules involving pathogenic variant MAF, effect impact distribution, CADD score thresholds, or a combination thereof.

On average, CADD scores were informative of pathogenicity. The mean benign variant CADD score across all genes was 23.08, while the mean pathogenic variant CADD score was 28.44, a mean difference of 5.36 ($\sigma = 4.80$). Of 3,237 genes that underwent the calibration process, we found 681 “CADD predictive” genes that had a significantly higher CADD score for pathogenic variants than for benign variants (Mann-Whitney U test, p value <0.05). Interestingly, we also found 732 “CADD less predictive” genes, for which there was no proven difference between benign and pathogenic variants (p value >0.05 despite having ≥ 5 pathogenic and ≥ 5 benign variants in the gene). For 774 genes there was very little calibration data available (<5 pathogenic or <5 benign variants), resulting in no significant difference (p value >0.05) between CADD scores of pathogenic and benign variants. We also found 159 genes for which effect impact alone was predictive, meaning that a certain impact category was unique for pathogenic variants compared to benign variants. For instance, if we observe HIGH impact pathogenic variants (frame shift, stopgain, etc.) for a given gene, whereas benign variants only reach MODERATE impact (missense, in-frame insertion, etc.), we use this criterion as a direct classifier. No further CADD calibration was performed on these genes. In summary, the total set of 3,237 genes comprises 681 “CADD predictive” genes + 732 “CADD less predictive” genes + 774 “little calibration data” genes + 159 “impact predictive” + 178 genes with only pathogenic MAF calibrated + 712 genes without calibration due to less than 2 ClinVar or ExAC variants available + 1 artifact where population CADD was greater than pathogenic CADD. See Additional file 1: Table S1 for details.

1
2
3
4
5
6
7
8

6.5.2 Variant sets for benchmarking

1 We obtained six variant sets that had been classified by human experts.
2 These data sets were used to benchmark the in silico variant pathogenicity
3 prediction tools mentioned in this paper. Variants from the original
4 sets may sometimes be lost due to conversion of cDNA/HGVS notation
5 to VCF.

6 The VariBench protein tolerance data set 7 ([http://structure
7 .bmc.lu.se/VariBench/](http://structure.bmc.lu.se/VariBench/)) contains disease-causing missense variations
8 from the PhenCode[123] database, IDbases[266], and 18 individual
LSDBs[239]. The training set we used contained 17,490 variants,
of which 11,347 were benign and 6,143 pathogenic. The test set
contained 1,887 variants, of which 1,377 were benign and 510 pathogenic.
We used both the training set and test set as benchmarking sets.

The MutationTaster2[304] test set contains known disease mutations
from HGMD[323] Professional and putatively harmless polymorphisms
from 1000 Genomes. It is available at [http://www.mutationtaster.org/info/Comparison_20130328_with_results
_ClinVar.html](http://www.mutationtaster.org/info/Comparison_20130328_with_results_ClinVar.html). This set contains 1,355 variants, of which 1,194
are benign and 161 pathogenic.

We selected 1,688 pathogenic variants from ClinVar that were added
between November 2015 and February 2016 as an additional benchmarking
set, since our method was based on the November 2015 release
of ClinVar. We supplemented this set with a random selection of 1,668
benign variants from ClinVar, yielding a total of 3,356 variants.

We obtained an in-house list of 2,359 variants that had been classified
by molecular and clinical geneticists at the University Medical Center
Groningen. These variants belong to patients seen in the context of
various disorders: cardiomyopathies, epilepsy, dystonia, preconception
carrier screening, and dermatology. Variants were analyzed according
to Dutch medical center guidelines[242] for variant interpretation, using
Cartagenia Bench LabTM (Agilent Technologies) and Alamut[®] software
(Interactive Biosoftware) by evaluating in-house databases, known

population databases (1000G[18], ExAC, ESP6500 at <http://evs.gs.washington.edu/EVS/>, GoNL[244]), functional effect and literature searches. Any ClinVar variants included in the November 2015 release were removed from this set to prevent circular reasoning, resulting in a total of 1,512 variants, with 1,176 benign/likely benign (merged as Benign), 162 VUS, and 174 pathogenic/likely pathogenic (merged as Pathogenic).

From the UMCG diagnostics laboratory we also obtained a list of 607 variants seen in the context of familial cancers. These were interpreted by a medical doctor according to ACMG guidelines[288]. We removed any ClinVar variants (November 2015 release), resulting in 395 variants, with 301 benign/likely benign (merged as Benign), 68 VUS and 26 likely pathogenic/pathogenic (merged as Pathogenic).

6.5.3 Variant data processing and preparation

We used Ensembl VEP (http://grch37.ensembl.org/Homo_sapiens/Tools/VEP/) to convert cDNA/HGVS notations to VCF format. Newly introduced N-notated reference bases were replaced with the appropriate GRCh37 base, and alleles were trimmed where needed (e.g. "TA/TTA" to "T/TT"). We annotated with SnpEff (version 4.2) using the following settings: hg19 -noStats -noLog -lof -canon -ud 0. CADD scores (version 1.3) were added by running the variants through the CADD webservice (available at <http://cadd.gs.washington.edu/score>). ExAC (release 0.3) allele frequencies were added with MOLGENIS annotator (release 1.16.2). We also merged all benchmarking sets into a combined file with 25,995 variants (of which 25,765 classified as benign, likely benign, likely pathogenic or pathogenic) for submission to various online in silico prediction tools.

6.5.4 Execution of in silico predictors

1 The combined set of 25,765 variants was classified by the in silico variant
2 pathogenicity predictors (MSC, CADD, SIFT, PolyPhen2, PROVEAN,
3 Condel, PON-P2, PredictSNP2, FATHMM, GWAVA, FunSeq, DANN).
4 The output of each tool was loaded into a program that compared the
5 observed output to the expected classification and which then calculated
6 performance metrics such as sensitivity and specificity. The tools that
7 we evaluated and the web addresses used can be found in Table 6.5. We
8 executed PROVEAN and SIFT, for which the output was reduced by
retaining the following columns: "INPUT", "PROVEAN PREDICTION
(cut-off = -2.5)" and "SIFT PREDICTION (cut-off = 0.05)". For
PONP-2, the output was left as-is. The Mutation Significance Cutoff
(MSC) thresholds are configurable; we downloaded the ClinVar-based
thresholds for CADD 1.3 at 95% confidence interval, comparable to our
method, as well as HGMD-based thresholds at 99% confidence inter-
val, the default setting. Variants below the gene-specific thresholds were
considered benign, and above the threshold pathogenic. Following the
suggestion of the CADD authors, scores of variants below a threshold
of 15 were considered benign, above this threshold pathogenic. We also
tested CADD thresholds 20 and 25 for comparison. The output of Con-
del was reduced by retaining the following columns: "CHR", "START",
"SYMBOL", "REF", "ALT", "MA", "FATHMM", "CONDEL", "CON-
DELLABEL". After running PolyPhen2, its output was reduced by
retaining the positional information ("chr2:220285283—CG") and the
"prediction" column. Finally, we executed PredictSNP2, which contains
the output from multiple tools. From the output VCF, we used the
INFO fields "PSNPE", "FATE", "GWAVAE", "DANNE" and "FUNNE"
for the pathogenicity estimation outcomes according to the PredictSNP
protocol for PredictSNP2 consensus, FATHMM, GWAVA, DANN and
FunSeq, respectively.

Tool	Used via web address	
MSC	http://pec630.rockefeller.edu/MS/	1
CADD	http://cadd.gs.washington.edu/	2
SIFT	http://provean.jcvi.org/index.php	3
PolyPhen2	http://genetics.bwh.harvard.edu/pph2/	4
PROVEAN	http://provean.jcvi.org/index.php	5
Condel	http://bg.upf.edu/fannsdB/query/condel	6
PON-P2	http://structure.bmc.lu.se/PON-P2/	7
PredictSNP2	http://loschmidt.chemi.muni.cz/predictsnp2/	8
FATHMM	http://loschmidt.chemi.muni.cz/predictsnp2/	
GWAVA	http://loschmidt.chemi.muni.cz/predictsnp2/	
FunSeq	http://loschmidt.chemi.muni.cz/predictsnp2/	
DANN	http://loschmidt.chemi.muni.cz/predictsnp2/	

Table 6.5: The tools used to evaluate our benchmark variant set and the web addresses used through which they were accessed.

6.5.5 Stratification of variants using Clinical Genomics Database

1 We downloaded Clinical Genomics Database (CGD; the .tsv.gz version
2 on 1 June 2016 from [http://research.nhgri.nih.gov/CGD](http://research.nhgri.nih.gov/CGD/download/)
3 /download/). A Java program evaluated each variant in the full
4 set of 25,765 variants and retrieved their associate gene symbols as
5 annotated by SnpEff. We matched the gene symbols to the genes
6 present in CGD and retrieved the corresponding physical manifestation
7 categories. Variants were then written out to separate files for each
8 manifestation category (cardiovascular, craniofacial, renal, etc.). This
means a variant may be output into multiple files if its gene was linked
to multiple manifestation categories. However, we did prevent variants
from being written out twice to the same file in the case of overlapping
genes in the same manifestation categories. We output a variant into
the “NotInCGD” file only if it was not located in any gene present in
CGD.

6.5.6 Implementation

GAVIN was implemented using Java 1.8 and MOLGENIS[328] 1.21 (<http://molgenis.org>). The calibration method is agnostic of the meaning of pathogenic or benign, resulting in thresholds that have balanced sensitivity and specificity. In our diagnostics practice, sensitivity is valued over specificity. We therefore adjusted the CADD and MAF thresholds to shift the balance towards sensitivity at the cost of specificity. We found a setting of 5 (adjustable in source code) achieved >90% sensitivity and this setting was used to generate final thresholds. The genome-wide classification thresholds based on CADD scores < 15 for benign and > 15 for pathogenic matched this high sensitivity. The full table of gene-specific thresholds used can be found at <http://www.molgenis.org/gavin> (for latest release) or Additional file 1: Table S1. They can be used to guide manual variant interpretation or

be re-used in other tools. Source code with tool implementation details can be found at <https://github.com/molgenis/gavin>. All benchmarking, bootstrapping and plotting tools can be found in this repository, as well as all data processing and calibration programs.

6.5.7 Binary classification metrics

Prediction tools may classify variants as benign or pathogenic, but may also fail to reach a classification or classify a variant as VUS. Because of these three outcome states, binary classification metrics must be used with caution. We define sensitivity as the number of detected pathogenic variants (true positives) over the total number of pathogenic variants, which includes true positives, false negatives (pathogenic variants misclassified as benign), and pathogenic variants that were otherwise "missed", i.e. classified as VUS or not classified at all. Therefore, $\text{Sensitivity} = \text{TruePositive} / (\text{TruePositive} + \text{FalseNegative} + \text{MissedPositive})$. We applied the same definition for specificity and define it as: $\text{Specificity} = \text{TrueNegative} / (\text{TrueNegative} + \text{FalsePositive} + \text{MissedNegative})$. Following this line, accuracy is then defined as $(\text{TruePositive} + \text{TrueNegative}) / (\text{TruePositive} + \text{TrueNegative} + \text{FalsePositive} + \text{FalseNegative} + \text{MissedPositive} + \text{MissedNegative})$.

Additional files

Additional file 1: Table S1. GAVIN gene-specific thresholds used in the benchmark. This table can be used to look up thresholds of individual genes and allow variant interpretation by following classification rules as indicated by the column names and provided explanation. (XLSX 198 kb) [available online at genomebiology.biomedcentral.com]

Additional file 2: Table S2. Detailed overview of all benchmark results. Each combination of tool and dataset is listed. We provide the raw counts of true positives (TP), true negatives (TN), false positives

(FP), and false negatives (FN), as well as of pathogenic and benign variants that were "missed", i.e. not correctly identified as such. From these numbers, we calculated the sensitivity and specificity. (XLSX 58 kb) [available online at genomebiology.biomedcentral.com]

Additional file 3: Table S3. *Included in this chapter as Table 6.4*

Additional file 4: Table S4. *Included in this chapter as Table 6.5*

Acknowledgements

We thank Jackie Senior, Kate McIntyre, and Diane Black for their editorial advice. We thank the MOLGENIS team for their assistance with the software implementation and the GAVIN user interface: Bart Charbon, Fleur Kelpin, Mark de Haan, Erwin Winder, Tommy de Boer, Jonathan Jetten, Dennis Hendriksen, and Chao Pang.

Funding

We thank BBMRI-NL for sponsoring the above software development via a voucher. BBMRI-NL is a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO), grant number 184.033.111. We also thank NWO VIDI grant number 016.156.455.

Availability of data and material

The datasets generated during and/or analysed during the current study are available in the GAVIN public GitHub repository, available at <https://github.com/molgenis/gavin>. We have released citable DOI objects for the full source code of both GAVIN, available at <https://doi.org/10.5281/zenodo.155254> and its MOLGENIS dependency at <https://doi.org/10.5281/zenodo.155255>.

Authors' contributions

KV, EB, and MS conceived the method. KV, EB, CD, BS, KA, LF, CW, RHS, RJS, and TK helped to fine-tune the method, accumulate relevant validation data, and evaluate the results. KV and MS drafted the manuscript. KV, EB, CD, BS, KA, AK, LF, FS, TK, CW, RHS, RJS, and MS edited and reviewed the manuscript. All authors read and approved the final manuscript.

1

2

3

Competing interests

4

The authors declare that they have no competing interests.

5

Consent for publication

6

Not applicable.

7

8

Ethics approval and consent to participate

The study was done in accordance with the regulations and ethical guidelines of the University Medical Center Groningen. Specific ethical approval was not necessary because this study was conducted on aggregated, fully anonymized data.

CHAPTER 6. VARIANT INTERPRETATION FOR MEDICAL SEQ.

1

2

3

4

5

6

7

8

Chapter 7

**A bioinformatics
framework for flexible
automation of
downstream genome
analysis**

unpublished

1 K. Joeri van der Velde^{1,2}, Lennart F. Johansson², Ellen Carbo³, Bart
2 Charbon¹, Martine Meems-Veldhuis², Dennis Hendriksen¹, Cleo C. van
3 Diemen², Freerk van Dijk^{1,2}, Fleur Kelpin¹, Kristin M. Abbott², Birgit
4 Sikkema-Raddatz², Richard J. Sinke² and Morris A. Swertz^{1,2,*}

5
6
7
8
1. University of Groningen, University Medical Center Groningen, Ge-
nomics Coordination Center, Groningen, The Netherlands

2. University of Groningen, University Medical Center Groningen, De-
partment of Genetics, Groningen, The Netherlands

3. University Medical Center Utrecht, Department of Genetics, Utrecht,
The Netherlands

* To whom correspondence should be addressed.

Abstract

With the popularity of next-generation sequencing rising, we expect thousands of individuals to soon have whole-genome profiling. However, implementation is a huge challenge for both genome research and diagnostic applications, with the primary roadblock not data acquisition or variant calling, but downstream interpretation.

Interpretation can be sped up using the huge amount of useful information collected by laboratories, public databases and biobanks. Unfortunately, for now, all these sources of useful data cannot be easily integrated and explored in unison. Further, while many innovative analysis methods emerge from research on a regular basis, a lack of standardization makes it difficult to adopt, share, compare and validate them in practice.

Here we report a lightweight framework for genome interpretation pipelines that aims to enable rapid implementation and adaptation of analysis protocols that integrate reference annotation data (e.g. ClinVar, ExAC, GoNL), run best-practice analysis tools (e.g. VAAST or

GAVIN), capture their outputs in a standardized way using a new VCF extension, and use those outputs to generate informative, customizable reports for human interpretation. Clear definitions of tools and standardization of outputs enable interoperability/flexibility and encourage members of the genomics community to jointly develop and reuse framework components in order to rapidly integrate new data and methods and develop and share new best practice protocols. Standardized validation and benchmarking enables rapid testing and uptake of these developments. We used MOLGENIS open source for its implementation but the framework can be readily added in other software.

Implementation of this framework in a genome diagnostic setting shows that we can successfully translate the latest knowledge and methods to medical practice. However, we also envision its usage for different types of genomic studies because the framework is straightforward and offers advantages in terms of storing and sharing results. Software downloads, manuals and source code can be found at www.molgenis.org/genomics.

7.1 Introduction

Sequencing of DNA and RNA has become pivotal in modern life science research, and we can now exploit our understanding of the genome to develop molecular diagnostic tests for many disorders. However, when it comes to making more discoveries in research, there is still a need for better data integration[290] in order to discover new disease genes. At the same time, in genome diagnostics, an increasing number of patients expect a reliable molecular diagnosis based on their complete genome[29] while diagnostic yield is still highly variable[356, 221, 74, 380].

To improve the situation, we can utilize a rapidly growing list of relevant methods, data and knowledge tools. Unfortunately, sharing and uptake of these new analysis resources is difficult because it takes con-

1 considerable effort to investigate, adapt and validate them into diagnostic
or research protocols. This is partly because the quality and reusabil-
2 ity of academic software remains low[274] while commercial software
prefers to incorporate widely used methods that lag behind innovation.
As a result, institutes develop their own interpretation strategies, with
3 varying degrees of success[42], to keep up with a quickly evolving field.

To encourage sharing and incorporation of community-built and
4 commercial tools they should be easy to adopt into the interpreta-
tion workflows used in practice. The first step to achieving this is
to agree on the definitions of each processing step instead of focusing
5 on implementations. This has already happened for NGS variant calling
pipelines[362], which are often composed of multiple commandline tools
6 that are loosely connected by scripts and intermediate output files. In
NGS variant calling, steps such as aligning sequenced reads and variant
7 calling are carried out by tools such as BWA, Samtools and GATK. In
this process, they use standard file formats FASTA/FASTQ, BAM/SAM
8 and VCF (Variant Call Format). This separation allows tools to be in-
terchanged and optimized as a distinct unit of functionality in a bigger
pipeline, driving innovations like Sambamba[332], which is specifically
developed for high-performance filtering of BAM files.

To better serve both research and routine diagnostic needs, we now
wished to expand our NGS pipelines further downstream with the same
modularity for fast integration and exchange of genome analysis meth-
ods such as the recently published GAVIN tool[345]. To facilitate this,
we present here a framework for standardized genomic analysis with in-
terchangeable components. It includes a new intermediate VCF-based
format to capture relevant findings as the basis for interoperability be-
tween the tools in the pipeline, a task for which there was no pre-existing
solution. We have tested and validated this framework with an open
source implementation for genome diagnostics including tools for vari-
ant annotation, interpretation and reporting.

7.2 Results

7.2.1 Framework for downstream genome analysis

We developed a variant analysis framework consisting of a number of intermediate files and tool roles (annotation, analysis, and reporting). The intermediate files use the well-known VCF specification, with additional extensions to support interoperability between alternative tools that fulfill the same role. A new VCF extension called rVCF (Report VCF) is used to describe variants of interest.

We implemented and validated the framework into an existing bioinformatics pipeline within UMCG clinical genome diagnostics laboratory using both existing and newly developed tools. See Figure 7.1 for an overview of the components and data flow within the framework. Below we first describe the file formats and tool roles involved and then provide examples of the implementation, usage and results of the framework as applied to genome diagnostics.

We define the following intermediate files and tool roles:

Regular VCF files capture SNVs and small indels with the option to also capture genomic coverage and structural variants using existing definitions. The VCF 4.2 standard¹ can describe any type of variation. For genomic coverage, the gVCF² extension may be used.

Annotation tools such as SnpEff[60], Ensembl VEP[228], Jannovar[175], VarioWatch[56] and CmdlineAnnotator (presented below). These tools accept VCF files as input and add more contextual information per variant.

¹<https://samtools.github.io/hts-specs/VCFv4.2.pdf>

²<https://software.broadinstitute.org/gatk/guide/article?id=4017>

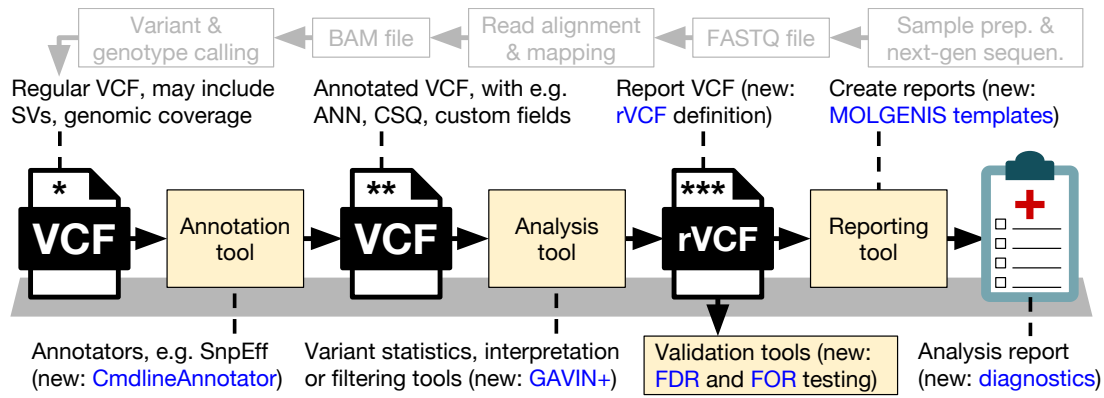


Figure 7.1: Overview of the framework for adaptable automation of downstream genome analysis. New tools and formats developed for the genome diagnostic implementation of the framework are highlighted in blue. Upstream pipeline steps that are not part of the framework are shown in grey.

Annotated VCF files contain the enrichment of variants with additional contextual information from population references, known disease genes/variants, in silico pathogenicity estimates, GWAS, pathways and more. This is stored in a standardized way to ensure tool interoperability (i.e. steps can be changed without needing to rewrite the pipeline). We reuse existing field definitions where possible, such as the SnpEff ANN field³ and Ensembl VEP CSQ field⁴. In addition, we added fields such as CADD_SCALED (for CADD scores) and EXAC_AF (for ExAC allele frequencies) for more annotations.

Analysis tools such as GEMINI[253], InterVar[202], VIKING[232], KGG-Seq[201], VAAST[181] and GAVIN+ (presented below). These tools filter and query annotated VCFs to find candidate variants. Ideally, these tools output their results in Report VCF which can be processed or visualized further by other tools.

Report VCF files store the outcome of a genome analysis, such as GWAS p-values or diagnostic interpretation. This intermediate file format stores relevant findings in a computer-readable format to increase pipeline flexibility by disconnecting results from reporting (see Figure 7.2). We build upon the VCF format by defining a specific extension for results, abbreviated 'rVCF'. This format is fully VCF-compliant but adds an extra INFO field named RLV (relevance) that ensures tool interoperability within this framework. This field contains the explanation for why this variant was thought to be relevant for the question imposed on the original data.

The RLV field was developed with the following criteria in mind: it should (i) be broad enough to allow many types of uses (e.g. for diagnostics, genome research, population studies); (ii) be specific enough for the results captured to be informative, at least for our diagnostic

³http://snpeff.sourceforge.net/VCFannotationformat_v1.0.pdf

⁴http://www.ensembl.org/info/docs/tools/vep/vep_formats.htm

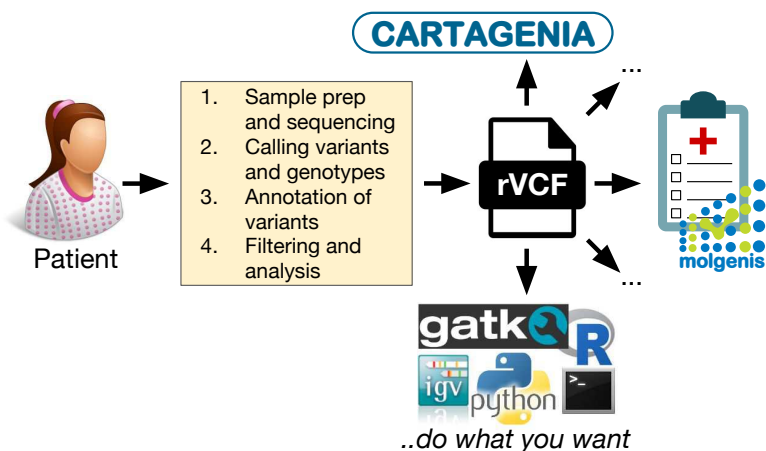


Figure 7.2: Creation and possible applications of the rVCF format. After the output has been generated by the analysis step, the results are stored in an rVCF file. This file can then be further analyzed or used to create a report. We have processed and visualized results in MOLGENIS[328] and Cartagena Bench Lab™ (Agilent Technologies), but any other tool or environment can be used including GATK[344], Integrative Genomics Viewer[341], scripting languages and command line.

use case; (iii) provide all necessary contextual information to explain why a variant is relevant for the question posed; (iv) be structured and simple enough to allow reports to be created in a straightforward way (e.g. by templating); and (v) not contain unnecessary fields that would bloat the specification beyond its intended purpose and make it harder to use. See Table 7.1 for a detailed breakdown of the fields in the rVCF specification and how they can be used for different use-cases.

Reporting tools turn Report VCF into result overviews that can be read by human users and tailored to a specific audience. A database system such as MOLGENIS, for example, can import, store, query and

Field	Definition	
allele	The alternative allele in question, since VCF has multi-allelic sites	
alleleFreq	Reference database minor allele frequencies, e.g. ExAC, GoNL or 1000G	1
gene	Gene name or identifier, e.g. HGNC symbol or MIM accession	
FDR	Any pre-calculated FDR thresholds for this variant, gene or transcript	2
transcript	Transcript used, e.g. SnpEff canonical transcript	3
phenotype	Trait of disorder in question, e.g. trichotillomania, BMI or ethosuximide resistant	4
phenotypeInheritance	Mode of inheritance, e.g. dominant, recessive, additive	4
phenotypeOnset	Age of onset, e.g. pediatric, adult, L2-stage	
phenotypeDetails	Extra phenotype info, e.g. treatment options or literature references	5
phenotypeGroup	Grouping of phenotypes, e.g. cardiovascular, neurological, oncological	6
< many > sampleStatus	How this sample is potentially affected based on genotype and inheritance e.g. HOMZYG, AFFECTED, COMPOUND, DENOVO, CARRIER	7
< many > samplePhenotype	The actual sample phenotype, taken from VCF 'SAMPLE' annotation field	8
< many > sampleGenotype	Sample variant or marker genotypes with possibly quality or probability data	
< many > sampleGroup	Grouping of samples, e.g. case, control, EUR, AFR, infant, adult	
variantSignificance	Type or value of variant significance, e.g. Reported pathogenic, Predicted pathogenic, pval 0.0035	
variantSignificanceSource	Tool or source used to discover this variant, e.g. your lab list, ClinVar, PolyPhen2, CADD, SIFT, GAVIN	
variantSignificanceJustification	The reason why source thought this variant was interesting, or the criteria used by prediction tool	
variantMultiGenic	Denote how this variant is potentially part of digenic or other complex forms of inheritance	
variantGroup	Grouping of variants, e.g. suggestive, significant, iarc_class_5	

Table 7.1: rVCF format. Interoperability is ensured by standardizing interpretation information within the framework using a new 'RLV' field. Within this field, specific interpretation data can be described using the VCF standard INFO sub-field structure. Any sub-fields marked with < many > may contain multiple values, each linked to specific sample identifiers. Multiple RLV values may be present to accommodate multi-allelic variants and overlapping gene annotations.

1 share genomic data and analysis results, then format it for diagnostics
2 reporting from a diagnostics lab in a way that the clinical geneticist can
3 use. The system can be used to create analysis reports from the stored
4 results.

5 **Analysis reports** are the final products of the downstream analysis.
6 They can be clinical patient reports for doctors or statistical genome-
7 wide reports for researchers. Multiple reports for different users and
8 questions may be generated from an analysis, and within a report there
is no limit on flexibility or interactivity, although use of single file doc-
uments with a simple and well-thought out graphical layout is encour-
aged.

6 **Validation tools** (automatically) test and evaluate pipeline results in a
7 standardized way against a gold-standard data set. Such validation is a
8 precondition for diagnostics implementation. Examples of possible val-
idation tests include a false omission test to count how many expected
hits were actually observed in the Report VCF output, a false discovery
test to see how many unexpected hits were found, or a combination of
the two. These tests can be executed on individual level, on gene level,
or across the whole genomes of many individuals simultaneously. Vali-
dations of the pipeline are (automatically) performed whenever software
versions or data sources have been updated.

7.2.2 Implementation for genome diagnostics

We used this framework to implement an automated downstream anal-
ysis and reporting pipeline for genome diagnostics. Below we describe
examples of annotation, analysis and reporting tools that are connected
by using the standardized interpretation data format.

Annotation tool: CmdlineAnnotator

We implemented a high-performance tool that performs annotation tasks that enable variant enrichment. It wraps common genomic annotation sources, such as ExAC[196], 1000 Genomes[18], Genome of the Netherlands[244], ClinVar[191] and CADD[185] in a standardized way. This enables us to quickly add and combine different annotation steps, which was one of our goals with this framework. The tool, MOLGENIS CmdlineAnnotator, can be run on command line making it easy to script into typical bioinformatic pipelines, but we have also implemented a web interface for interactive use. The tool requires no installation other than downloading an executable, and its use is straightforward. MOLGENIS CmdlineAnnotator supports any valid VCF file, handles multi-allelic variants and can match equivalent variants even when their notation is different (see Methods and Materials).

Analysis tool: GAVIN+

We also created a new analysis tool to replace the analysis protocol we had used before. GAVIN+ is a diagnostic interpretation tool that prioritizes DNA variants of potential clinical relevance in the genome. It achieves this by using GAVIN[345], a sensitive tool to predict pathogenic variants based on gene-specific CADD scores calibrated on ExAC, ClinVar and SnpEff. In addition, GAVIN+ matches against candidates against known pathogenic variants in ClinVar but removes potential false positives with a GoNL/ExAC >5% MAF. The tool then queries the Clinical Genomics Database[319] to find affected and carrier individuals depending on sample genotype and mode of inheritance. For uncharacterized genes, the default heterozygous, compound heterozygous and homozygous states are assigned. Depending on the input VCF, additional knowledge is automatically used. This includes trio-aware sample filtering and genotype phasing to check validity of compound heterozygous hits, or reassigning status from, e.g., 'homozygous by compound mutation' back to 'multiple hits on the same allele'. Hemizygous geno-

types for chromosome X and Y are also taken into consideration. The tool even works on mitochondrial genotypes, albeit with limited references. We run GAVIN+ as a command line runnable standalone tool that takes an annotated VCF file and returns a Report VCF file.

Report VCF: novel format

The Report VCF (rVCF) format is used to mark variants that may be of clinical interest. The field that captures the clinical relevance, which is similar to SnpEff's ANN field, consists of a single string with multiple sub-fields separated by pipe symbols. The values are described in the VCF header and include the alternative allele in question, gene, and transcript (just as in ANN), but rVCF also includes associated phenotype, inheritance mode, onset, genotype and affected/carrier status of samples, reason why this variant was relevant and according to whom.

There are currently 19 values within the RLV field, but in practice the notation is quite compact because unused sub-fields add only one character to the notation. See Table 7.2 for examples of comparable VCF INFO field extensions for the same variant, including an example of the RLV field.

Because all applicable sample genotypes in question are now stored in the RLV field, information on the genotype used is left out of the rVCF file. The data reduction achieved by selecting only variants of potential relevance is substantial, and turns out very relevant for usage in a health care setting. For example, on a data set of samples from 69 patients that were sequenced for panel of 96 dystonia genes, we reduce 2,154 variants (6.3 megabyte, MB) to 63 variants (90 kilobyte, kB). A patient exome of 108,004 variants (77 MB) was reduced to 449 variants (625 kB), and a combined VCF with 282 exomes of 790,297 variants (7.3 GB) was reduced to 19,572 variants (17 MB). In research one might still want to retain all data, therefore we developed a helper tool to merge the rVCF files back into the full list of variants and genotypes in the original VCF file, which provides more flexibility for how the format

VCF INFO field	Data example	
Consequence annotations from Ensembl VEP	CSQ=T 5_prime_UTR_variant MODIFIER UROS ENSG00000188690 Transcript ENST00000368797 protein_coding 1/10 7 -1 HGNC 12592;	1 2 3 4
Functional annotations from SnpEff	ANN=T 5_prime_UTR_variant MODIFIER UROS UROS transcript NM_000375.2 Coding 1/10 c.-219C>A 6722 ;	5 6
Relevance annotations from GAVIN+	RLV=T 0.0 UROS 0.0 0.011980830670926517 NM_000375.2 Porphyria congenital erythropoietic RECESSIVE Pediatric DNA7654321:CARRIER DNA7654321:0s1 Reported pathogenic ClinVar NM_000375.2(UROS):c.-219C>A UROS Pathogenic ;	7 8

Table 7.2: Examples of VCF INFO field definitions. The VEP consequence and SnpEff annotation fields describe the functional effects of a variant on genes and transcripts at its locus. The relevance field denotes why this particular variant and effect was included in the analysis result, such as screening for candidates that may explain a clinical phenotype.

can be used. Another helper tool can split the compound RLV field into many separate info fields for ease of import and filtering in other tools.

1 **Reporting tool: patient report for genome diagnostics**

2 We have implemented a report generator to visualize the information
3 contained within rVCF using the MOLGENIS web database, although
4 interested readers can easily write their own. The rVCF files can be
5 uploaded and stored in the database just like any other VCF file. After
6 importing, the genomic data can be browsed, queried, visualized and fil-
7 tered using standard MOLGENIS UI components in the Data Explorer.
8 In addition, users can generate reports based on a simple template lan-
guage (FreeMarker) and use R, Python or JavaScript and web services
to make very interactive reports. These templates can be uploaded,
customized, changed, and reused within the database itself.

We have defined a patient report template that transforms the data into an overview showing the main findings of potential clinical relevance. This report ranks the variants by importance for medical interpretation based on the evidence from the data and how well genes and variants are clinically characterized[229]. Users can apply a number of post-filters within the report if needed. For instance, they may wish to exclude or include certain genes, e.g. those for late-onset disorders in the case of a young patient, or adjust the variant MAF inclusion threshold. See Figure 7.3 for an example of this report.

7.2.3 Validation tool: evaluation for diagnostics

Establishing a molecular diagnosis is only possible when enough data is filtered out to allow human interpretation of the remainder while maintaining reasonably confidence that computational pre-filtering did not remove the causal variant. To estimate the number of variants that were not detected (false negatives or missed), and the number of variants incorrectly implicated (false positives) in our pipeline implementation

molgenis Upload Data Explorer Catalogue Data Integration Plugins Admin About Account [Sign out](#)

RVCF_Dystonia_r1_0 RVCF_Dystonia_r1_0 Delete -

Data Aggregates Charts Annotators **PatientReport**

Select patient Allele frequency: < 2% Late onset exclusion: UMCG Minimum variant impact: Moderate Filter by genes:

PATIENT REPORT FOR [REDACTED]

Name: [REDACTED] MRN: [REDACTED] Patient #: [REDACTED]
 DOB: [REDACTED] Specimens: [REDACTED] DNA #: [REDACTED]
 Sex: [REDACTED] Received: [REDACTED] Family #: [REDACTED]
 Ethnicity: [REDACTED] Referring physician: [REDACTED]
 Indication of testing: [REDACTED] Referring facility: [REDACTED]
 Test: [REDACTED]

GENOME REPORT

STRONG CAUSAL SUSPECTS FOR MONOGENIC DISORDER

Variants are ranked from most relevant (known clinical genes, known pathogenic) towards lesser relevance (uncharacterized genes, predicted pathogenic).

CAT. I: KNOWN PATHOGENIC VARIANT, IN CLINICAL GENE, AFFECTED STATUS
 No variants found.

CAT. II: PREDICTED PATHOGENIC VARIANT, IN CLINICAL GENE, AFFECTED STATUS

Gene Transcript cDNA	AA change Consequence Impact	Genotype Frequency FDR_affected/carr	Disorder Inheritance Onset	Source Justification
CACNA1B NM_000718.3 c.6161A>C	p.His2054Pro missense variant MODERATE	0/1 0.0 35% / 0%	Dystonia 23 DOMINANT Pediatric	GAVIN Variant MAF of 0.0 is rare enough to be potentially pathogenic and its CADD score of 23.9 is greater than a global threshold of 15.
ATP1A3 NM_001256214.1 c.2305C>T	p.Arg769Cys missense variant&splice region variant MODERATE	0/1 0.0 0% / 0%	Alternating hemiplegia of childhood 2 DOMINANT Pediatric	GAVIN Variant CADD score of 34.0 is greater than 23.39 in a gene for which CADD scores are informative.

CAT. III: KNOWN PATHOGENIC VARIANT, IN UNCHARACTERIZED GENE, HOMOZYGOUS GENOTYPE
 No variants found.

CAT. IV: PREDICTED PATHOGENIC VARIANT, IN UNCHARACTERIZED GENE, HOMOZYGOUS GENOTYPE
 No variants found.

171

Figure 7.3: Example screenshot of a patient report generated in MOLGENIS. Sensitive information has been blacked out. Reports are created fully automatically using a template engine on an imported rVCF data set. The buttons at the top offers post-render customization options.

7.2. RESULTS

for genome diagnostics, we set up an automated validation procedure. These validations can be quickly run as many times as needed to check performance and expected output of pipelines whenever they undergo any change or update. This allows efficient testing and uptake of new methods for clinical use, which should increase diagnostic yield.

Estimation of false omission

First we investigate how many of the pathogenic variants that we want to find are actually detected. A missed detection, or false omission, is the most worrisome type of error for molecular geneticists and other experts because a diagnosis cannot be established.

Public benchmark data

We performed a false omission rate (FOR) analysis on the GAVIN+ interpretation tool (version 1.0) using known pathogenic variants. First, we calculated gene-specific FOR on the GAVIN[345] benchmark set, a comprehensive gold-standard consisting of 8,087 unique pathogenic variants from various sources (VariBench, ClinVar, MutationTaster and UMCG clinic), in 1,113 genes. In total we detected 7,598 of the 8,087 (94%) pathogenic variants. For 889 (out of 1,113) genes we recovered all their variants, meaning these genes have a FOR of 0%.

In-house patient variant list

As an additional analysis, we exported the most recent controlled in-house list of interpreted variants from our current clinical diagnostic interpretation software. In this list there were 980 unique variants classified as Pathogenic or Likely pathogenic. We recall 936 of these 980 variants, or 95.5%, consistent with the previous result. See Figure 7.4 for an overview of false omission counts per gene.

In-house patient cases

Lastly, we gathered diagnostic results of 31 patients for whom whole

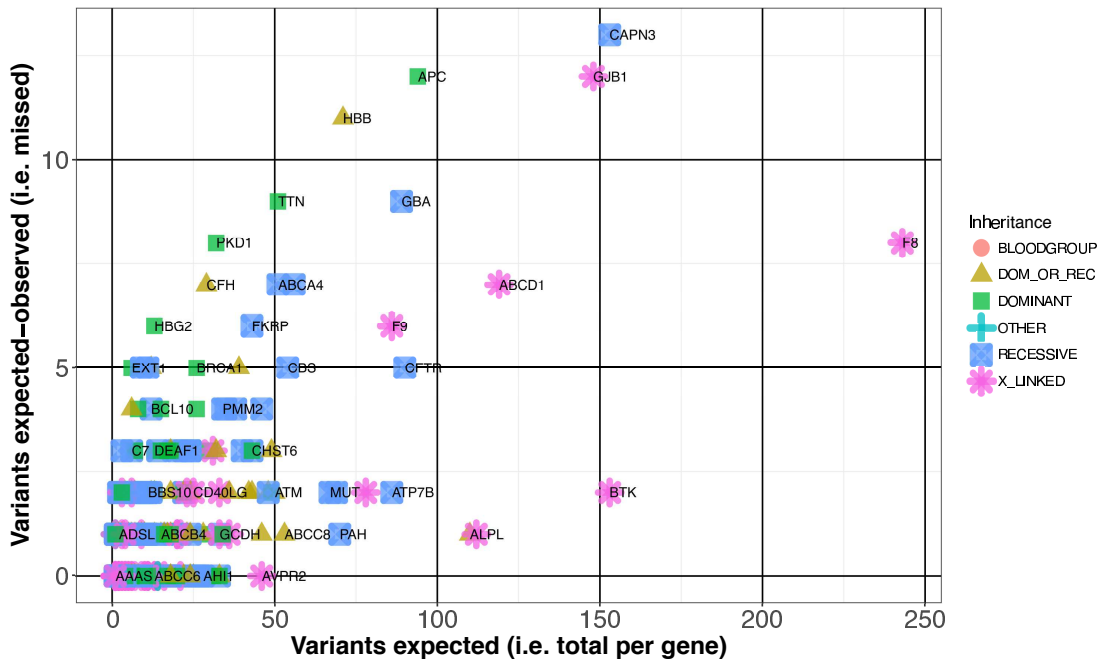


Figure 7.4: Counts of GAVIN+ false omission on known pathogenic benchmark variants. Shown are 1,048 (out of 1,113) genes with known inheritance modes from the Clinical Genomics Database.

1 genome or whole exome sequencing was performed to find the molec-
2 ular cause of their diseases. These include 21 adults from our clinic
3 (2x pulmonary arterial hypertension, 2x familial cancer, 1x familial hy-
4 percholesterolemia, 2x epilepsy, 2x dystonia, 2x epidermolysis bullosa,
5 10x cardiomyopathy) and 10 critically ill newborns and infants from
6 our rapid genome sequencing program[348]. Executing our diagnostic
7 implementation of the framework resulted in a small number of hits
8 that was checked by molecular geneticists. In the adult patients, we
retrieved the causal pathogenic variant in 18 of 21 cases. For the 10
newborns, we found the correct mutation in all 10 cases. In total, we
recalled 28 out of 31 variants, or 90%. The variants that were missed
can be found in Table 7.3.

6 **Estimation of false discovery**

7 To gain confidence in our method to predict a variant to be pathogenic
8 for a patient, we estimated how often it would return false hits in
genomes of healthy individuals. A hit here means a potentially pathogenic
variant under an acting genotype, e.g. heterozygous for a dominant dis-
order or homozygous for a recessive disorder. A very low or even zero
number of positive hits in healthy individuals would increase the chance
that a hit found in a patient is indeed causal for disease.

Public reference genomes

To estimate the gene-specific false discovery rate (FDR) of the inter-
pretation pipeline, we used 2,504 healthy individuals who were whole-
genome sequenced in the 1000 Genomes Project[18]. In total, we used
38,097,906 non-intergenic variants observed across chromosomes 1-22,
X, Y, and MT. Per gene, we counted how many unique samples would
have one or more variants detected as potentially pathogenic. This
may happen in affected status, meaning that the genotype matches
the known inheritance mode of a clinical gene, or if the sample has a
homozygous genotype. For carrier status, the sample is heterozygous,

Gene	Variant	Expert opinion	Reason missed	Gene FOR bench- mark
TTN	c.68225-1G>C (splice acceptor-site variant)	Likely pathogenic	Variant CADD score of 24.3 is less than 26.93 in a gene for which CADD scores are informative	51 / 43 / 15.7%
NPC1	c.3011C>T (protein change p.Ser1004Leu)	Pathogenic	Variant MAF of 9.39E-4 is greater than 4.36E-4	3 / 3 / 0.0%
LDLR	c.-135C>G (5' UTR promoter variant)	Pathogenic	Variant CADD score of 12.88 is less than 16.13 in a gene for which CADD scores are informative	43 / 41 / 4.7%

Table 7.3: Variants that were missed by the GAVIN+ interpretation tool. In two instances, CADD scores were in the benign range and in one case the MAF was just too high to be considered pathogenic. The gene FOR benchmark column shows the false omission test results for the corresponding gene as number of variants expected, number recalled and percentage missed (E / R / M%).

1 which cannot occur for dominant acting genes. When applied to the
2 GAVIN+ interpretation tool (version 1.0) described above, we find a
3 mean affected fraction of 0.26% with a median of 0%, and a mean
4 carrier fraction is 1.85% with a median of 0.72%. For an overview of
5 gene-specific false discovery rates, see Figure 7.5.

In-house patient variant list

6 We also performed an additional assessment on an in-house list of
7 variants classified by experts (for details, see Methods and Materials).
8 These variants were rare enough for careful assessment by clinical genet-
9 icists as candidates for potential disease-causing effects but were found
10 to be plausibly harmless. Of the 9,145 variants classified as benign or
11 likely benign, the tool reports 336 hits, resulting in a false positive rate
12 of 3.7%.

7.3 Discussion

13 We have reported our development a conceptual framework for struc-
14 tured, automated interpretation of variants with interchangeable com-
15 ponents loosely connected by the VCF format. In addition, we created
16 a first implementation of the framework building on the existing open
17 source MOLGENIS software. We integrated existing commandline tools
18 as well as a number of new software tools for annotation, analysis and
19 filtering into an executable pipeline for genome diagnostics. The newly
20 proposed rVCF extension provides an interoperability backbone and,
21 importantly, adds the previously missing link by including information
22 that explains why variants were part of the analysis results. Further-
23 more, all rVCF files can be loaded into a web database that generates
24 a report for researchers that provides an overview of clinically relevant
25 findings for medical use. The content of the reports can be adjusted
26 using option menus, or fully customized by programmers using a tem-
27 plate system. By default, the reports rank variants from most relevant

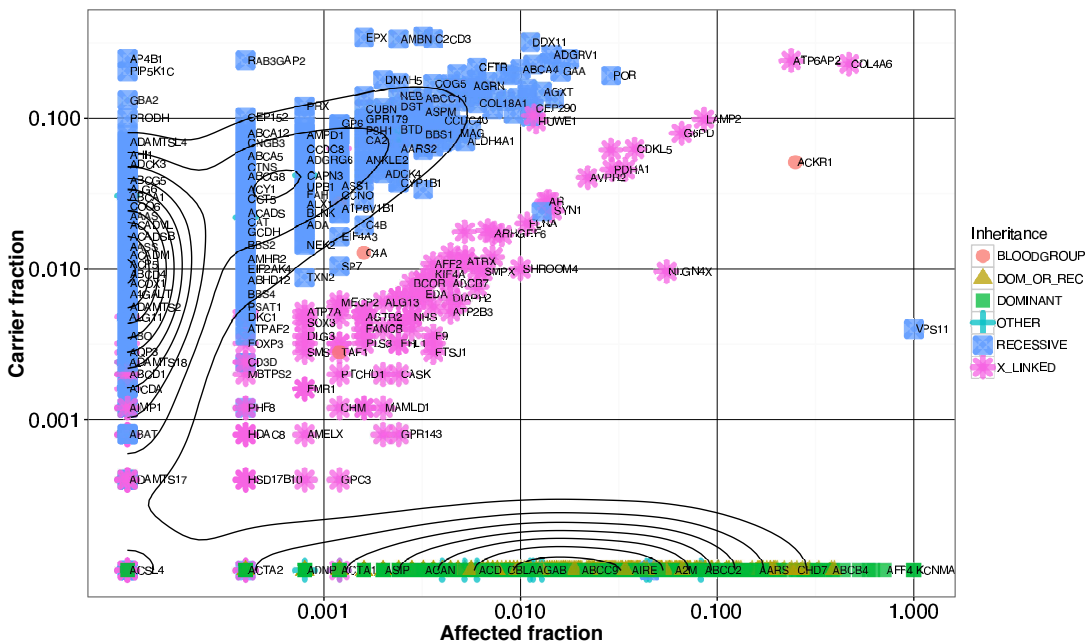


Figure 7.5: Estimations of GAVIN+ false discovery rate using data from the 1000 Genomes Project. Each point represents the fraction of affected vs. carrier samples, i.e. the number of samples for which a potentially pathogenic genotype was detected under that inheritance mode divided by the total number of samples. Only genes with known inheritance modes in the Clinical Genomics Database are shown.

to least relevant.

We have proven our approach with a functional result, but both the framework and its implementation have sparked many ideas for further improvement. Below we discuss the framework itself, evaluate the implementation results and finally provide direction for future work.

7.3.1 Framework considerations

Tool interoperability

The framework defines how tools and data fulfill specific roles within the context of genome interpretation. However, it does not specify if and how the output of one tool can be connected to the expected inputs of another. While some output fields are standardized to a degree, such as SnpEff's 'ANN' field, others are custom or even ambiguous. For example, an 'AF' field probably refers to allele frequency, but does not make explicit from which population reference this frequency is taken. An internationally recognized and maintained list of annotation fields might solve this ambiguity, but this agreement needs support from the entire community and requires software changes in most existing tools. A subtler but easier to achieve strategy would be to use an ontology of genomic annotations to which user can map their software output fields. To use a familiar reference point, Sequence Ontology[89] could be extended with genomic annotations, organized in categories such as 'computational predictions' and 'population allele frequencies'. In this way, different output annotations could point towards a common reference and software implementations could migrate over time without breaking backwards compatibility.

Report VCF format

rVCF specification was defined to capture the relation of any phenotype to any variant including the context in which this association was established. The information included in rVCF was based on clinically

relevant information such as disease phenotype and inheritance, affected individuals, their genotypes, and the reason why this variant was selected. The data captured in this format can be used to generate genomic reports. We believe the specification can be easily adopted by other genomic use-cases such as genome-wide association studies (GWAS), quantitative trait loci (QTL) studies, linkage studies or epigenetic studies. Examples of values for domains other than genome diagnostics, such as model organisms or statistical associations, are shown in Table 7.1. While the format should be able to accommodate all these applications, additional fields may need to be incorporated to capture important information. We are curious to find out what these extra requirements are, and we cordially invite the community to adopt, standardize, provide feedback about and improve upon the format.

7.3.2 Implementation enhancements

Detailed validation output

The validation tools presented produce false positives and false negatives percentages either overall or per gene. However, this does not provide insight into the exact differences between the old and a new pipeline or the types of errors made. To gain more insight, such tools should create a report of their own that compares the errors of the previous pipeline version with the current version. Information such as 'there is a 14% increase in false negative splice variants but a 20% decrease in false positive missense variants' can be very helpful for further improvements or a second tier test to mitigate a particular flaw.

False discovery analysis

The GAVIN+ tool was also applied to 2,504 healthy genomes to estimate the rate of false discovery, i.e. to give an indication of how many falsely accused variants can be expected for each gene (see Figure 7.5). However, since the 1000 Genomes data is based on low-coverage se-

quencing, it is unclear whether the false positives we found are caused by flaws in our method or by false genotypes present in the data. Thus, the resulting plot might reflect sequencing bias or genotype calling errors instead of the actual limitations of our method. To make these estimates more comparable and representative of the data used in a diagnostic sequencing, we need high-coverage whole-genome data from many thousands of healthy individuals from all ethnic backgrounds. With the current data, causal variants might be dismissed because of undeservedly high false discovery estimates of the genes they are in. Another approach that could be used is to use population genomes as a background from which we may calculate diagnostic significance P values for individual patient genomes[365].

Phenotype-matched reports

The framework implementation we have presented uses only genomic information to generate a patient or research report. Of course, the clinical features of the sample offer vital clues as to which gene is likely responsible for the disease. It would therefore make sense to include phenotype-based gene filtering or prioritization to the report. To make this possible, associations of Human Phenotype Ontology (HPO) terms[292] to their known disease genes could be integrated into the system. Users can enter HPO terms that match the phenotypes observed in a patient to shorten their list of candidate genes.

7.3.3 Increasing diagnostic yield

Reducing false positives

False positives in the context of genome diagnostics are harmless variants that are mistaken for pathogenic. Predicting many variants as pathogenic translates to a high workload for human experts, who must manually investigate variants before communicating the results to the patient. To decrease the number of false positives, we can also use spe-

cific populations like those provided within ExAC and 1000 Genomes instead of population-wide allele frequency thresholds. Variants that may be relatively common in one sub-population but not present in others could be filtered out, having low overall MAFs but would logically be considered harmless. However, with fewer individuals used to ascertain the frequency of a variant within a specific population, the potential bias introduced by randomly including non-representative samples also increases. This can be addressed by calculating confidence intervals for allele frequencies, and using those in practice instead of the direct allele frequency values.

In addition, population reference databases such as ExAC offer not only allele counts but also counts of observed homozygous and heterozygous genotypes. For recessive disease genes, the number of observed homozygous genotypes may be more informative than allele frequency, as the variant may be carried within the population without pathogenic effect.

Lastly, another improvement that could reduce the number of false positives is checking for variants or sequencing artifacts previously classified as benign in ClinVar or other sources such as in-house variant lists.

Reducing false negatives

False negatives in the context of genome diagnostics are pathogenic variants that are not detected, a type of error that must be strongly avoided. The number of false negatives could be reduced by built-in consideration of pathogenic founder mutations. These can be common enough for a MAF cutoff filter to accidentally remove them before interpretation. Reporting these variants in international or local databases for use as an interpretation safety net, which is already an optional input for the GAVIN+ interpretation tool, would resolve this issue to a degree. UMCG genome diagnostics uses an in-house list to identify such variants, but an internationally shared and curated list would be

an improvement.

While it is unfortunate that some variants can be missed, estimating miss rates does provide an *a priori* measure of the difficulty of finding pathogenic variants in their respective genes. Variants in genes with high estimated miss rates can be double checked when patient symptoms point toward these genes as likely candidates. This knowledge turns *unknown unknowns* into *known unknowns*, empowering the interpretation process and shedding light on potential uncertainties.

Using structural variation

This study was focused on processing single-nucleotide variants (SNVs) and small indels, but the structural variation (SV) output of tools such as Manta[55] and Delly[281] would also be an important addition to automated interpretation framework. The regions indicated to be deletions, insertions, duplications, inversions or translocations may be complemented with any SNVs and small indels called by conventional variant callers to increase overall diagnostic yield or to obtain a more complete genomic picture for research projects.

7.4 Conclusion

We have developed and evaluated a framework for structured, stepwise downstream variant analysis. The aim was a structure that links the different tools and data in this process to enable exchange, reuse and improvement of components of equal scope across institutes. The novel rVCF intermediate format allows standardized representation of analysis results, and these can be used to quickly create patient or research reports. We expect that this modular structure will also make it easier to integrate the additional omics technologies that will soon support next-generation sequencing, such as allele specific expression and splicing effects from RNA-sequencing expression, epigenetic markers and metabolomics.

Geneticists have already adopted and learned to rely on best-practice pipelines for NGS variant- and genotype-calling, partly because these tools have matured by the efforts and support of the community but primarily because there is too much raw data to assess by hand. Given the ever-growing numbers of whole-genome patient sequences we must extend this mentality further downstream towards analysis and interpretation. To deal with false positives and negatives more effectively, our approach includes automated validation and error estimation tools applied to large benchmark sets to assess the quality and pitfalls of such a pipeline. We have shown that our diagnostic framework implementation combines and automates the latest knowledge, tools and practices to reduce the time and effort spent on easily resolvable patient cases, a times savings that will provide human experts with the time they need to solve puzzling cases that will further extend our knowledge of human genetics.

7.5 Methods and Materials

7.5.1 MOLGENIS annotation tool

We developed MOLGENIS CmdlineAnnotator as an extensible annotation framework that works seamlessly in both web and commandline environments. It has some smart features to maximize the match of input variants (patient) to resource data (context) such as population references. As an example we can consider the case where the annotation resource has a variant $AGG > A, ATGG$ to denote both the deletion of GG and the insertion of T and our input VCF file has a $A > AT$ variant at the same location. Though this variant is present in the resource, it would likely be missed because the notation is different. The CmdlineAnnotator matches this variant by removing but remembering GG from AGG to match the input A . The input alternative allele AT would subsequently be postfixed with GG to form $ATGG$, which matches against the resource successfully. These slight but rele-

1 vant differences can be responsible for misinterpretation during analysis.
2 CmdlineAnnotator version 1.21.1 source code and release are available
3 at [https://github.com/molgenis/molgenis/releases/t](https://github.com/molgenis/molgenis/releases/tag/v1.21.1)
4 ag/v1.21.1.

5 **7.5.2 Population reference for false discovery analysis**

6 We downloaded the 1000 Genomes Project phase 3[18] release data
7 from [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/rele](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/)
8 ase/20130502/. We annotated genes using SnpEff version 4.2 with
9 these settings: hg19 -noStats -noLog -lof -canon -no-intergenic -ud
10 0. Allele frequencies of Genome of the Netherlands[244] release 5 and
11 ExAC[196] release 0.3, and CADD scores[185] version 1.3 were an-
12 notated using MOLGENIS CmdlineAnnotator v1.21.1. All VCF FOR-
13 MAT fields except genotype (GT) were removed from chromosome
14 X and Y to harmonize the data with chromosomes 1-22. Also, the
15 genotyped samples for chromosomes Y and MT are different from
16 those in chromosomes 1-22. We wrote a simple tool (SampleFix-
17 For1000GchrYandMT.java) to harmonize the samples for Y and MT,
18 available at <https://github.com/molgenis/gavin-plus>. These
19 fixes now allow all data to be merged by stripping the headers and
20 concatenating the files in the order 1-22, X, Y and MT. The header
21 from chromosome 1 was added to the merged file with a few INFO
22 lines added that were specific for other chromosomes: LEN, TYPE and
23 OLD_VARIANT from chromosome X, and VT from chromosome MT.
24 The resulting file was compressed to 8.4G with bgzip and indexed us-
25 ing tabix -p vcf. It contains 38,097,906 non-intergenic variants and
26 95,397,156,624 genotypes. The file is available at [http://molgeni](http://molgenis.org/downloads/gavin/)
27 s.org/downloads/gavin/.

7.5.3 Pathogenic variants for false omission analysis

We used the GAVIN[345] variant classification benchmark set available at <https://github.com/molgenis/gavin>. This set comprises 25,995 variants from which we select 8,087 pathogenic variants after filtering duplicate genomic positions. We annotated these variants with SnpEff 4.2, ExAC r0.3, CADD 1.3 and GoNL r5 using MOLGENIS 1.21.1 CmdlineAnnotator. A heterozygous genotype was added to each variant to enable running of the GAVIN+ automated interpretation tool. This dataset is available for download at <http://molgenis.org/downloads/gavin/>.

We also used a list of variants interpreted by molecular and clinical geneticists at the University Medical Center Groningen according to Dutch medical center guidelines[242]. More details on the interpretation criteria are provided in Van der Velde *et al.*[345]. This list contained 980 likely pathogenic or pathogenic variants and 9,145 benign or likely benign variants after filtering for duplicate genomic positions. These variants were annotated and processed as above, and access to these data can be requested.

7.5.4 GAVIN+ interpretation tool

We developed the GAVIN+ tool to automate sample genome interpretation. In a stepwise process, interesting variants are selected (based on hits from GAVIN[345], ClinVar[191], or a user-supplied list of variants), followed by a MAF filter, match of genotype to gene inheritance mode, checks for compound heterozygosity and the use of trio or duo sample genotype phasing and de novo variant finding. GAVIN+ supports multiple alleles per variant that may be present in multiple overlapping gene annotations. It is implemented in Java 1.8 (<https://www.java.com>) as free open source software at <https://github.com/molgenis/gavin-plus>. A comprehensive TestNG (<http://testng.org>) test suite ensures correctness and allows further development with a limited chance of introducing bugs. Dependencies are managed by

1 Apache Maven (<https://maven.apache.org/>). A precompiled,
2 command line runnable version of the tool can be downloaded at <http://molgenis.org/downloads/gavin/>. A demo and manual
3 are available, as are the bundled resources needed to run the tool: Clin-
4 Var (any variant matching 'pathogenic' from combined TSV and VCF
5 representations of 11 oct. 2016, 1.5 MB), CGD (version 11 oct. 2016,
6 380 kB), FDR (version 1.0, 946 kB) and GAVIN calibrations (r0.3, 331
7 kB). The implementation insures high performance by using a streaming
8 architecture with as little in-memory buffering as necessary and as much
9 output as possible immediately written to disk. This results in speeds
10 of millions of genotypes per second. GAVIN+ can analyze 300 whole
11 exomes in 2 minutes or the full 1000G FDR analysis (95,397,156,624
12 genotypes) in 2 hours on commodity hardware.

7.5.5 Running false omission analysis

8 We ran the GAVIN+ tool in a first pass (using `-m CREATEFILEFOR-`
9 `CADD`) to get a list of 34 indel variants that are not yet scored by
10 CADD. These variants were then scored by a local CADD 1.3 which
11 was installed offline following the instructions at <http://cadd.gs.washington.edu/download>. After scoring, GAVIN+ was run for
12 a second and final pass using arguments: `-i GAVIN_FOR_benchmark_`
13 `goldstandard_nodup_gonl.vcf -g bundle_r1.0/GAVIN_calibrations_r0.3.tsv`
14 `-c bundle_r1.0/clinvar.patho.fix.11oct2016.vcf.gz -d bundle_r1.0/CGD_`
15 `11oct2016.txt.gz -a fromCadd.tsv -f bundle_r1.0/FDR_allGenes_r1.0.tsv`
16 `-m ANALYSIS -o RVCF_GAVIN_FOR_benchmark_goldstandard_nodup_`
17 `gonl_r1.0.vcf`. This was followed by a simple tool (`FOR.java`, available at
18 <https://github.com/molgenis/gavin-plus>) to report the
19 false omission rate using original VCF and the rVCF file produced by the
20 GAVIN+ tool. For each gene, we counted the number of pathogenic
21 variants in the original VCF that we expect to recover. We divide this
22 number by the observed number of variants in the rVCF as a missed
23 fraction for each gene to estimate how well GAVIN+ detection works

for that gene. All files and results are available at <http://molgenis.org/downloads/gavin/>.

7.5.6 Running false discovery analysis

We ran the GAVIN+ tool in a first pass on the 1000G population reference set and got 1,136,050 variants that were not yet scored by CADD. The local CADD 1.3 tool scored 1,110,509 (97.75%) of these, meaning that 25,541 of 38,097,906 total variants (0.067%) remained un-scored. This allowed GAVIN+ to be run in a second and final pass (using -m ANALYSIS) of the data, resulting in an rVCF file with 381,482 selected variants. To obtain false discovery rate estimates, we wrote FDR.java, a simple tool that assumes that the hits in the rVCF file are false. These hits are counted per gene as the number of samples that would have at least one matching genotype under one of two inheritance modes. A sample is counted as affected when the match is homozygous or compound heterozygous or heterozygous in a known dominant disease causing gene, and counted as carrier when heterozygous in either a recessive disease causing gene or an unknown gene. A check on phasing may revert compound heterozygotes back to carrier heterozygous multihit status. The FDR tool outputs a list of 19,230 genes, each with four columns: affected count, carrier count, affected fraction ($\text{aff.count}/2504$) and carrier fraction ($\text{carr.count}/2504$). Among these 19,230 genes, we find 8,399 with one or more affected samples, 17,878 with one or more carriers, and an overlap of 7,047 genes ($17878+(8399-7047)=19230$). This list is further processed to include all 26,023 SnpEff gene names present in the original VCF for which no affected or carrier status was detected. This was done by first extracting all gene names using GetAllGeneNamesFromVCF.java, followed by using CombineFDRwithAllGenes.java to get the final FDR result file with 26,044 genes. Note that 21 mitochondrial genes were present in the rVCF file due to ClinVar hits, but these were not annotated by SnpEff in the original VCF file, hence the final result includes

slightly more genes. All result files and intermediates are available at <http://molgenis.org/downloads/gavin/>.

7.5.7 Visualizing FOR and FDR analysis results

We wrote a small R script (FDR_plot.R) that uses the FDR result file with 26,044 genes as well as gene annotations from Clinical Genomics Database[319] (file date: August 31, 2016) to plot the observed affected versus carrier fractions. After merging with CGD we could plot 3,232 of these genes with known inheritance modes. The color and shape of the points is dependent on the inheritance mode, scales are base 10 logarithmic. Zero values were replaced with 1e-4 to allow logarithmic scale. Some genes appear in unexpected plot locations, but can be logically explained. For instance, CSF2RA appears in the recessive band while being on chromosome X but it is located in the X-PAR1 region. Conversely, TENM1 appears in the X_LINKED band with a blue icon because it has a mistakenly RECESSIVE annotation. For only the plotted CGD genes, we find a mean affected fraction of 1.51% with a median of 0.08% and a mean carrier fraction of 1.43% with a median of 0.20%. Another R script (FOR_plot.R) visualizes the FOR result file with 1,113 genes. After merging with CGD we can plot 1,048 of these genes with known inheritance modes. Again the color and shape of the points is dependent on the inheritance mode. All scripts and required files are available at <https://github.com/molgenis/gavin-plus>.

7.5.8 MOLGENIS reporting tool

To store and visualize the rVCF files produced, we used a modified version of MOLGENIS 1.21.2[328]. Users can simply import rVCF files via the standard Data Importer, which can then be viewed in the Data Explorer. To create the custom reports, we used the FreeMarker template engine (<http://freemarker.org>). The templates should be

placed in the molgenis/molgenis-app/src/main/resources/templates/ folder, and adhere to this naming scheme: view[reportname]entitiesreport.ftl. For instance, a PatientReport should be named viewPatientReport-entitiesreport.ftl. The templates are then added as an additional tab for a dataset by clicking the Configure button, and in the Reports box define this link using [reportname]:[datasetname]. For instance, we connect the PatientReport to an rVCF named Cardio using Cardio:-PatientReport. Multiple links can be separated using commas, e.g. My-ResearchExomes:GeneReport,Dystonia:PatientReport,Cardio:PatientReport. The used MOLGENIS and created templates are available at <http://github.com/joerivandervelde/molgenis>.

Acknowledgements

We thank Kate McIntyre for editing.

Funding

We thank BBMRI-NL for sponsoring the above software development via a voucher. BBMRI-NL is a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO), grant number 184.033.111. We also thank NWO VIDI grant number 016.156.455. We acknowledge the support from the Netherlands CardioVascular Research Initiative: "the Dutch Heart Foundation, Dutch Federation of University Medical Centres, the Netherlands Organisation for Health Research and Development and the Royal Netherlands Academy of Sciences" for the GENIUS project "Generating the best evidence-based pharmaceutical targets for atherosclerosis" (CVON2011-19).

Authors' contributions

1 KJV, MS designed the conceptual framework. KJV, BC, DH and FK
2 developed the software and processed the data. MMV, CCvD, KMA
3 and BSR performed patient DNA variant interpretation and diagnostic
4 validation. KJV drafted the manuscript with input from LFJ, EC, DH,
5 FvD, FK, KMA, BSR, RJS and MAS. All authors contributed to the
6 development of the framework and evaluated one or more components.
7 All authors read and approved the final manuscript.

Competing interests

8 The authors declare that they have no competing interests.

Chapter 8

Discussion and Perspectives

1
2
3
4
5
6
7
8

Abstract

1 The enormous wealth of data generated in the life sciences presents us
2 with incredible opportunities to improve medical genetics, but also with
3 an equally big bioinformatics challenge to fulfill this promise. The scope
4 and complexity of this challenge is exacerbated by the increasing speed
5 at which new methods, tools and data sets become available across a
wide range of disciplines ranging from statistics to computer science
and from model organisms to clinical validation. In this thesis, I con-
tributed a number of bioinformatics models, methods, and integrated
systems thereof as infrastructure to enable rapid translation of these
new resources into medical applications.

Introduction

6 In this thesis I first developed new data management and processing
7 models in chapter 2 and implemented these to store, integrate and
8 visualize model organism data in chapter 3. By connecting human
disease phenotypes to model organisms, I showed that this approach
can be used to discover new disease leads in chapter 4.

I then investigated how existing methods for computational esti-
mation of variant deleteriousness can be used to predict pathogenicity
classification with high precision and recall in chapter 5. In chapter 6
I generalized this method to thousands of genes, and molecular geneti-
cists can now use an integrated online system to classify patient DNA
variants in the context of large reference data.

While new methods and data become available at ever increasing
speeds, their implementation in clinical practice lags behind because
of the time needed to validate and implement them into clinical prac-
tice. To implement and validate new analysis protocols rapidly and
efficiently in research and clinical practice, we need a streamlined auto-
mated pipeline for data processing, decision making, and reporting of
results. I therefore developed a software system for structured variant
interpretation, currently adopted in routine diagnostics, that combines

new methods and models with existing tools and knowledge databases in chapter 7.

In this chapter I will consider the meaning and implications of the work presented in this thesis and look to the future potential and on-coming challenges of the field. I first discuss in section 8.1 the successes and challenges of flexible models to capture, integrate (8.1.1), share and reuse life science data (8.1.2). I address how more people could benefit from these innovations (8.1.3), and if perhaps smarter technologies are needed to get more value from complex data (8.1.4).

Second, I consider the challenges of method development, in which data plays a critical role in section 8.2. The importance of data quantity and quality is exemplified (8.2.1), as well as pitfalls in method benchmarking (8.2.2). I highlight future ways to overcome difficulties in finding (8.2.3) and running appropriate methods (8.2.4).

Finally, I examine how to better implement complex systems for application to medical genetics in section 8.3. Crucial aspects such as sharing of workflows (8.3.1) and community expertise (8.3.2) are discussed, and I suggest future work on multi-omics analysis (8.3.3) and semantic protocols (8.3.4). For each of the sections I will summarize key question and key points.

8.1 Flexible models for life science omics data

Life science data can be stored, managed and queried in a multitude of different ways, each with their own advantages and drawbacks. In this section, I discuss various models and aspects involved in confronting the life science data challenge.

1
2
3
4
5
6
7
8

Key question and points of this section

How should data models be designed to integrate, reuse and share data from life science experiments in order to extract knowledge that benefits medical genetics?

Key points

- We have researched and evaluated different data integration models that make genotype-phenotype analyses more contextual and insightful (8.1.1).
- Some applications demand greater data model flexibility, but the loss of predefined data classes presents new issues (8.1.2).
- Ontologies can be used to give explicit meaning back to the data, allowing methods and tools to again perform cross-set analyses (8.1.2).
- Managing growing quantities of life science data in spreadsheets and flat files is problematic, therefore more should be done to increase uptake of better alternatives (8.1.3).
- Data is being shared, but without smart storage algorithms it remains difficult and time consuming to ask even basic questions (8.1.4).

8.1.1 Integration of heterogeneous omics data

Methodological (re)use of all available life science data is difficult. Making data suitable for reuse requires structured storage with sufficient metadata to allow retrieval and interpretation, which is troublesome and time consuming to achieve. In addition, various storage paradigms are needed to complement traditional databases be-

cause data volumes are large and database structures often complex and highly heterogeneous[329], which limits interoperability and integrateability of these data. Data covers a wide variety of phenotypic measurements including age, gender and height; answers given in questionnaires; detailed clinical observations; and large-scale molecular measurements such as DNA sequencing, gene expression, metabolomics and proteomics. Moreover, these data may be collected from many subjects, at different timepoints, from multiple tissues, using a various wet and dry laboratory protocols. Finally, rapid development of new profiling techniques and analysis methods requires data infrastructure to rapidly change to accommodate.

The eXtensible Genotype and Phenotype model

We investigated how data models and supporting software systems should be (re)designed to accommodate this heterogeneity and to be effective in handling these data. The first result was the XGAP data model developed to capture a variety of life science data that is described in chapter 2 and summarized in Box 1.

1
2
3
4
5
6
7
8

Box 1: Brief explanation of the XGAP data model

1 The innovative XGAP model facilitates integration of a wide
2 range of data sources into one conceptual framework by en-
3 abling researchers to define observations as any combination of
4 subjects (the thing being observed) and trait (the measurable
5 quality). As a result, data can be flexibly stored, eliminating
6 the need to define the exact storage requirements for exper-
7 imental data, which is impossible beforehand and moot after
8 project completion. For example, definitions of concepts such
as 'Gene', 'Marker', 'Individual' or 'Metabolite' remain constant,
but they can be used to create any dataset while the application
is running. Then gene expression data can be defined as 'Gene'
 \times 'Individual' with a numeric value at each combination and a
genotype map as 'Marker' \times 'Individual' with categorical values,
e.g. 'AA', 'CC', 'AC'. After QTL analysis, the result then can
be defined as 'Gene' \times 'Marker' where each value indicates the
statistical strength of the association between gene expression
and genotype.

XGAP's flexibility is a sharp contrast to traditional applications built on relational databases that need to be taken offline for redesign when a new data modality comes in. A life science database can now be created when a research project starts and experimental measurements plus contextual data from external providers, e.g. gene annotations or pathway definitions, added naturally as the project progresses.

We implemented XGAP into the MOLGENIS[328] software toolkit that generates database software infrastructure from data model and user interface specifications. By combining the MOLGENIS software and XGAP datamodel as a foundation, we now can implement generic life science databases that can handle almost any omics/QTL data. We published this system as xQTL workbench in chapter 3. This system can handle any genotype-to-phenotype experiments and can be used as

a template to create data portals for specific research areas. We demonstrated the added value of xQTL first in *C. elegans* research[315], then added a translation from model organism to human disease genetics. The resulting WormQTL^{HD} database[346] is described in chapter 4. Its built-in visualization tools can be used to find clues for the molecular workings of human disease in almost 100 online-accessible datasets.

When even more flexibility is needed

The XGAP model's power comes from its fifty-fifty balance between static data structure (the underlying structure does not change when new data is loaded) and dynamic modeling (the structure can be easily extended and adapted and depends on the genes and phenotypes in the experimental data). The static structure acts as a stable template that enables development of new software tools that then will work on all data loaded because the tools know what data structure to expect. At the same time new data modalities can be rapidly accommodated using the flexible structure.

However, there are drawbacks to using (partially) static data structures. It requires users to familiarize themselves with a data model, and limits the attributes that can be used to express information while simultaneously burdening the user with often unnecessary attributes. An attribute is a property of the object type being described. An object of the type 'car', for instance, can have the attributes 'brand', 'model', 'color', and 'year of construction'. Pre-defining the attributes for data in life sciences can be very convenient for some uses, for instance if they want to automatically connecting genomic data to a genome browser.

Nevertheless, we found that more flexibility is better in some other cases. Therefore, when we developed MOLGENIS 2.0, we created the even more flexible EMX (entity model extensible) storage model. In MOLGENIS 2.0, the user uploading the data has full control over all aspects of the data model, meaning that tabular data can be defined at column and data type level, as well as cross-linked in any way. This

allows use of XGAP or other data models if desired.

1 Together with collaborators, we evaluated the models in many other
2 online databases for various domains within life sciences[4]. Figure 8.1
3 shows the two main paths in evolution (XGAP and EMX data modeling)
4 and a selection of currently active software applications. These appli-
5 cations are powered by the MOLGENIS platform which allows flexible
6 generation and configuration of database application.

8.1.2 Making omics data reusable across systems

4 The most recent versions of MOLGENIS allow the user to define and
5 import any data structure. This is highly appreciated by users as it
6 provides complete freedom to upload whatever data they like. MOL-
7 GENIS uses a simple tabular format that contains both the data itself
8 and its meta-data describing the flexible attributes with strongly-typed
values. After import, data set columns can be added, deleted or redef-
ined if needed. Data values can cross-reference to other datasets or
rows within datasets to create a complex ad hoc data model. Users can
thus create a database perfectly tailored to their storage needs that can
be re-tailored whenever those needs change. The learning threshold of
creating and managing such a database is low, and it encourages people
to upload and connect any data they find relevant. However, because
the semantics of the data are now no longer explicit, this presents a
new challenge.

Need standard model building blocks

While it is now much easier to bring data together into one MOLGENIS
system, the cost of this freedom is the loss the explicit meaning of the
data, as compared to XGAP. This greatly limits data re-use because
data cannot be easily integrated with other datasets and analysis tools
cannot be used. In other words, for data to be reusable its semantics
must be clear so humans as well as software can understand and know

8.1. FLEXIBLE MODELS FOR LIFE SCIENCE OMICS DATA

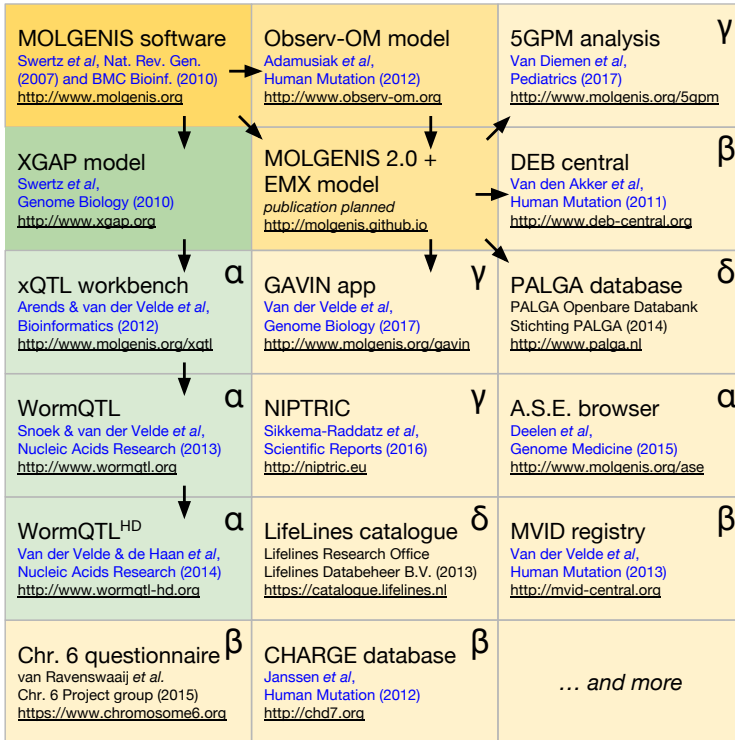


Figure 8.1: The MOLGENIS family of software, data models and applications. The core software and models (top left corner) form the basis for a variety of applications, with those in green boxes directly built from the XGAP data model. The types of applications are indicated with an α for research portals, β for patient registries, γ for diagnostics support, and δ for biobank catalogues. Blue text indicates a peer-reviewed published article.

1 what to do with it. Users A and B may both upload a set named
2 'Genes', but the system does understand whether these actually refer
3 to the same concept. The same problem applies to any attributes within
4 these data sets. Perhaps they both have a 'Position' column, but one
5 may be measured in centimorgan and the other in base pairs. Even
6 if the units are the same there may be crucial contextual differences,
7 for example base pair positions that are derived from different genome
8 builds. This uncertainty makes it impossible to reuse tools and methods
because certain necessary attributes cannot be automatically connected.

To deal with this issue, the latest MOLGENIS versions use small data models to enable implementation of standard analysis protocols without limiting the flexibility of the system. Some data has a very predictable and often reoccurring structure. A good example is the 'genomic location' tied to variants and other features of the DNA. The attributes of these genomic locations are usually chromosome, position, genome build, reference and alternative base, which are automatically mapped to a micro-model and understood by the system so that tools such as the genome browser can immediately visualize the data.

Use of ontologies

The micro-model solution is an efficient way to deal with highly predictable data. However, most flexible data will still not be understood by the system and therefore not easily connected, integrated and analyzed. Note that the difference between a data model and an ontology is subtle but significant, and therefore we clarify this in Box 2.

Box 2: Difference between data model and ontology

Ontologies can be thought of as dictionaries that define terms, with the special addition that terms can be related to each other. Data models can be thought of as floor plans that offer explicit structure, while the definitions of the concepts used are implicit. For example, a floor plan could specify how a living room is connected to the other rooms and show the arrangement of the furniture, whereas an ontology would define the concepts of a living room and elaborate on known types of furniture. People who used different building plans to construct their house can use the ontology to refer to their now shared understanding of a living room, and find out if they both have a couch in it.

To overcome this problem, we enable users to annotate or 'tag' their free-form data with additional meta-information that explains what the data means. The database must contain concepts such as 'Position' measured as 'Integers' on reference build 'GRCh37' for 'Homo sapiens'. These concepts can then be mapped on the sets, rows and columns of data sets imported. Tools use these tags to understand the meaning of the data, which ensures that queries and tools can be reused between heterogeneous datasets.

The meta-data used to annotate data with meaning are called *ontologies*, which are common dictionaries of agreed-upon, well-defined terms and their relationships. Building an ontology through input from an international community of domain experts ensures a shared point of reference for the clear communication of meaning. Annotating data with ontologies offers many advantages in terms of connectivity and reasoning, but requires a significant investment of expert's time. To drastically lower this burden, we have developed strategies for automated matching of terms to ontologies[255] and for smart matching values to coding systems[256]. The resulting systems allow users to harmonize data items and values in a fraction of the time it would nor-

1
2
3
4
5
6

7

8

mally cost, thereby enabling pooled data analysis for higher statistical significance. The MOLGENIS online data platform can assist users to quickly interconnect their data via semantic annotation[256].

Ontologies for medical genetics

Structured ontologies also provide computational advantages. Ontologies may be expressed as graphs of connected terms, typically a tree-shaped hierarchies where a broad root term branches out into more specific terms. Famous ontologies used in medical genetics include ICD¹ and SNOMED-CT² for diseases, and Gene Ontology[16] to describe genes.

The Human Phenotype Ontology[292] (HPO), shown in Figure 8.2, is an ontology of particular usefulness. HPO terms are becoming an integral component of clinical work in many medical centers, including the UMCG[348], where they are used to convey patient symptoms to the decision support software. Symptoms can be expressed in HPO terms that may be broad or specific. Likewise, diseases may be expressed as collections of multiple HPO terms. Computer algorithms can accept HPO terms as inputs and take advantage of the underlying graph structure, for example, to find a known disease that best matches a set of input symptoms. The paths between the terms are used as a distance measure, and advanced methods can even calculate semantic distance between collections of terms weighted by information content[285]. This clears the way for advanced tools[125] that can guide or support a genomic diagnosis with a robust phenotypic match of patient symptoms to a known disorder for which the causal gene is known.

¹<http://www.who.int/classifications/icd/en>

²<http://www.snomed.org>

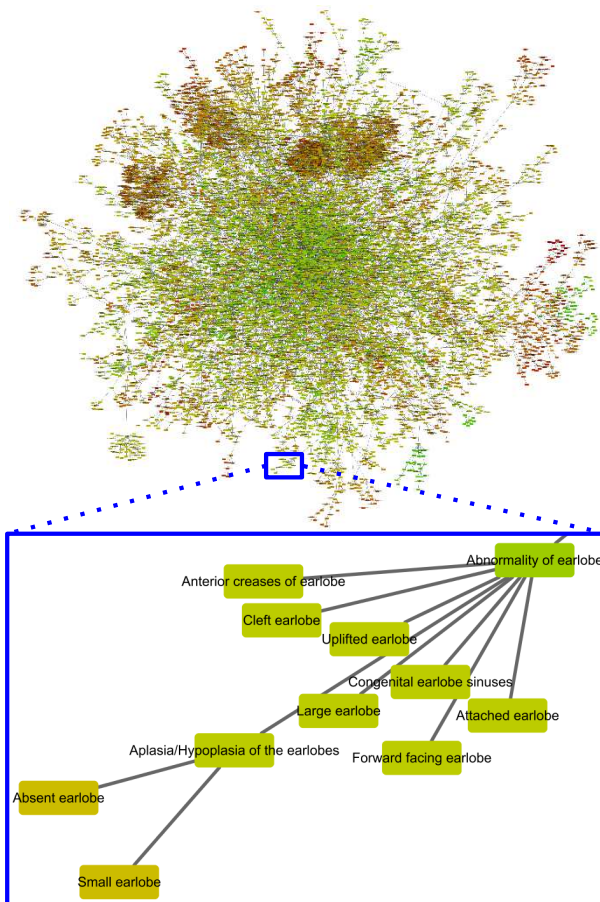


Figure 8.2: Graph of the Human Phenotype Ontology colored by eccentricity, i.e. the distance to root node, from 0 (green) to 13 (red). The blue box shows a zoomed-in view so that labels can be read and the hierarchical structure becomes apparent. This graph has 11,044 vertices and was visualized using CytoScape (<http://www.cytoscape.org>) on the OBO file of HPO downloaded June 2015.

8.1.3 Spreadsheets in the era of big complex data

1 Many researchers, clinicians and other specialists use unsophisticated
2 data storage methods such as spreadsheet documents. While spread-
3 sheets are a useful means for interacting with tabular data, they are
4 not intended to be used as local databases for serious long-term data
5 management. As datasets in spreadsheets grow in size and complexity,
many problems arise regarding data consistency[390], corruption (e.g.
the 'autocorrection' of gene names that look like dates[383]), availabil-
ity, versioning, performance, backups, multi-tenancy and security - nor
do local spreadsheets support FAIR principles[366, 368] that encourage
data to be Findable, Accessible, Interoperable and Reusable.

6 Making relational database technology accessible

7 Relational database technology has developed over the many decades
8 since its invention[62] to specialize in highly structured and consistent
management of high-dimensional tabular data. Frameworks such as
MOLGENIS enable the creation of front-end web interfaces that allow
relational databases to be operated in a more visual and user-friendly
way. Many software applications that were developed with the MOL-
GENIS framework, shown in Figure 8.1, prove that this approach brings
the advantages of relational databases to a variety of life science data
applications that might otherwise remain hidden and vulnerable in local
spreadsheets.

The challenge of big data

Relational databases are not always an appropriate storage solution.
Huge files are currently being produced by automated data processing
tools by high-throughput technologies such as whole-genome DNA se-
quencing. With costs dropping and thousands of samples sequenced
daily, the data is rapidly growing from terabytes to petabytes and be-
yond. It would therefore be highly impractical to store each atomic

value from these data in a relational structure because there is no need for this data to be query-able and a significant amount of additional disk space would be needed to index the data.

However, when thousands of individuals are profiled over many years and their data analyzed by many researchers in different projects, it is inevitable that data is lost track of. Retrieving specific samples in that situation would involve a costly exercise in 'forensic bioinformatics'. Projects that want make use of the data, for example a combined re-analysis of undiagnosed patients with a cardiomyopathy indication, would have to spend a significant amount of time to simply retrieve the right samples.

How to find and access large files

A catalogue system powered by a relational database can be used to store sample metadata and file locations, along with detailed provenance of how the sample was processed and analyzed, and of the results. Using information such as patient phenotype, tissue sampled, sequencing platform and processing software used, data of interest can be quickly found and used for analysis.

A publicly available example of such a big data catalogue is the European Genome-phenome Archive [192], which currently contains everything from raw sequencing files to genotypes called to phenotypes. The data is organized in studies and data sets, and enriched with the provenance of samples and the technology used for analysis, e.g. "Affymetrix 500K" or "Illumina HiSeq 2000". This allows researchers to find data appropriate for their (meta-)analysis amongst the petabytes of deposited files based on biological-, laboratory- and digital provenance.

Hybrid solutions for large data queries

Combining relational and file-based storage can also be an effective solution. The applications developed based on the XGAP data model, such as WormQTL^{HD}, employ a hybrid file-relation storage strategy.

1 The rows and columns link to entries in a relational database, such as
2 Markers and Genes, which can be queried as usual, while the large data
3 in matrix form are stored as a two-dimensionally indexed binary file.
4 Using the query results, data selections from the matrix based on rows
5 and columns can be made very quickly.

6 The result of this hybrid design is great performance with mini-
7 mal overhead and disk space requirements, but its drawback is that
8 sorting and filtering operations on non-indexed matrix values are slow.
9 These queries are however not important for the main use-cases of this
10 database, so the overall solution worked out very well. This shows that
11 no solution is perfect, but the success of a system does depend on the
12 storage strategy chosen.

6 **Basic data management training for all life science researchers**

7 We have shown many different models, methods and tools to manage
8 life science data, from relational databases with static and dynamic
9 models to relational-file hybrids and file-catalogue systems. There is
10 no single “best” solution: each of these approaches represents a valid
11 solution for different storage and query requirements, which just un-
12 derlines the need for flexible systems that can adapt to future data
13 structure needs and switch to storage backends that scale to bigger data
14 volumes when required.

15 However, the average researcher has little interest in the technical
16 background of these solutions and simply wants a system that capable
17 of serious data management that is still as comfortable to use as a
18 spreadsheet program. Making the transition requires an investment of
19 time and energy that is sometimes not well understood and/or seen as
20 too burdensome. Parties that offer data management solutions may
21 have a responsibility to underscore the importance of helping people
22 to use better data management. We suggest number of actions to
23 facilitate the uptake of better data management tools in Box 3.

Box 3: Actions towards uptake of better data management tools

1. Raising awareness of the dangers and limitations of using spreadsheets and other inappropriate solutions for data management. This is also the mission of the European Spreadsheet Risks Interest Group^a.
2. Increasing the visibility of alternatives by shifting the focus of publications, workshops and presentations from specific applications back towards the importance of underlying technologies such as the MOLGENIS platform.
3. Making demonstrations publicly usable in an unrestricted but private way to those who are interested. Subsequently, non-technical users should be able to immediately create secure instances in the cloud suitable for sensitive data. For technical users, it should be simply to run the software on their own servers.
4. To keep initial interest alive, the thresholds of starting to use these applications must be as low as possible. Data management should be user-friendly overall, but it is critical that systems be fault-tolerant so aspiring users are not punished by having to fix many small mistakes when importing their data. Data should also be importable in the simplest of formats and even via direct entry, i.e. similar to using spreadsheet software.
5. Those exploring the system more deeply must experience clear benefits and advantages. For instance, a user-friendly data explorer should offer powerful options to find, filter, sort and plot the data, as well as guarantee consistency, security and easy sharing with colleagues.

^awww.eusprig.org

8.1.4 Future perspectives of sharing life science data

1 The life science and molecular medicine community is gathering, using
2 and sharing tremendous amounts of data. Popular published data re-
3 sources include the Genome of the Netherlands[244], ArrayExpress[298],
4 1000 Genomes Project[18], VariBench[239], Blood eQTL browser[364],
5 database of Genotypes and Phenotypes[216], ClinVar[190], Exome Ag-
6 gregation Consortium[196], genome Aggregation Database³, RNA-seq
7 ASE browser[78], European Nucleotide Archive[195] and Genotype-Tis-
8 sue Expression[213]. Making these data open and freely available gen-
erates a higher number of citations and increases their overall prestige,
which should clearly outweigh any benefits of 'keeping data to yourself'.
Further, an increasing number of (often high-impact) scientific journals
and public funding bodies require any data in publications to be openly
accessible (the ideal) or at least available on request. In fact, it is now
possible to publish data as standalone resources, for example in jour-
nals such as *Scientific Data* or initiatives such as *DataCite*. The field
of genetics is indeed a frontrunner "whose sharing of data markedly
accelerated progress" [297] compared to fields such as clinical trials.

Indeed, we ourselves opened up dozens of research datasets and
results in the WormQTL^{HD} database, free to download without re-
strictions for anyone interested . Conversely, the methods presented in
chapter 5, the CADD scores for MMR genes, and GAVIN in chapter
6, could not have been developed without free and publicly available
datasets.

New query options are needed for humans and computers

While incredible datasets are currently being generated, the develop-
ment of software tools to interact with these data seems to be lagging
behind. Even relatively simple questions such as "How many variants
at exon 10 of the TTN gene are exclusive for individuals of Asian de-

³<http://gnomad.broadinstitute.org/>

scent?" cannot easily be asked because (i) most resources do not have an adequate query interface, (ii) query interfaces do not understand or support the question, and (iii) there is no cross-database "Google-like" interface to query all potential sources at once. We could fulfill the promise of systems biology research to create understanding of life in a bigger context by being able to routinely run deep complex queries across our complete knowledge of organisms, phenotypes, populations, tissues, metabolites, genetics, drugs, and so on.

Ontologies, semantics and FAIR solutions

This situation can be improved by increased usage of ontologies. Data may be expressed as *nanopublications*, i.e. predicated relationships between two ontological concepts with provenance attached. For example, we can express knowledge into formal terms as follows: the CENPJ gene (<http://bio2rdf.org/geneid:55835>) has a statistical association (http://semanticscience.org/resource/SIO_000765.rdf) to Seckel Syndrome (<http://bio2rdf.org/omim:210600>) of 0.00006562⁴. Data can be freely reasoned upon once expressed in a formal and semantic way, as demonstrated in case studies[233]. The underlying RDF[218] (Resource Description Framework, a way to semantically describe data) and SPARQL[1] (Simple Protocol And RDF Query Language, a way to question semantic data) technologies are supported by the Linked Data initiative⁵, which aims to connect structured data on the internet, essentially turning it into a giant database. It has also been shown that semantic queries can indeed be used for both generating and evaluating hypotheses[233, 47], and assist in genomic variant prioritization[33].

Although some large organizations offer proper RDF access[174] and third-party tools are written for others[13], this technology is hardly a cornerstone of modern life science databases. It can be quite labor-

⁴Example taken from: http://nanopub.org/wordpress/?page_id=57.

⁵<http://linkeddata.org>

1 intensive to completely ontologize a database for RDF support and setting
2 up an active SPARQL endpoint. The FAIR principles[366, 368] offer
3 an alternative with much lower barriers. FAIR is essentially a checklist to
4 guide technical solutions and their practical applications to make data
5 more Findable, Accessible, Interoperable and Reusable. When fully realized,
6 every aspect of data would be annotated with semantic metadata for complete
7 and seamless reusability. The MOLGENIS/XGAP work presented in this thesis
8 are in that sense FAIR systems, created before that term was coined.

Given the challenges of building and using datamodel based storage systems this is not always attainable. However, the guidelines can be followed to make data discoverable and searchable in simpler ways, which is always better than not at all. An example of an easy and lightweight solution to make resources better findable by search engines is to add Bioschemas⁶ semantic markup. Other databases, platforms and initiatives such as Dataverse[70], FAIRDOME[373], ISA[301] and Open PHACTS[144] showcase other implementations and applications that are based on FAIR principles.

Federated queries require an attitude shift

In addition to new data models and tools, I think we also need to change our attitude towards sharing and integration of query results. Data sizes are rapidly increasing and much data is privacy sensitive, factors that make it unlikely that all data relevant to a given study will always be available in one place. New data analysis paradigms are being developed to address this, of which so-called 'federated analysis', is a prominent example[118, 169]. A federated query is executed on multiple independent locations, after which the partial results are combined into the final answer. The nature of federated queries on multiple external sources is that the answer of today may be different from the answer of tomorrow. This can be seen as a limitation, but can also be considered

⁶<http://bioschemas.org>

an opportunity where researchers and even clinicians always have access to the latest and greatest knowledge. Instead of statically versioning complete data sources as a reference point, we must consider making our data dynamic and compensate for perceived uncertainties by storing the questions together with detailed data and source provenance used to compile the answer.

Legal aspects of data sharing

Until here I have discussed life science data sharing from a mostly technical point of view. It is however important to realize that even the most brilliant solution is pointless if the law precludes sharing and reuse of data.

Currently, data and privacy laws are being revised in light of social media and the other 'tech giants' who are gathering massive amounts of sensitive user data. While citizens will benefit from better privacy protection, the same rules impede cohort-based biomedical research by requiring participant re-consent every time the data is used[2].

While scientific representatives are gaining concessions on behalf of research, their legal struggle is not yet over. Establishing fitting international legislation with many stakeholders across cultural barriers is difficult, as is the practical implementation of new rules by the science community[209].

Perhaps by overcoming these legal growing pains, current hindrances for sharing and reusing biomedical data across the international community can be lifted, benefitting researchers and ultimately patients.

8.2 Developing computational methods for medical genetics

So far I have investigated models to store, access, share and query data measured and calculated in life science experiments to gain new insights.

In this section, I discuss challenges and solutions in the development and, more importantly, validation of new methods and algorithms to make most use of all these new data.

Key question and points of this section

How can we develop, characterize, find, share and reuse high-quality computational methods as part of new analysis protocols for medical genetics?

Key points

- The performance of newly developed methods depends on both the quality and quantity of available data. We must therefore cherish and share gold standard data (8.2.1).
- Reporting the strengths and weaknesses of a method in greater detail next to the overall performance is crucial for choosing the right analysis method to obtain better research or diagnostic results (8.2.2).
- We need detailed catalogs where researchers and clinicians can seamlessly find these assessments (8.2.3) and run the right tool for their data and hypothesis (8.2.4).

8.2.1 Method dependence on high quality data

The genome of an average person contains around 5,000 unique mutations[355]. To establish a molecular diagnosis in an individual with a suspected genetic problem, we need to quickly reduce the number of candidate variants from thousands to just a few. Computer algorithms can predict how harmful mutations are, but the strength of these predictions depends on the availability and quality of reference data. These

gold standard reference data are used to develop new algorithms, which are then validated on a similar but independent data set.

In this section we show that both quality and quantity of data have an effect on the predictive power of new methods. Sharing data is therefore crucial for developing the best possible tools, resulting in more accurate prioritization and less time spent on manual assessment. We therefore must engage in active international collaboration to set up systems for joint interpretation and open sharing of variants. To accomplish this goal we need to overcome technical and legal barriers, but it is also a social challenge, as labs can be distrustful of other labs that operate under different guidelines or less stringent quality standards.

CADD and GAVIN

CADD scores[185] are a promising method to estimate pathogenicity of any mutation in the genome. We calibrated CADD in silico pathogenicity estimates to help classification of mismatch repair gene variants in chapter 5. We used variants assessed by an international committee of domain experts to establish the relationship between CADD scores and classification outcome. This was very successful, and this success was helped by the outstanding quality of the reference data set we used. Re-application of this model to the original data showed only a few discrepancies, and they could all be explained in favor of the original human expert classification.

This work was followed up in chapter 6 where we present the GAVIN variant classification tool for >3,000 clinical genes. Interestingly, we found that CADD-based predictions work better for some genes than for others. This may be explained by currently unknown biological differences that are not captured in the in silico pathogenicity estimates we used, however a simpler explanation is that the expert variant classifications were of lower quality for certain genes, and those mistakes distorted the calibration. We have already shown that for genes with

1 enough training data we get better classification accuracy than we do
2 for genes for which scarce data is available, but here we can investigate
3 the relationship of both quantity and quality of gold-standard data with
4 calibration success in more detail.

5 **Association of p-value with variant quantity and quality**

6 The GAVIN gene calibrations include a Mann-Whitney U test p-value for
7 the tested significance of the CADD score difference between pathogenic
8 and matching benign variants. The lower a p-value for a gene, the more
reliable and useful the gene calibration becomes for automated variant
classification. These gene calibration p-values can be plotted against
the number of expert-assessed pathogenic variants available per gene.
For this we used the ClinVar variant_summary.txt file of 1/12/2016,
downloaded from <ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar>
(ar). The result can be seen in Figure 8.3. Indeed, a simple log-linear
model shows a trend in which gene p-values become more significant
as the quantity of available variant classifications increases.

To measure variant interpretation quality, we can use the review
status of ClinVar variants. The review status shows the amount of
supporting evidence for the clinical significance (i.e. classification) of
a variant. The text values also correspond to a 'star' rating from 0
to 4 (see Table 8.1 on how the terms are mapped), which can be
used quantitatively. We can take the mean of this rating per gene as
a measure for interpretation quality. When plotted against the gene
calibration p-values (Figure 8.4), there is a trend in which gene p-
values become more significant as the quality of variant classifications
increases.

Both trends seem to indicate that both quality and quantity of data
are important when developing new methods based on previous ob-
servations to help us interpret unprecedented amounts of new data.
Therefore, we need to treasure the results of expert interpretation and
analysis that has been made freely available, but at the same time put

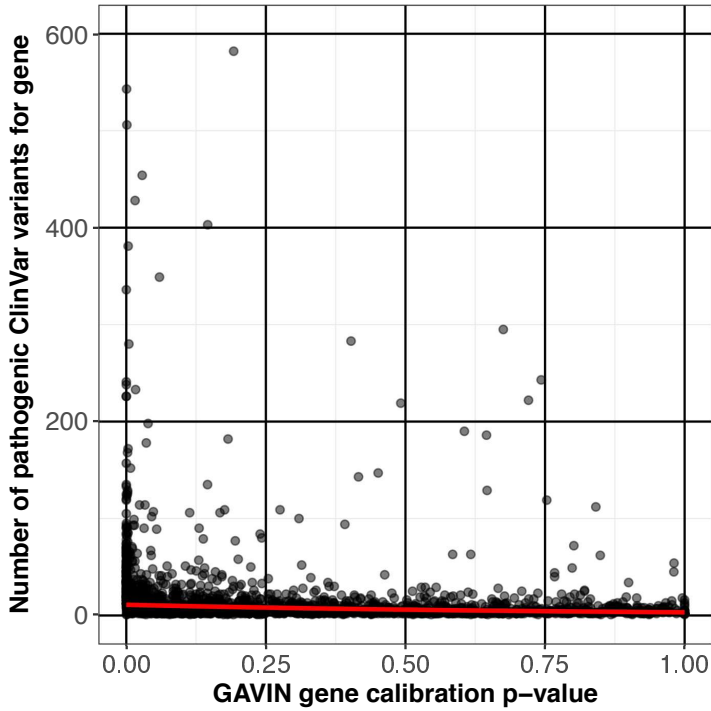


Figure 8.3: Relation between calibration success and the number of pathogenic variants available for that gene. Log_{10} linear regression (shown in red) resulted in $R^2 = 0.11$ and $p\text{-value} = 7.28\text{e-}57$.

ClinVar review status	'Star' rating
No assertion provided	0
No assertion criteria provided	0
No assertion for the individual variant	0
Criteria provided, single submitter	1
Criteria provided, conflicting interpretations	1
Criteria provided, multiple submitters, no conflicts	2
Reviewed by expert panel	3
Practice guideline	4

Table 8.1: ClinVar review status and how this translates to a numeric range, i.e. the corresponding review status 'star' rating. See https://www.ncbi.nlm.nih.gov/clinvar/docs/variation_report

more effort into integration and 'FAIR-ification' of the many sources to achieve the best and most complete reference set possible[40] for human health and disease. This need will become more pressing as the amount of data grows quickly, and the upcoming demand for powerful methods that can deal with new diagnostic data modalities such as non-coding DNA, RNA-sequencing, metabolomics and epigenomics.

8.2.2 Benchmarking and characterization of methods

While overall performance is a good metric to judge a method's ability to screen a large set non-specific data, for specific data we should characterize a method in more detail. In order for methods to become trusted and accepted by users such as clinical geneticists, they must know how well a method behaves and if it succeeds in situations they are familiar with.

A good example is our analysis of CADD scores for MMR genes in chapter 5, where we delved into the strengths and limitations of this method when applied to four specific genes. Developing a new method as a black box, with just an overall reported accuracy (e.g. "93% AUC")

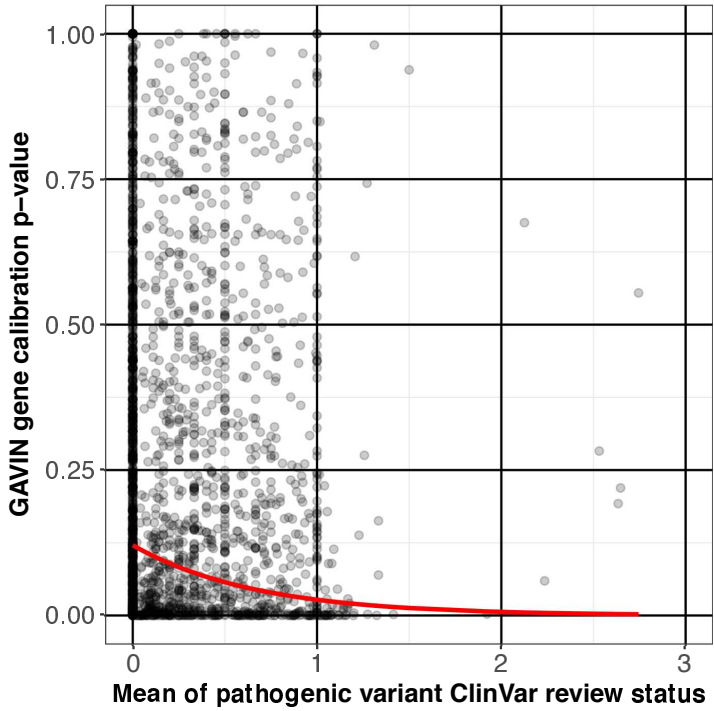


Figure 8.4: Relation between calibration success and the average review quality for that gene. Log₁₀ linear regression (shown in red) resulted in $R^2 = 0.03$ and p -value = $5.95e-16$.

1
2
3
4
5
6
7
8

1 may lead to some skepticism from those using the methods in practice,
2 especially since most methods claim to be the best. Instead we should
3 characterize and report the strengths and limitations of a method, and
4 communicate clearly that method performance may depend on the con-
5 text it is used in.

6 The GAVIN gene calibrations in chapter 6 are reported in categories
7 where 'C1' indicates a high degree of separation between pathogenic and
8 benign variants and 'C4' indicates a poor separation. These indications
show that classifications in certain genes are better than in others, even
though the overall performance is high. Reporting the performance on
a gene-level helps researchers or clinicians to select the best method for
their specific question.

It must be noted that GAVIN reports no formal information beyond
the gene level, and that there are indeed instances where more sophis-
ticated definitions are beneficial. We will now show three examples of
genes where in-depth characterization is indeed relevant and how this
may lead to tailored or optimized usage.

Examples of in-depth gene characterization

The gene *SCN5A*, which encodes sodium channel type V and is asso-
ciated to dominant atrial fibrillation/long QT syndrome, is an example
where there is a high degree of separation and an overall great calibra-
tion. However, notice the cluster of pathogenic variants around location
38655000 that dips below the calibration line in Figure 8.5. This gene-
specific model may be improved by local correction of the threshold for
this effect.

In the gene *TTN*, which encodes the protein titin and is associated
to dominant cardiomyopathy, this effect is far more pronounced with a
massive cluster of extremely high score pathogenic variants in the first
30% of the gene. See Figure 8.6. The remaining pathogenic variants
appear to be distributed amongst the benign variants, making a poten-
tial two-part model perhaps complicated, but much more powerful.

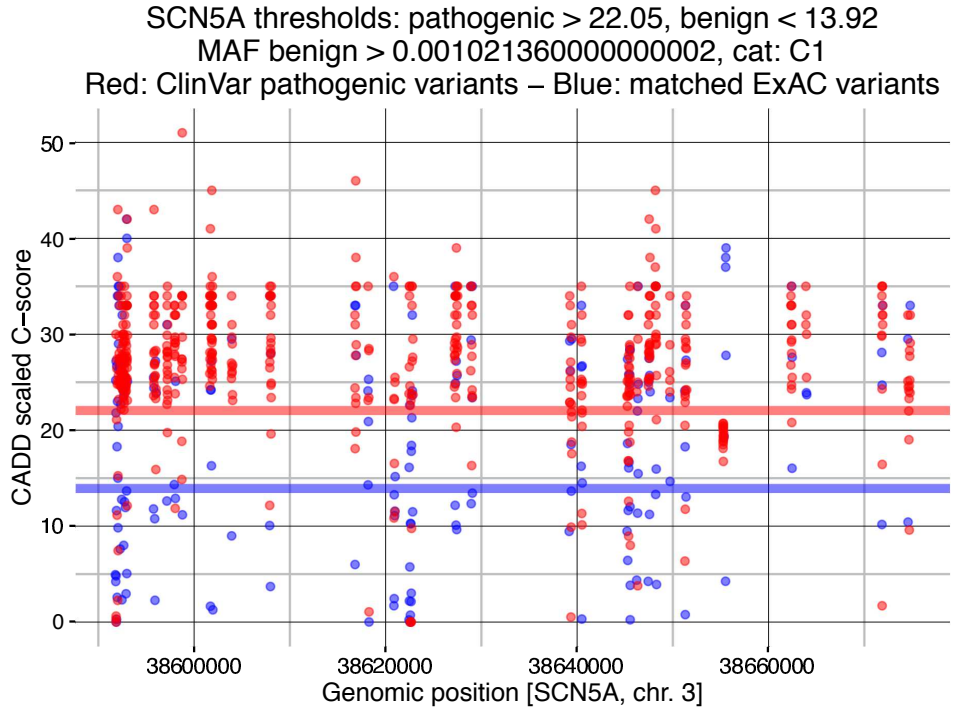


Figure 8.5: GAVIN (r0.3) calibration plot for SCN5A.

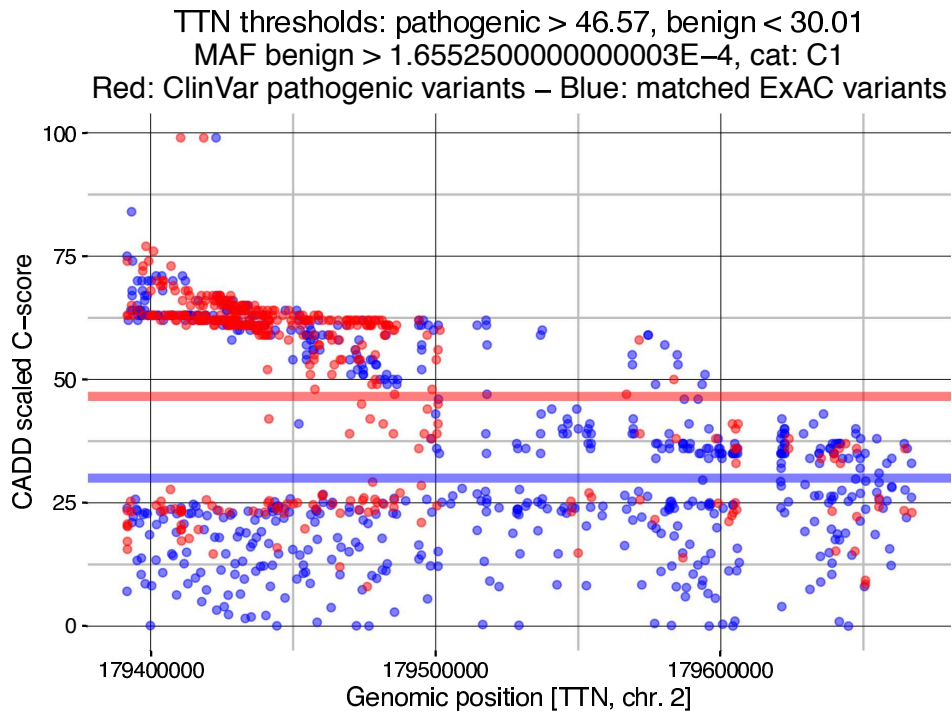


Figure 8.6: GAVIN (r0.3) calibration plot for TTN.

An example where the method offers little predictive value is the gene F11, which is associated to blood coagulation factor XI deficiency and is often recessive. The benign variants that remain after the filtering stage of GAVIN calibration are distributed uniformly across the pathogenic variants as can be seen in Figure 8.7. A simple explanation of this might be that this is a relatively mild and recessive disorder under apparently little selective pressure, which is consistent with it being relatively common among Ashkenazi Jews[305]. What this example shows is that some pathogenic variants seem to be quite tolerated in the general population, blurring the line between 'benign' and 'pathogenic', therefore deleteriousness may be hard to estimate computationally.

Implications of method characterization

These examples illustrate the advantages and limitations of a gene-based variant classification method. This characterization will also apply to other data modalities including gene expression and metabolites and across different conditions such as tissue type, cell types, age, and ethnicity. If we want to develop methods that bring these new types of information to the clinic, we must have transparent and widely accepted validation procedures and be clear on what methods can and cannot do, to prevent disappointed users. The context for which the tool has been developed, as well as its strengths and weaknesses, should be made clear even on a gene-specific level because this knowledge may be more important than the overall performance of the method. I think it may be worthwhile to create a truly standardized benchmark to assess and compare the methods currently available as well as those that will be developed in the future.

8.2.3 Finding appropriate methods in repositories

The sequencing of thousands of patients and healthy individuals worldwide, combined with the sequenced genomes from thousands of different organisms, has spurred the development of countless methods that

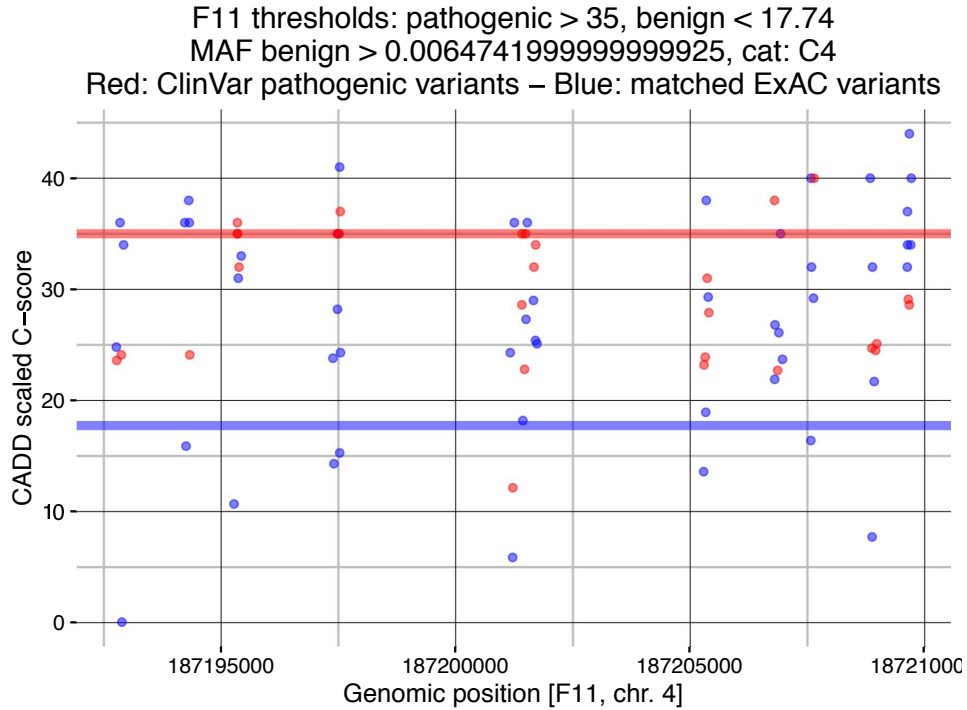


Figure 8.7: GAVIN (r0.3) calibration plot for F11.

predict variant pathogenicity, mostly based on estimated protein conservation. Examples of various scope and quality include: SIFT[187], PolyPhen2[5], PROVEAN[59], PON-P2[241], MutationAssessor[287], FATHMM-MKL[308], Condel[129], PhyloP[271], UMD-Predictor[109], Grantham[131], ENTPRISE[386], PFAST[388], FitCons[136], MutPred[199], EIGEN[160], GERP++[76], VAAST[181], AlignGVGD[333], MAPP[334], MutationTaster[304], VIPUR[22], REVEL[159], CADD[185], LINSIGHT[155], FATHMM-XF[296] and GAVIN[345]. These tools are being improved and invented in a fast competitive cycle as more benchmark data becomes available every day due to the interpretation help of the previous tool generation.

Similarly, thousands of methods for all kinds of applications, analyses and data types have been created across all branches of life sciences. However, for a researcher at the beginning of a project, the question remains: How do I find the best methods for my analysis question?

Search engines to find methods such as OmicsTools⁷ have emerged to let users find methods of interest. In OmicsTools, users can browse methods via search box or by drilling down in categories. Users can also post reviews and rating to let others know how well they liked the tool. OmicsTools does not, however, allow more fine-grained searches that can:

- Find tools that are applicable to your data, e.g. quickly finding which analyses can be run on your data.
- Find tools that produce a specific type of output, e.g. if you are interested in a specific type of output such as gene annotations and want to list any tools that can provide this.
- Find tools that perform a specific role, e.g. when you want to benchmark your method to any tool that performs the same function regardless of input or output.

⁷<https://omictools.com>, currently hosting >17,000 entries

1 The Elixir Tools and Data Services Registry[163]⁸ tries to solve this
2 issue by attaching EDAM[162] ontology terms to the function, topic,
3 inputs and outputs of each method. Tags, documentation, publication,
4 links and contact information are also provided, together forming a com-
5 prehensive and highly structured collection of bioinformatics software.

6 These solutions are a step in the right direction, but they do not
7 address two major needs: (i) providing a standardized and detailed
8 benchmark of tool performance (as discussed earlier) and (ii) helping
the user to install and run the tool of interest. Therefore, there is still
a duty for the bioinformatics community to create a central repository
of documented, tested and runnable bioinformatics tools, in the same
spirit as e.g. the CRAN repository for R packages.

8.2.4 Integrating and running methods for evaluation

7 Evaluating new tools in practice requires a quick process of installation
8 and running. This is difficult for methods that can not be offered as web
services due to transfer limitations or patient confidentiality, and must
therefore be compiled or installed locally and subsequently integrated
into an analysis protocol. BioContainers[72] solve some of these issues
by wrapping individual tools in container engines (Docker and rkt). The
tools retain their identity while being easier to run cross-platform.

The fast uptake of methods is much easier and quicker when they
are wrapped or created for an existing workflow engine. MOLGENIS
Compute protocols⁹ and Galaxy tools¹⁰ are method libraries for their
respective workflow engines that come with descriptions and technical
definitions for their inputs and outputs. Taverna[372] workflows can be
constructed and shared on MyExperiment[127]¹¹ using components¹²

⁸<https://bio.tools>

⁹https://github.com/molgenis/NGS_DNA/tree/master/protocols

¹⁰<https://toolshed.g2.bx.psu.edu/>

¹¹<https://www.myexperiment.org>

¹²<http://www.taverna.org.uk/documentation/taverna-2-x/components/>

8.2. TOWARDS BETTER SYSTEMS FOR (GEN)OMIC MEDICINE

that can be linked to ontological terms for their input, output and activity.

However, there does not seem to be much emphasis on semantic description of workflow engine components in general. Taverna components can only be published within the definitions of conventional workflows, making them difficult to find. Methods wrapped for Compute and Galaxy do not explicitly link to ontologies at all.

These limitations were recognized and addressed by the BioMOBY method and its successor SADI[367, 370]. SADI is a method to set up discoverable semantic web services from regular databases, and has a Taverna plugin to access these services. Unfortunately this project has been silent since 2014 and the plugin is only available for the outdated Taverna versions 2.1.2 and 2.2 (current version is 2.5).

Taken together, we find many great initiatives that collect, describe and offer methods focused on different aspects. Unfortunately there is currently no standardized complete solution that allows a user to seamlessly discover, run and integrate methods into their analysis workflows.

There is a growing focus on the FAIR principles to make data reusable, but the exact same principles should also apply to tools. I think we should treat tools as 'runnable data', meaning that they must be as easy to find, understand and use as data itself. In practice, the most popular tools are simply those that are easily runnable, and these are not necessarily the best at what they do. By applying FAIR principles to remove some of the barriers to use, we could exchange and adopt more appropriate tools to the tasks at hand.

8.3 Towards better systems for (gen)omic medicine

Thus far I have discussed models to store, manage, share and query life science data in smarter and better ways. I then looked at the challenges of developing, characterizing and discovering new methods. In this

section I focus on challenges and solutions in translating new data and methods to health research and patient care. Note that the terminology used in this section can be confusing and is therefore explained in Box 4.

Box 4: Clarification of terminology

In a typical data processing scenario, multiple tools or methods are connected in a workflow, also called a pipeline, which is a sequence of events through which data is processed to reach a final state. A workflow that has been formalized into an official procedure that is agreed upon and usually versioned is what we call a protocol. The steps within a protocol can be implemented using different tools for each step, where tools are implementations of methods. Lastly, a system is a piece of software that executes workflows or protocols and manages their inputs, tools, outputs, provenance and other related data. Some of these terms are used interchangeably when context allows it, for instance, workflow and protocol are in often in practice not that different.

Key question and points of this section

How can we bring data and methods together in flexible and scalable multi-omics analysis protocols and software systems for future patient care?

Key points

- There is a plethora of academic and commercial software for DNA analysis, but their protocols and data sources are quite static (8.3.1).
- Workflow engines offer greater flexibility for data processing and are far more future-proof, but the use of a common language must be encouraged (8.3.1).
- Protocol implementations of best practice guidelines should be a community effort including common automated benchmarking and validation (8.3.2).
- As multi-omics analysis starts to complement routine DNA diagnostics, experts from the disciplines involved will need to contribute their best practices in a combined approach (8.3.3).
- To keep multi-omics diagnostic protocols up-to-date, we should consider using smart workflows with abstract step definitions that automatically select the best or most appropriate tools for the job (8.3.4).

1
2
3
4
5
6
7
8

8.3.1 Reusable and flexible DNA analysis workflows

Since DNA sequencing has become popular, plenty of integrated systems have been developed that cover complete genome analysis workflows. Commercial examples include products such as Alamut¹³, SeqPi-

¹³<http://www.interactive-biosoftware.com>

lot¹⁴, Omicia¹⁵, Sophia¹⁶, NextBio¹⁷, Cartagenia¹⁸, MedGenome¹⁹, VariantStudio²⁰, GoldenHelix²¹, Ingenuity²², Bina²³, Enlis²⁴ and Genomatix²⁵. Alternatives developed in academia, often free to use, include SpeedSeq[58], GEMINI[253], InterVar[202], Genomiser[313], eXtasy[309], VAAST[181], SG-ADVISER[262], IMPACT, SeqMule[137], TAPERTM [126], ClinLabGeneticist[357], WGSa[211] and wANNOVAR[53]. These systems are quite specific for the types of data and questions they can handle. Either they offer a built-in analysis or they allow the user to select which of the preconfigured data and filters should be used. While these products may perform their function well, the lack of freedom can be a serious restriction. The GAVIN+ tool presented in chapter 7 is guilty of the same, although it is part of a bigger genome interpretation framework with options for customization and tool replacement.

With many new sources of data, knowledge and tools are quickly becoming available, typical genomics analysis software cannot keep up with the demand for the latest and greatest developments, let alone support integrating completely new data modalities such as RNA-seq, metabolomics and epigenetics. Switching to different software or adopting multiple tools to use the best features of each is time consuming and expensive, while using an incomplete solution means missing out on the best research results or diagnostic answers.

¹⁴<http://www.jsi-medisys.de>

¹⁵<https://www.omicia.com>

¹⁶<http://www.sophiagenetics.com>

¹⁷<https://www.nextbio.com>

¹⁸<http://www.agilent.com>

¹⁹<https://www.medgenome.com>

²⁰<http://variantstudio.software.illumina.com>

²¹<http://goldenhelix.com/>

²²<https://www.qiagenbioinformatics.com>

²³<http://www.bina.com>

²⁴<https://www.enlis.com>

²⁵<https://www.genomatix.de>

Workflow engines, a better way?

Workflow engines offer part of the solution to these issues. They are a type of software not tied to specific data types or analysis protocols, but instead offering the flexibility of letting users define and share their own analyses. Examples include Galaxy[128], Taverna[372], Anduril[249], UGENE[245], GenePattern[284], VisTrails[23], Arvados²⁶, AWE, Toil, Rabix and MOLGENIS Compute[43]. In these software the user has full control and freedom over analysis steps such as choosing which tool and data dependencies are used. A user can, for instance, choose GATK[344] for genotype calling then subsequently choose PLINK[277] for association analysis. The flexibility originates from using simple tool input/output structures that are typically file based. This agnostic approach to handling data makes it easy to hook up new tools and file formats. Due of their adaptability these solutions are more future-proof than software shipped with built-in analyses. However, the graphical user interfaces of workflow engines, if present at all, are not optimized towards a domain-specific task, and this may deter users who are used to hand-designed interfaces. These generic workflow engines are often quite technical to use and have therefore not caught on in mainstream molecular diagnostics.

Another issue is that workflows are usually created for a specific workflow engine, making cross-platform reuse harder or impossible. This leads to reimplementing of workflows for different engines, which costs time that could have been spent improving the already existing workflow. Some relief is on the horizon in efforts such as the Common Workflow Language[10] (CWL) that encourage the use of a cross-engine language to define workflows, and CWL is currently supported by 9 different workflow engines. I believe that uptake of one standard by the community would enable sharing and collaborative improvements of workflows. In addition, easier and more customizable user interfaces could make them more popular with non-technical users in clinical ap-

²⁶<https://arvados.org>

1
2
3
4
5
6
7
8

1 plication. Whether the standard exchange format should become CWL
2 or something else is up for debate, but any standard here would surely
3 further boost this field to innovate and collaborate around it, which
4 is what we have seen happen with the VCF format that has been an
5 incredible catalyst for variant data exchange and common usage across
6 tools.

3 **8.3.2 Community sharing of protocols and expertise**

4 Medical centers that perform molecular diagnostics use national or
5 international guidelines to implement their protocols for the interpreta-
6 tion of DNA variants. Examples include guidelines established by clini-
7 cal genetics associations in the Netherlands[242], United Kingdom[353]
8 and United States[288]. The guidelines are implemented by configu-
ring existing software, for instance by running an automated MAF filter
followed by manual interpretation that may assign 'likely pathogenic'
status to a private stopgain variant with PolyPhen verdict 'damaging'.
While these guidelines are established and agreed upon by experts, their
implementation and validation is typically not shared amongst centers.
This is a pity because sharing of expertise with an international commu-
nity would reduce redundant work and increase quality of results. We
should therefore unlock the knowledge that is now kept in the protocols
of local configurations or precompiled software.

Luckily, many of these protocols consist of objective interpretation
criteria that can be turned into automated workflows, and these are
easier to share. The recently published InterVar[202] tool, for example,
has implemented the ACMG 2015 guideline in a Python script. Au-
tomated analysis workflows created by experts can also be shared via
initiatives that support communities such as MyExperiment[127]. By
using a common language to describe these workflows we can try work
towards an interactive online catalog of best practice workflows main-
tained by the international genetics community. In open collaborative
development these protocols can be updated, amended, expanded or

8.3. TOWARDS BETTER SYSTEMS FOR (GEN)OMIC MEDICINE

merged as our insight and resources increase, leading to higher quality guidelines and best practices. This will increase sharing of specific expertise and lessons learned, in addition to preventing duplicate implementation and validation efforts.

The genomics platform in chapter 7 proposes implementation of an automated interpretation protocol based on a template connected by existing and sharable formats that allows modularity and reuse of individual components. It features built-in methods and gold standard data for automated validation, and benchmarking that can be applied in any new implementation of the protocol to verify that the performance is still the same, or has changed. The automated re-validation of interpretation workflows also drives innovations as added value of enhancements can be objectively proven and mistakes avoided with very little effort, and I think this is a subject where much can be gained that will allow genome diagnostics to scale up for future needs. The concept of sharing tools, templates and validation strategies gives centers many options to exchange best practices and benchmarking tools that can be reused fully or in part.

8.3.3 Towards integrated multi-omics analyses

Life science research and healthcare is starting to complement the sequencing of genes with new data modalities such as non-coding DNA and RNA expression. A number of efforts have shown great potential for moving past 'DNA only' analyses, as demonstrated by studies that integrate multiple omics layers to better understand human health[272]. Following these experiments, new computational methods such as REMM scores[313] for estimating variant pathogenicity in non-coding DNA are being developed to leverage wet laboratory advances into new high-throughput tools. An brief overview of the current state and diagnostic potential use of various omics data types can be found in Table 8.2.

Using all these omics data modalities in an integral way requires

flexible databases and tool systems. But, more importantly, it also requires best-practice data processing protocols to be shared by experts. An inspiring example of multi-omics bioinformatics that turned fundamental research into clinical practice is the discovery of the PCSK9 gene. In this case a combination of identification of protective alleles, classical family studies, discovery of cellular pathways using model organisms, metabolic measurements and gene sequencing led to discovery of a drug target[3] that is now successfully used in clinics[94].

Studies that are focused on a particular molecular mechanism can afford to manually integrate relevant data and run additional experiments to complete the puzzle[261]. For routine use of multi-omics, such as in diagnostics offered to thousands of patients, we need a more systematic and automated approach.

While better databases for multi-omics data are emerging, our omics data integration methods lag behind. The usual strategy in data integration for DNA analysis is using software that glues data together, where all the complex logic to the join data resides within the software and the data itself is agnostic. For example, we have implemented 'annotators', which enrich genomic variants with information such as allele frequencies, pathogenicity scores and transcript annotations[345]. There are many such annotation tools[100], integration platforms[257] and precompiled annotation sources[210], all of these can have notable differences[223] but all generally work reasonably well for this purpose.

Moving the 'smartness' from tools to the data

The problem is that for every new application of the data, or new data source that needs to be integrated, the community needs to develop and maintain new software. This means that the programming logic for integrating specific sources is re-implemented many times, with problems and limitations in each separate implementation. This will become a serious issue when moving from DNA alone to 10+ omics data layers in the future. Developing the ultimate tool that can handle all omics

8.3. TOWARDS BETTER SYSTEMS FOR (GEN)OMIC MEDICINE

Molecular assessment	Application as diagnostic technique	
Coding DNA sequencing	Commonly used technique with a yield that varies from 30% for difficult cases[348] to 60% as a first line tool[232].	1
Non-coding DNA sequencing	Methods are currently emerging[313, 138] based on known non-coding pathogenic variants located in promoters, enhancers, 5'UTR, 3'UTR, RNA genes and topological domains[81].	2 3
RNA sequencing	Powerful complement to DNA sequencing that can to detect aberrant expression levels, gene fusion, allele specific expression, expressed lncRNA and viral DNA[78, 44, 318, 186, 71].	4 5
Epigenetics profiling	Methylation profiling is now limited to cancer but may soon be applied to neurological and autoimmune disorders[151]. Histone modification is assessed in colorectal cancer[116].	6 7
Microbiome metagenomic sequencing	May soon be used in diagnosis and treatment of bowel related disorders[279] in combination with host genome[32].	8
Metabolomics	A reliable technique[149] to screen for 900 compounds[231] but is currently limited to 80 inborn errors of metabolism and other metabolic disorders[130].	
Proteomics	Though mass spectrometry provides high-throughput potential[143] and there are proteomic cancer biomarkers[113], uncertainties and difficulties need to be resolved for diagnostic use[135].	

Table 8.2: Brief review of different omics data types and how they are currently used, or could be used, for diagnostic applications.

data perfectly is a pipe dream, but there is an alternative.

I believe we should make the data itself 'smarter', i.e. more self-aware of what it is, through semantic (meta)data enrichment - while the software can be made 'dumber', only knowing how to process standardized definitions and rules, unaware of arbitrary details of the underlying data sources. By keeping the meaning of the data contained within itself, it would be much simpler to connect datasets to each other without the need for complicated software. Generic software to facilitate data integration is better suited for community-driven development, shared usage, higher code quality and future updates. It also saves the time that would otherwise be spent building and maintaining multiple implementations that do the same work.

8.3.4 Future work on semantic analysis systems

So far I have discussed ways to standardize workflows, to share protocols and validation tools, and to use semantics to better integrate data within complex analyses. The challenge of data integration also applies to organizing tools, protocols and systems created thereof. It therefore makes sense to extend the use of semantics to these factors as well. We propose a protocol template in chapter 7 but, while it has been implemented, the template itself is not formally defined. If we would do this, a logical choice would be to express the protocol as a series of semantic concepts such as "annotating", "reporting", "validation", and so on. This would allow these components to be swapped out for tools with the same definition, assuming there are no further compatibility issues. In effect, we would have parameterized the choice of method for each step and could offer the user a selection of matching tools that can fulfill that step. A validation procedure for that step can be implemented, automated and shared, enabling fast and objective comparisons between variations.

Similarly, the complete protocols may be semantically annotated to perform a role that complements another protocol. For example, a




8.3. TOWARDS BETTER SYSTEMS FOR (GEN)OMIC MEDICINE

"variant calling" protocol may be seamlessly connected to a "diagnostic interpretation" protocol. Other protocols that use "variant calling" as input may be downloaded and executed on the fly for maximum flexibility. How all these components could now interact is illustrated in Figure 8.8.

Disconnecting tools and data from the protocols allows for designing and sharing of protocols without simultaneously dealing with implementation details, and prevents new protocols from having to be defined when a better tool that performs the same role is introduced. When we extend this concept to tools that retrieve their own source data via semantic connection, the result is a moldable, future-proof workflow that ships with built-in validation and benchmarking.

The results created by tools or workflows may be automatically annotated with the appropriate semantics and can be fed back into knowledge repositories. For example, after running a chain of tools that leads to variant classifications, an expert should be able to easily upload these results and share them with the community. Such an approach helps us get the most value out of human experts, as their knowledge and decisions flow back into the system, catalyzing development of new and improved tools.

An immediate challenge to implement this concept is finding a place to keep the necessary tools, data and their metadata. There are central repositories to help make patient mutations or sequencing reads meet FAIR principles, but there is no such place for the results of computational analyses. While the model organism data in chapter 4 can be found through the major community hub <http://www.wormbase.org>, the data in chapter 6 is not so FAIR. Both the article and data are free and open to the world, but the existence of the data set is not explicitly advertised or registered on a community hub, limiting its Findability. Bioinformaticians and computational biologists must put more effort into placing their work in context with the original data, and we need take responsibility as a community to create new means of enabling this integration if there are no suitable solutions available.

-  *Basic scientist: query hypotheses, find tools, protocols*
-  *Translational tool developer: discover latest resources*
-  *Clinical geneticist: discover and run clinical protocols*

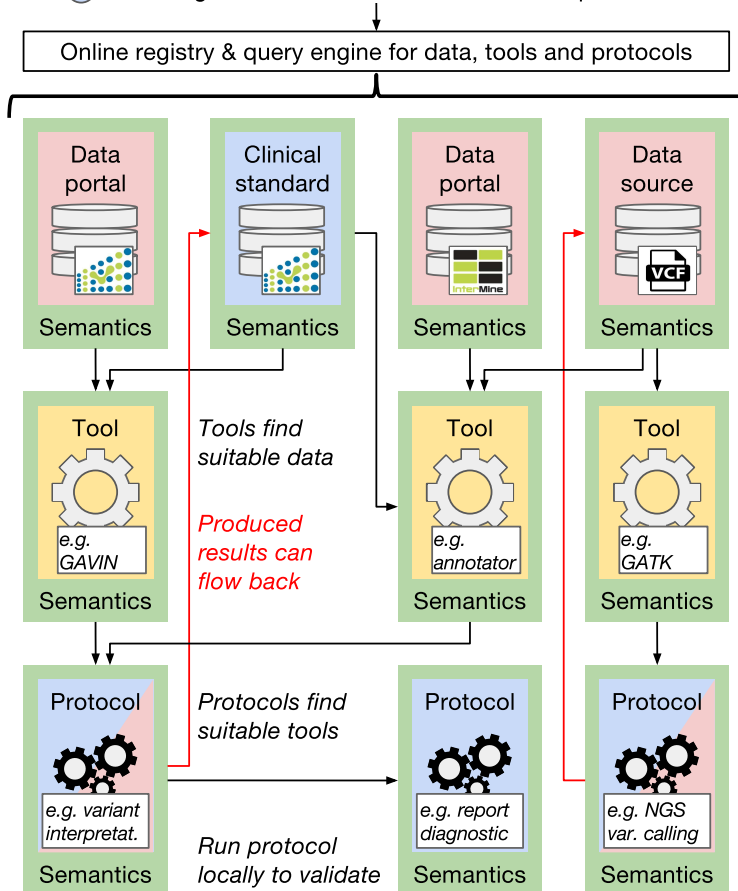


Figure 8.8: Overview of potential interplay between data, tools and protocols, all mediated by semantic definitions. The resources are self-descriptive and can be used in different scenarios by multiple disciplines in a synergistic knowledge feedback loop.

8.4 Conclusion

Each topic in this thesis is part of the same mission: to let patients with genetic disorders benefit more from rapidly increasing biological data production and new knowledge. More specifically, we want to obtain predictors and biomarkers from research and patient data that can quickly establish the best and earliest diagnosis or prognosis. To achieve this goal, we integrated and make available big reference data in chapters 2 and 3, bridged model organism to human data in chapter 4, translated generic methods into clinical applications in chapters 5 and 6, and developed a platform to bring innovations into practice in chapter 7.

The resources currently available are already plentiful, and both the amount and types of molecular life science data is growing at a tremendous pace. This present us with incredible opportunities to develop new and exciting methods that are more powerful and better tailored to patients than ever before, but simultaneously introduces the huge challenge of integrating and understanding these data. To keep up, we must work smarter by investing in development and implementation of techniques that bring data together and stimulate collaboration such as FAIR principles, sharing platforms and semantic web. Our research, development and hands-on experience of MOLGENIS flexible databases with ontology annotation tools are ready to play a significant role here.

New molecular data modalities such as non-coding DNA, RNA expression and metabolic profiles are emerging, and corresponding tools for clinical application will continuously improve as the quantity and quality of gold standard data increases over time. To keep complex multi-omics data pipelines manageable in practice, we need modular best practice workflows that have self-managing and self-validating properties. With many methods available, and their strengths and limitations characterized, the most appropriate tools could be automatically selected to diagnose a patient with an individualized multi-biomarker approach. The genomics platform that we are developing in combina-

1
2
3
4
5
6
7
8

tion with the MOLGENIS Compute engine is being used in a diagnostic setting already and should provide the flexibility to scale up and expand to future technologies.

1 In the future, I envision a seamless international collaboration of
2 experts in an online community-based decision support system where
3 research data, gold standards, tools, workflows, benchmarks and best
4 practices may be shared freely and openly. This will increase the ef-
5 fectiveness of clinical molecular diagnostics at a maximum speed and
6 unlock the potential of all measurable omics data types. We can fur-
7 ther integrate these results with findings that may currently be difficult
8 to interpret such as risk factors from genome wide association studies,
effects in quantitative trait loci or allele-specific expression, changes in
the microbiome, epigenetic marks, or multigenic inheritance. Automat-
ically generated reports will then present prioritized findings and other
relevant insights along with any known limitations and uncertainties, so
researchers and doctors have clear and honest understanding of the re-
sults. Taken together, we will be able to translate the knowledge gained
from research data, expert communities and computational methods
into medical practice for fast, accurate and personalized patient care.

Bibliography

- [1] SPARQL query language for RDF. Technical report, World Wide Web Consortium, January 2008.
- [2] Data overprotection. *Nature*, 522(7557):391–392, Jun 2015.
- [3] Marianne Abifadel, Sandy Elbitar, Petra El Khoury, Youmna Ghaleb, Mélody Chémaly, Marie-Line Moussalli, Jean-Pierre Rabès, Mathilde Varret, and Catherine Boileau. Living the pcsk9 adventure: from the identification of a new gene in familial hypercholesterolemia towards a potential new class of anticholesterol drugs. *Current Atherosclerosis Reports*, 16(9), Jul 2014.
- [4] Tomasz Adamusiak, Helen Parkinson, Juha Muiilu, Erik Roos, Kasper Joeri van der Velde, Gudmundur A. Thorisson, Myles Byrne, Chao Pang, Sirisha Gollapudi, Vincent Ferretti, and et al. Observ-om and observ-tab: Universal syntax solutions for the integration, search, and exchange of phenotype and genotype information. *Human Mutation*, 33(5):867–873, Apr 2012.

BIBLIOGRAPHY

- [5] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, Apr 2010.
- [6] Alan Agresti. *Categorical data analysis*, volume 359. John Wiley & Sons, 2002.
- [7] Frank W. Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, Feb 2015.
- [8] R. Alberts, G. Vera, and R. C. Jansen. affygg: computational protocols for genetical genomics with affymetrix arrays. *Bioinformatics*, 24(3):433–434, Dec 2007.
- [9] Akram Alyass, Michelle Turcotte, and David Meyre. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics*, 8(1), Jun 2015.
- [10] Peter Amstutz, Michael R. Crusoe, Nebojša Tijanić, Brad Chapman, John Chilton, Michael Heuer, Andrey Kartashov, Dan Leehr, Hervé Ménager, Maya Nedeljkovich, Matt Scales, Stian Soiland-Reyes, and Luka Stojanovic. Common workflow language, v1.0. *Figshare*, 7 2016.
- [11] Orphanet: an online database of rare diseases and orphan drugs. Copyright with INSERM 1997. Available: <http://www.orpha.net>. [accessed 14 nov 2016]. *URL*, 2016.
- [12] AndroMDA. Available: <http://www.andromda.org/>. *URL*, 2010.
- [13] Alberto Anguita, Miguel García-Remesal, Diana de la Iglesia, and Victor Maojo. Ncbi2rdf: Enabling full rdf-based access to ncbi databases. *BioMed Research International*, 2013:1–9, 2013.

- [14] D. Arends, K. J. van der Velde, P. Prins, K. W. Broman, S. Moller, R. C. Jansen, and M. A. Swertz. xqtl workbench: a scalable web environment for multi-level qtl analysis. *Bioinformatics*, 28(7):1042–1044, Feb 2012.
- [15] Danny Arends, Pjotr Prins, Ritsert C. Jansen, and Karl W. Broman. R/qtl: high-throughput multiple qtl mapping. *Bioinformatics*, 26(23):2990–2992, Oct 2010.
- [16] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, and et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.
- [17] Charles Auffray, Rudi Balling, Inês Barroso, László Bencze, Mikael Benson, Jay Bergeron, Enrique Bernal-Delgado, Niklas Blomberg, Christoph Bock, Ana Conesa, and et al. Making sense of big data in health research: Towards an eu action plan. *Genome Med*, 8(1), Jun 2016.
- [18] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, and et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Sep 2015.
- [19] O. T. Avery. 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. *Journal of Experimental Medicine*, 79(2):137–158, Feb 1944.
- [20] José L. Badano, Stephen J. Ansley, Carmen C. Leitch, Richard Alan Lewis, James R. Lupski, and Nicholas Katsanis.

- Identification of a novel bardet-biedl syndrome protein, *bbs7*, that shares structural features with *bbs1* and *bbs2*. *The American Journal of Human Genetics*, 72(3):650–658, Mar 2003.
- [21] J. S. Bailey, L. Grabowski-Boase, B. M. Steffy, T. Wiltshire, G. A. Churchill, and L. M. Tarantino. Identification of quantitative trait loci for locomotor activation and anxiety using closely related inbred strains. *Genes, Brain and Behavior*, 7(7):761–769, Oct 2008.
- [22] Evan H. Baugh, Riley Simmons-Edler, Christian L. Mueller, Rebecca F. Alford, Natalia Volfovsky, Alex E. Lash, and Richard Bonneau. Robust classification of protein variation using structural modeling and large-scale data integration. Oct 2015.
- [23] L. Bavoil, S.P. Callahan, P.J. Crossno, J. Freire, C.E. Scheidegger, C.T. Silva, and H.T. Vo. Vistrails: Enabling interactive multiple-view visualizations. *VIS 05. IEEE Visualization, 2005*.
- [24] G. W. Beadle and E. L. Tatum. Genetic control of biochemical reactions in *neurospora*. *Proceedings of the National Academy of Sciences*, 27(11):499–506, Nov 1941.
- [25] Wesley G. Beamer. Quantitative trait loci for bone density in *c57bl/6j* and *cast/eij* inbred mice. *Mammalian Genome*, 10(11):1043–1049, Nov 1999.
- [26] Ashwin Belle, Raghuram Thiagarajan, S. M. Reza Soroushmehr, Fatemeh Navidi, Daniel A. Beard, and Kayvan Najarian. Big data analytics in healthcare. *BioMed Research International*, 2015:1–16, 2015.
- [27] Daniel W. Belsky, Terrie E. Moffitt, Karen Sugden, Benjamin Williams, Renate Houts, Jeanette McCarthy, and Avshalom Caspi. Development and evaluation of a genetic risk score for

- obesity. *Biodemography and Social Biology*, 59(1):85–100, Jan 2013.
- [28] Jaroslav Bendl, Miloš Musil, Jan Štourač, Jaroslav Zendulka, Jiří Damborský, and Jan Brezovský. Predictsnp2: A unified platform for accurately evaluating snp effects by exploiting the different characteristics of variants in distinct genomic regions. *PLoS Comput Biol*, 12(5):e1004962, May 2016.
- [29] Jonathan S Berg, Muin J Khoury, and James P Evans. Deploying whole genome sequencing in clinical practice and public health: Meeting the challenge one bin at a time. *Genetics in Medicine*, 13(6):499–504, May 2011.
- [30] Sanjiv V Bhave, Cheryl Hornbaker, Tzu L Phang, Laura Saba, Razvan Lapadat, Katherina Kechris, Jeanette Gaydos, Daniel McGoldrick, Andrew Dolbey, Sonia Leach, and et al. The phenogen informatics website: tools for analyses of complex traits. *BMC Genetics*, 8(1):59, 2007.
- [31] A. Bird. Dna methylation patterns and epigenetic memory. *Genes and Development*, 16(1):6–21, Jan 2002.
- [32] Marc Jan Bonder, Alexander Kurilshikov, Etti F Tigchelaar, Zlatan Mujagic, Floris Imhann, Arnau Vich Vila, Patrick Deelen, Tommi Vatanen, Melanie Schirmer, Sanne P Smeeckens, and et al. The effect of host genetics on the gut microbiome. *Nature Genetics*, 48(11):1407–1412, Oct 2016.
- [33] Imane Boudellioua, Rozaimi B. Mahamad Razali, Maxat Kulmanov, Yasmeen Hashish, Vladimir B. Bajic, Eva Goncalves-Serra, Nadia Schoenmakers, Georgios V. Gkoutos, Paul N. Schofield, and Robert Hoehndorf. Semantic prioritization of novel causative genomic variants. *PLOS Computational Biology*, 13(4):e1005500, Apr 2017.

BIBLIOGRAPHY

- [34] Sarah Bowdin, Peter N. Ray, Ronald D. Cohn, and M. Stephen Meyn. The genome clinic: A multidisciplinary approach to assessing the opportunities and challenges of integrating genomic analysis into clinical care. *Human Mutation*, 35(5):513–519, Apr 2014.
- [35] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186, Jun 2017.
- [36] Alvis Brazma, Maria Krestyaninova, and Ugis Sarkans. Standards for systems biology. *Nature Reviews Genetics*, 7(8):593–605, Aug 2006.
- [37] R. B. Brem. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, Mar 2002.
- [38] K. W. Broman, H. Wu, S. Sen, and G. A. Churchill. R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*, 19(7):889–890, May 2003.
- [39] A. J. Brookes. Hgbase: a database of snps and other variations in and around human genes. *Nucleic Acids Research*, 28(1):356–360, Jan 2000.
- [40] Anthony J. Brookes and Peter N. Robinson. Human genotype–phenotype databases: aims, challenges and opportunities. *Nature Reviews Genetics*, 16(12):702–715, Nov 2015.
- [41] S.D.M. Brown, P. Chambon, and M. Hrabé de Angelis. Empress: standardized phenotype screens for functional annotation of the mouse genome. *Nature Genetics*, 37(11):1155–1155, Nov 2005.
- [42] Catherine A Brownstein, Alan H Beggs, Nils Homer, Barry Merriam, Timothy W Yu, Katherine C Flannery, Elizabeth T DeChene, Meghan C Towne, Sarah K Savage, Emily N Price, and

- et al. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the clarity challenge. *Genome Biology*, 15(3):R53, 2014.
- [43] H. Byelas, M. Dijkstra, P. Neerincx, F. van Dijk, A. Kanterakis, P. Deelen, and M. Swertz. Scaling bio-analyses from computational clusters to grids. *Proceedings of the 5th International Workshop on Science Gateways, Zurich, Switzerland*, 2013.
- [44] Sara A. Byron, Kendall R. Van Keuren-Jensen, David M. Engelthaler, John D. Carpten, and David W. Craig. Translating rna sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics*, 17(5):257–271, Mar 2016.
- [45] Leonid Bystrykh, Ellen Weersing, Bert Dontje, Sue Sutton, Mathew T Pletcher, Tim Wiltshire, Andrew I Su, Edo Vellenga, Jintao Wang, Kenneth F Manly, and et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using “genetical genomics”. *Nature Genetics*, 37(3):225–232, Feb 2005.
- [46] Combined Annotation Dependent Depletion (CADD). Available: <http://cadd.gs.washington.edu/info>. [accessed 1 oct 2015]. URL, 2015.
- [47] Alison Callahan, Michel Dumontier, and Nigam H Shah. Hyque: evaluating hypotheses using semantic web technologies. *Journal of Biomedical Semantics*, 2(Suppl 2):S3, 2011.
- [48] V. J. Carey, M. Morgan, S. Falcon, R. Lazarus, and R. Gentleman. Ggtools: analysis of genetics of gene expression in bioconductor. *Bioinformatics*, 23(4):522–523, Dec 2006.
- [49] Christopher A. Cassa, Mark Y. Tong, and Daniel M. Jordan. Large numbers of genetic variants considered to be pathogenic

- are common in asymptomatic individuals. *Human Mutation*, 34(9):1216–1220, Aug 2013.
- [50] Christopher A Cassa, Donate Weghorn, Daniel J Balick, Daniel M Jordan, David Nusinow, Kaitlin E Samocha, Anne O’Donnell-Luria, Daniel G MacArthur, Mark J Daly, David R Beier, and et al. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nature Genetics*, 49(5):806–810, Apr 2017.
- [51] Fabrice Caudron and Yves Barral. A super-assembly of whi3 encodes memory of deceptive encounters by single cells during yeast courtship. *Cell*, 155(6):1244–1257, Dec 2013.
- [52] Sohini Chakrabortee, James S. Byers, Sandra Jones, David M. Garcia, Bhupinder Bhullar, Amelia Chang, Richard She, Laura Lee, Brayon Fremin, Susan Lindquist, and et al. Intrinsically disordered proteins drive emergence and inheritance of biological traits. *Cell*, 167(2):369–381.e12, Oct 2016.
- [53] Xiao Chang and Kai Wang. wannovar: annotating genetic variants for personal genomes via the web. *Journal of Medical Genetics*, 49(7):433–436, Jun 2012.
- [54] Rong Chen, Lisong Shi, Jörg Hakenberg, Brian Naughton, Pamela Sklar, Jianguo Zhang, Hanlin Zhou, Lifeng Tian, Om Prakash, Mathieu Lemire, and et al. Analysis of 589,306 genomes identifies individuals resilient to severe mendelian childhood diseases. *Nature Biotechnology*, 34(5):531–538, Apr 2016.
- [55] Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–1222, Dec 2015.

- [56] Y.-C. Cheng, F.-C. Hsiao, E.-C. Yeh, W.-J. Lin, C.-Y. L. Tang, H.-C. Tseng, H.-T. Wu, C.-K. Liu, C.-C. Chen, Y.-T. Chen, and et al. Variowatch: providing large-scale and comprehensive annotations on human genomic variants in the next generation sequencing era. *Nucleic Acids Research*, 40(W1):W76–W81, May 2012.
- [57] Elissa J Chesler, Lu Lu, Siming Shou, Yanhua Qu, Jing Gu, Jintao Wang, Hui Chen Hsu, John D Mountz, Nicole E Baldwin, Michael A Langston, and et al. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics*, 37(3):233–242, Feb 2005.
- [58] Colby Chiang, Ryan M Layer, Gregory G Faust, Michael R Lindberg, David B Rose, Erik P Garrison, Gabor T Marth, Aaron R Quinlan, and Ira M Hall. Speedseq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, 12(10):966–968, Aug 2015.
- [59] Yongwook Choi, Gregory E. Sims, Sean Murphy, Jason R. Miller, and Agnes P. Chan. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE*, 7(10):e46688, Oct 2012.
- [60] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff. *Fly*, 6(2):80–92, Apr 2012.
- [61] W. R. Cnossen, R. H. M. te Morsche, A. Hoischen, C. Gilissen, M. Chrispijn, H. Venselaar, S. Mehdi, C. Bergmann, J. A. Veltman, and J. P. H. Drenth. Whole-exome sequencing reveals lrp5 mutations and canonical wnt signaling associated with hepatic cystogenesis. *Proceedings of the National Academy of Sciences*, 111(14):5343–5348, Mar 2014.

BIBLIOGRAPHY

- [62] E. F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, Jun 1970.
- [63] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Oct 2012.
- [64] CASIMIR EU consortium for Coordination and Sustainability of International Mouse Informatics Resources. Available: <http://www.casimir.org.uk>. URL, 2010.
- [65] GEN2PHEN EU consortium to unify human Genotype-To-Phenotype databases. Available: <http://www.gen2phen.org>. URL, 2010.
- [66] David N. Cooper, Michael Krawczak, Constantin Polychronakos, Chris Tyler-Smith, and Hildegard Kehrer-Sawatzki. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human Genetics*, 132(10):1077–1130, Jul 2013.
- [67] Gregory M. Cooper. Parlez-vous vus? *Genome Research*, 25(10):1423–1426, Oct 2015.
- [68] Gregory M. Cooper and Jay Shendure. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, 12(9):628–640, Aug 2011.
- [69] Manuel Corpas, Willy Valdivia-Granda, Nazareth Torres, Bastian Greshake, Alain Coletta, Alexej Knaus, Andrew P. Harrison, Mike Cariaso, Federico Moran, Fiona Nielsen, and et al. Crowdsourced direct-to-consumer genomic analysis of a family quartet. *BMC Genomics*, 16(1), Nov 2015.
- [70] Mercè Crosas. The rise of data publishing (and how dataverse 4 can help). *Journal of Technology Science*, Working Paper.

- [71] Beryl B. Cummings, Jamie L. Marshall, Taru Tukiainen, Monkol Lek, Sandra Donkervoort, A. Reghan Foley, Veronique Bolduc, Leigh B. Waddell, Sarah A. Sandaradura, Gina L. O'Grady, and et al. Improving genetic diagnosis in mendelian disease with transcriptome sequencing. *Science Translational Medicine*, 9(386):eaal5209, Apr 2017.
- [72] Felipe da Veiga Leprevost, Björn A. Grüning, Saulo Alves Aflitos, Hannes L. Röst, Julian Uszkoreit, Harald Barsnes, Marc Vaudel, Pablo Moreno, Laurent Gatto, Jonas Weber, and et al. Biocontainers: an open-source and community-driven framework for software standardization. *Bioinformatics*, Mar 2017.
- [73] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, and et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, Jun 2011.
- [74] H. Daoud, S. M. Luco, R. Li, E. Bareke, C. Beaulieu, O. Jarinova, N. Carson, S. M. Nikkel, G. E. Graham, J. Richer, and et al. Next-generation sequencing for diagnosis of rare diseases in the neonatal intensive care unit. *Canadian Medical Association Journal*, 188(11):E254–E260, May 2016.
- [75] XGAP data sets. Available: <http://www.xgap.org/wiki/datasets>. URL, 2010.
- [76] Eugene V. Davydov, David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS Computational Biology*, 6(12):e1001025, Dec 2010.
- [77] Patrick Deelen, Androniki Menelaou, Elisabeth M van Leeuwen, Alexandros Kanterakis, Freerk van Dijk, Carolina Medina-Gomez,

BIBLIOGRAPHY

- Laurent C Francioli, Jouke Jan Hottenga, Lennart C Karssen, Karol Estrada, and et al. Improved imputation quality of low-frequency and rare variants in european samples using the “genome of the netherlands”. *European Journal of Human Genetics*, 22(11):1321–1326, Jun 2014.
- [78] Patrick Deelen, Daria V Zhernakova, Mark de Haan, Marijke van der Sijde, Marc Jan Bonder, Juha Karjalainen, K Joeri van der Velde, Kristin M Abbott, Jingyuan Fu, Cisca Wijmenga, and et al. Calling genotypes from public rna-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Medicine*, 7(1), Mar 2015.
- [79] Johan T. den Dunnen and Mark H. Paalman. Standardizing mutation nomenclature: Why bother? *Human Mutation*, 22(3):181–182, Aug 2003.
- [80] Anna L Dixon, Liming Liang, Miriam F Moffatt, Wei Chen, Simon Heath, Kenny C C Wong, Jenny Taylor, Edward Burnett, Ivo Gut, Martin Farrall, and et al. A genome-wide association study of global gene expression. *Nature Genetics*, 39(10):1202–1207, Sep 2007.
- [81] Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, Apr 2012.
- [82] Neil Donald, Salim Malik, Joshua L. McGuire, and Kevin J. Monahan. The association of low penetrance genetic risk modifiers with colorectal cancer in lynch syndrome patients: a systematic review and meta-analysis. *Familial Cancer*, May 2017.
- [83] A. Doroszuk, L. B. Snoek, E. Fradin, J. Riksen, and J. Kam-menga. A genome-wide library of cb4856/n2 introgression

- lines of *caenorhabditis elegans*. *Nucleic Acids Research*, 37(16):e110–e110, Jun 2009.
- [84] Frank Dudbridge. Correction: Power and predictive accuracy of polygenic risk scores. *PLoS Genet*, 9(4), Apr 2013.
- [85] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Fretze, Jennifer Harrow, Rajinder Kaul, and et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.
- [86] C. Durrant, M. A. Swertz, R. Alberts, D. Arends, S. Moller, R. Mott, P. Prins, K. J. van der Velde, R. C. Jansen, and K. Schughart. Bioinformatics tools and database resources for systems genetics analysis in mice—a short review and an evaluation of future needs. *Briefings in Bioinformatics*, 13(2):135–142, Jul 2011.
- [87] A. Eberharter. Histone acetylation: a switch between repressive and permissive chromatin: Second in review series on chromatin dynamics. *EMBO Reports*, 3(3):224–229, Mar 2002.
- [88] Editorial. Pinpointing expression differences. *Nature Genetics*, 39(10), Sep 2007.
- [89] Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. *Genome Biology*, 6(5):R44, 2005.
- [90] Karen Eilbeck, Aaron Quinlan, and Mark Yandell. Settling the score: variant prioritization and mendelian disease. *Nature Reviews Genetics*, 18(10):599–612, Aug 2017.
- [91] The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *c. elegans*: A platform for investigating biology. *Science*, 282(5396):2012–2018, Dec 1998.

BIBLIOGRAPHY

- [92] Mark Elvin, Laurens B Snoek, Martin Frejno, Ulrike Klemstein, Jan E Kammenga, and Gino B Poulin. A fitness assay for comparing rnai effects across multiple *c. elegans* genotypes. *BMC Genomics*, 12(1), Oct 2011.
- [93] Maria Eriksson, W. Ted Brown, Leslie B. Gordon, Michael W. Glynn, Joel Singer, Laura Scott, Michael R. Erdos, Christiane M. Robbins, Tracy Y. Moses, Peter Berglund, and et al. Recurrent de novo point mutations in lamin a cause hutchinson–gilford progeria syndrome. *Nature*, 423(6937):293–298, Apr 2003.
- [94] Brendan M. Everett, Robert J. Smith, and William R. Hiatt. Reducing ldl with pcsk9 inhibitors — the clinical benefit of lipid drugs. *New England Journal of Medicine*, 373(17):1588–1591, Oct 2015.
- [95] NHLBI GO Exome Sequencing Project (ESP) at Seattle WA Exome Variant Server. Available: <http://evs.gs.washington.edu/evs/>. [accessed 20 oct 2014]. URL, 2014.
- [96] XGAP eXtensible Genotype and Phenotype platform. Available: <http://www.xgap.org>. URL, 2010.
- [97] Antonio Fabregat, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, Steven Jupe, Florian Korninger, Sheldon McKay, and et al. The reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1):D481–D487, Dec 2015.
- [98] David S Fay. Classical genetics goes high-tech. *Nature Methods*, 5(10):863–864, Oct 2008.
- [99] Rudolf S N Fehrmann, Juha M Karjalainen, Małgorzata Krajewska, Harm-Jan Westra, David Maloney, Anton Simeonov, Tune H Pers, Joel N Hirschhorn, Ritsert C Jansen, Erik A Schultes, and

- et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genetics*, 47(2):115–125, Jan 2015.
- [100] Feng, R Bao, L Huang, J Andrade, W Tan, W Kibbe, and Hongmei Jiang. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Informatics*, page 67, Sep 2014.
- [101] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, and et al. Complete nucleotide sequence of bacteriophage ms2 rna: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507, Apr 1976.
- [102] Helen V. Firth, Shola M. Richards, A. Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M. Pettett, and Nigel P. Carter. Decipher: Database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics*, 84(4):524–533, Apr 2009.
- [103] MOLGENIS flexible biosoftware generation toolkit. Available: <http://www.molgenis.org>. URL, 2010.
- [104] MIQAS Minimum Information for QTLs and Association Studies. Available: <http://miqas.sourceforge.net/>. URL, 2010.
- [105] Alistair R. R. Forrest, Hideya Kawaji, Michael Rehli, J. Kenneth Baillie, Michiel J. L. de Hoon, Vanja Haberle, Timo Lassmann, Ivan V. Kulakovskiy, Marina Lizio, Masayoshi Itoh, and et al. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470, Mar 2014.
- [106] Eric J Foss, Dragan Radulovic, Scott A Shaffer, Douglas M Ruderfer, Antonio Bedalov, David R Goodlett, and Leonid Kruglyak.

BIBLIOGRAPHY

- Genetic basis of proteome variation in yeast. *Nature Genetics*, 39(11):1369–1375, Oct 2007.
- [107] Martin Franke, Daniel M. Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, Rieke Kempfer, Ivana Jerković, Wing-Lee Chan, and et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624):265–269, Oct 2016.
- [108] D. Fredman. Hgvbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Research*, 30(1):387–391, Jan 2002.
- [109] Mélissa Yana Frédéric, Marine Lalande, Catherine Boileau, Dalil Hamroun, Mireille Claustres, Christophe Bérout, and Gwenaelle Collod-Bérout. Umd-predictor, a new prediction tool for nucleotide substitution pathogenicity-application to four genes: fbn1, fbn2, tgfbr1, and tgfbr2. *Human Mutation*, 30(6):952–959, Jun 2009.
- [110] Jingyuan Fu, Joost J B Keurentjes, Harro Bouwmeester, Twan America, Francel W A Verstappen, Jane L Ward, Michael H Beale, Ric C H de Vos, Martijn Dijkstra, Richard A Scheltema, and et al. System-wide molecular evidence for phenotypic buffering in arabidopsis. *Nature Genetics*, 41(2):166–167, Jan 2009.
- [111] Jingyuan Fu, Morris A Swertz, Joost JB Keurentjes, and Ritsert C Jansen. Metanetwork: a computational protocol for the genetic study of metabolic networks. *Nature Protocols*, 2(3):685–694, Mar 2007.
- [112] Yao Fu, Zhu Liu, Shaoke Lou, Jason Bedford, Xinmeng Jasmine Mu, Kevin Y Yip, Ekta Khurana, and Mark Gerstein. Funseq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology*, 15(10), Oct 2014.

- [113] Anna K Füzéry, Joshua Levin, Maria M Chan, and Daniel W Chan. Translation of proteomic biomarkers into fda approved cancer diagnostics: issues and challenges. *Clinical Proteomics*, 10(1):13, 2013.
- [114] Bryn E. Gaertner and Patrick C. Phillips. Caenorhabditis elegans as a platform for molecular quantitative genetics and the systems biology of natural variation. *Genet. Res.*, 92(5-6):331–348, Dec 2010.
- [115] M. Y. Galperin and G. R. Cochrane. Nucleic acids research annual database issue and the nar online molecular biology database collection in 2009. *Nucleic Acids Research*, 37(Database):D1–D4, Jan 2009.
- [116] Antonios N. Gargalionis, Christina Piperi, Christos Adamopoulos, and Athanasios G. Papavassiliou. Histone modifications as a pathogenic mechanism of colorectal tumorigenesis. *The International Journal of Biochemistry & Cell Biology*, 44(8):1276–1289, Aug 2012.
- [117] Archibald E. Garrod. The incidence of alkaptonuria : A study in chemical individuality. *The Lancet*, 160(4137):16161620, Dec 1902.
- [118] A. Gaye, Y. Marcon, J. Isaeva, P. LaFlamme, A. Turner, E. M. Jones, J. Minion, A. W. Boyd, C. J. Newby, M.-L. Nuotio, and et al. Datashield: taking the analysis to the data, not the data to the analysis. *International Journal of Epidemiology*, 43(6):1929–1944, Sep 2014.
- [119] Andrea M. Gazzo, Dorien Daneels, Elisa Cilia, Maryse Bonduelle, Marc Abramowicz, Sonia Van Dooren, Guillaume Smits, and Tom Lenaerts. Dida: A curated and annotated digenic diseases database. *Nucleic Acids Research*, 44(D1):D900–D907, Oct 2015.

BIBLIOGRAPHY

- [120] Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.
- [121] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, and et al. *Genome Biology*, 5(10):R80, 2004.
- [122] M. B. Gerstein, Z. J. Lu, E. L. Van Nostrand, C. Cheng, B. I. Arshinoff, T. Liu, K. Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, and et al. Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science*, 330(6012):1775–1787, Dec 2010.
- [123] Belinda Giardine, Cathy Riemer, Tim Hefferon, Daryl Thomas, Fan Hsu, Julian Zielenski, Yunhua Sang, Laura Elnitski, Garry Cutting, Heather Trumbower, and et al. Phencode: connecting encode data with mutations and phenotype. *Human Mutation*, 28(6):554–562, 2007.
- [124] Geoffrey Ginsburg. Medical genomics: Gather and use genetic data in health care. *Nature*, 508(7497):451–453, Apr 2014.
- [125] Marta Girdea, Sergiu Dumitriu, Marc Fiume, Sarah Bowdin, Kym M. Boycott, Sébastien Chénier, David Chitayat, Hanna Faghfoury, M. Stephen Meyn, Peter N. Ray, and et al. Phenotips: Patient phenotyping software for clinical and research use. *Human Mutation*, 34(8):1057–1065, May 2013.
- [126] Brigitte Glanzmann, Hendri Herbst, Craig J. Kinnear, Marlo Möller, Junaid Gamielien, and Soraya Bardiën. A new tool for prioritization of sequence variants from whole exome sequencing data. *Source Code for Biology and Medicine*, 11(1), Jul 2016.

- [127] C. A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, and et al. myexperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research*, 38(Web Server):W677–W682, May 2010.
- [128] Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.
- [129] Abel González-Pérez and Nuria López-Bigas. Improving the assessment of the outcome of nonsynonymous snvs with a consensus deleteriousness score, condel. *The American Journal of Human Genetics*, 88(4):440–449, Apr 2011.
- [130] GA Nagana Gowda, Shucha Zhang, Haiwei Gu, Vincent Asiago, Narasimhamurthy Shanaiah, and Daniel Raftery. Metabolomics-based methods for early disease diagnostics. *Expert Review of Molecular Diagnostics*, 8(5):617–633, Sep 2008.
- [131] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–864, Sep 1974.
- [132] Dmitry Grapov, Johannes Fahrman, and Kwanjeera Wanichtharak. Genomic, proteomic, and metabolomic data integration strategies. *BMI*, page 1, Sep 2015.
- [133] J W M Green, L B Snoek, J E Kammenga, and S C Harvey. Genetic mapping of variation in dauer larvae development in growing populations of *caenorhabditis elegans*. *Heredity*, 111(4):306–313, May 2013.
- [134] Robert C. Green, Jonathan S. Berg, Wayne W. Grody, Sarah S. Kalia, Bruce R. Korf, Christa L. Martin, Amy L. McGuire, Robert L. Nussbaum, Julianne M. O'Daniel, Kelly E. Ormond,

BIBLIOGRAPHY

- and et al. Acmg recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine*, 15(7):565–574, Jun 2013.
- [135] Matthias Gstaiger and Ruedi Aebersold. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nature Reviews Genetics*, 10(9):617–627, Sep 2009.
- [136] Brad Gulko, Melissa J Hubisz, Ilan Gronau, and Adam Siepel. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature Genetics*, 47(3):276–283, Jan 2015.
- [137] Yunfei Guo, Xiaolei Ding, Yufeng Shen, Gholson J. Lyon, and Kai Wang. Seqmule: automated pipeline for analysis of human exome/genome sequencing data. *Scientific Reports*, 5(1), Sep 2015.
- [138] Ayal B. Gussow, Brett R. Copeland, Ryan S. Dhindsa, Quanli Wang, Slavé Petrovski, William H. Majoros, Andrew S. Allen, and David B. Goldstein. Orion: Detecting regions of the human non-coding genome that are intolerant to variation using population genetics. *PLOS ONE*, 12(8):e0181604, Aug 2017.
- [139] E W Gutteling, A Doroszuk, J A G Riksen, Z Prokop, J Reszka, and J E Kammenga. Environmental influence on the genetic correlations between life-history traits in *caenorhabditis elegans*. *Heredity*, 98(4):206–213, Jan 2007.
- [140] E W Gutteling, J A G Riksen, J Bakker, and J E Kammenga. Mapping phenotypic plasticity and genotype–environment interactions affecting life-history traits in *caenorhabditis elegans*. *Heredity*, 98(1):28–37, Sep 2006.
- [141] Harald H H Göring, Joanne E Curran, Matthew P Johnson, Thomas D Dyer, Jac Charlesworth, Shelley A Cole, Jeremy B M

- Jowett, Lawrence J Abraham, David L Rainwater, Anthony G Co-muzzie, and et al. Discovery of expression qtls using large-scale transcriptional profiling in human lymphocytes. *Nature Genetics*, 39(10):1208–1216, Sep 2007.
- [142] Ada Hamosh, Alan F. Scott, Joanna Amberger, David Valle, and Victor A. McKusick. Online mendelian inheritance in man (omim). *Human Mutation*, 15(1):57–61, Jan 2000.
- [143] Sam Hanash. Disease proteomics. *Nature*, 422(6928):226–232, Mar 2003.
- [144] Lee Harland. Open phacts: A semantic knowledge infrastructure for public and commercial drug discovery research. *Knowledge Engineering and Knowledge Management*, page 1–7, 2012.
- [145] Andrew R. Harper, Shalini Nayee, and Eric J. Topol. Protective alleles and modifier variants in human health and disease. *Nature Reviews Genetics*, 16(12):689–701, Oct 2015.
- [146] Mary Harris. Would you want to know the secrets hidden in your baby's genes? *National Public Radio*, 2016 October 31, 20163:42 AM ET.
- [147] Simon C Harvey, Alison Shorto, and Mark E Viney. Quantitative genetic analysis of life-history traits of *caenorhabditis elegans* in stressful environments. *BMC Evolutionary Biology*, 8(1):15, 2008.
- [148] Graham A Heap, Gosia Trynka, Ritsert C Jansen, Marcel Bruinenberg, Morris A Swertz, Lotte C Dinesen, Karen A Hunt, Cisca Wijmenga, David A vanHeel, and Lude Franke. Complex nature of snp genotype effects on gene expression in primary human leucocytes. *BMC Medical Genomics*, 2(1), Jan 2009.

BIBLIOGRAPHY

- [149] M. Rebecca Heiner-Fokkema, Frédéric M. Vaz, Ronald Maatman, Leo A. J. Kluijtmans, Francjan J. van Spronsen, and Dirk-Jan Reijngoud. Reliable diagnosis of carnitine palmitoyltransferase type ia deficiency by analysis of plasma acylcarnitine profiles. *JIMD Reports*, 2016.
- [150] A. D. Hershey. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology*, 36(1):39–56, Sep 1952.
- [151] Holger Heyn and Manel Esteller. Dna methylation profiling in the clinic: applications and challenges. *Nature Reviews Genetics*, 13(10):679–692, Sep 2012.
- [152] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, May 2009.
- [153] Jacob Shujui Hsu, Johnny S.H. Kwan, Zhicheng Pan, Maria-Mercè Garcia-Barcelo, Pak Chung Sham, and Miaoxin Li. Inheritance-mode specific pathogenicity prioritization (ispp) for human protein coding genes. *Bioinformatics*, 32(20):3065–3071, Jun 2016.
- [154] Z.-L. Hu, E. R. Fritz, and J. M. Reecy. Animalqtl db: a livestock qtl database tool set for positional qtl information mining and beyond. *Nucleic Acids Research*, 35(Database):D604–D609, Jan 2007.
- [155] Yi-Fei Huang, Brad Gulko, and Adam Siepel. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature Genetics*, 49(4):618–624, Mar 2017.

- [156] Norbert Hubner, Caroline A Wallace, Heike Zimdahl, Enrico Petretto, Herbert Schulz, Fiona Maciver, Michael Mueller, Oliver Hummel, Jan Monti, Vaclav Zidek, and et al. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*, 37(3):243–253, Feb 2005.
- [157] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(Web Server):W729–W732, Jul 2006.
- [158] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, Sep 1996.
- [159] Nilah M. Ioannidis, Joseph H. Rothstein, Vikas Pejaver, Sumit Middha, Shannon K. McDonnell, Saurabh Baheti, Anthony Musolf, Qing Li, Emily Holzinger, Danielle Karyadi, and et al. Revel: An ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, 99(4):877–885, Oct 2016.
- [160] Iuliana Ionita-Laza, Kenneth McCallum, Bin Xu, and Joseph D Buxbaum. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet*, 48(2):214–220, Jan 2016.
- [161] R. A. Irizarry. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, Apr 2003.
- [162] J. Ison, M. Kalas, I. Jonassen, D. Bolser, M. Uludag, H. McWilliam, J. Malone, R. Lopez, S. Pettifer, and P. Rice. Edam: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, 29(10):1325–1332, Mar 2013.

BIBLIOGRAPHY

- [163] Jon Ison, Kristoffer Rapacki, Hervé Ménager, Matúš Kalaš, Emil Rydza, Piotr Chmura, Christian Anthon, Niall Beard, Karel Berka, Dan Bolser, and et al. Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Research*, 44(D1):D38–D47, Nov 2015.
- [164] Yuval Itan, Lei Shang, Bertrand Boisson, Michael J Ciancanelli, Janet G Markle, Ruben Martinez-Barricarte, Eric Scott, Ishaan Shah, Peter D Stenson, Joseph Gleeson, and et al. The mutation significance cutoff: gene-level thresholds for variant predictions. *Nat Meth*, 13(2):109–110, Jan 2016.
- [165] Yuval Itan, Lei Shang, Bertrand Boisson, Etienne Patin, Alexandre Bolze, Marcela Moncada-Vélez, Eric Scott, Michael J. Ciancanelli, Fabien G. Lafaille, Janet G. Markle, and et al. The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc Natl Acad Sci USA*, 112(44):13615–13620, Oct 2015.
- [166] Daniel Jameson, Kevin Garwood, Chris Garwood, Tim Booth, Pinar Alper, Stephen G Oliver, and Norman W Paton. Data capture in bioinformatics: requirements and experiences with pedro. *BMC Bioinformatics*, 9(1):183, 2008.
- [167] R Jansen. Genetical genomics: the added value from segregation. *Trends in Genetics*, 17(7):388–391, Jul 2001.
- [168] Biola M. Javierre, Oliver S. Burren, Steven P. Wilder, Roman Kreuzhuber, Steven M. Hill, Sven Sewitz, Jonathan Cairns, Steven W. Wingett, Csilla Várnai, Michiel J. Thiecke, and et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, 167(5):1369–1384.e19, Nov 2016.
- [169] Arthur Jochems, Timo M. Deist, Johan van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and

- Andre Dekker. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – a real life proof of concept. *Radiotherapy and Oncology*, 121(3):459–467, Dec 2016.
- [170] W. Johannsen. Elemente der exakten erblichkeitslehre. *Gustav Fischer Verlag, Jena*, 1909.
- [171] Andrew R Jones, Michael Miller, Ruedi Aebersold, Rolf Apweiler, Catherine A Ball, Alvis Brazma, James DeGreef, Nigel Hardy, Henning Hermjakob, Simon J Hubbard, and et al. The functional genomics experiment model (fuge): an extensible framework for standards in functional genomics. *Nature Biotechnology*, 25(10):1127–1133, Oct 2007.
- [172] Andrew R Jones and Norman W Paton. *BMC Bioinformatics*, 6(1):235, 2005.
- [173] Jan DH Jongbloed, Anna Pósfalvi, Wilhelmina S Kerstjens-Frederikse, Richard J Sinke, and J Peter van Tintelen. New clinical molecular diagnostic methods for congenital and inherited heart disease. *Expert Opinion on Medical Diagnostics*, 5(1):9–24, Dec 2010.
- [174] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, and et al. The ebi rdf platform: linked open data for the life sciences. *Bioinformatics*, 30(9):1338–1339, Jan 2014.
- [175] Marten Jäger, Kai Wang, Sebastian Bauer, Damian Smedley, Peter Krawitz, and Peter N. Robinson. Jannovar: A java library for exome annotation. *Human Mutation*, 35(5):548–555, Apr 2014.
- [176] Titus Kaletta and Michael O. Hengartner. Finding function in novel targets: *C. elegans* as a model organism. *Nature Reviews Drug Discovery*, 5(5):387–399, Apr 2006.

BIBLIOGRAPHY

- [177] Jan E. Kammenga, Agnieszka Doroszuk, Joost A. G. Riksen, Esther Hazendonk, Laurentiu Spiridon, Andrei-Jose Petrescu, Marcel Tijsterman, Ronald H. A. Plasterk, and Jaap Bakker. A *Caenorhabditis elegans* wild type defies the temperature–size rule owing to a single nucleotide polymorphism in *tra-3*. *PLoS Genet*, 3(3):e34, 2007.
- [178] Jan E. Kammenga, Patrick C. Phillips, Mario De Bono, and Agnieszka Doroszuk. Beyond induced mutants: using worms to study natural variation in genetic pathways. *Trends in Genetics*, 24(4):178–185, Apr 2008.
- [179] Peter Kampstra. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software*, 28(1):1–9, 11 2008.
- [180] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, Oct 2015.
- [181] Brett Kennedy, Zev Kronenberg, Hao Hu, Barry Moore, Steven Flygare, Martin G. Reese, Lynn B. Jorde, Mark Yandell, and Chad Huff. Using *vaast* to identify disease-associated variants in next-generation sequencing data. *Current Protocols in Human Genetics*, page 6.14.1–6.14.25, Apr 2014.
- [182] J. J. B. Keurentjes, J. Fu, I. R. Terpstra, J. M. Garcia, G. van den Ackerveken, L. B. Snoek, A. J. M. Peeters, D. Vreugdenhil, M. Koornneef, and R. C. Jansen. Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences*, 104(5):1708–1713, Jan 2007.
- [183] Joost J B Keurentjes, Jingyuan Fu, C H Ric de Vos, Arjen Lommen, Robert D Hall, Raoul J Bino, Linus H W van der Plas, Rit-

- sert C Jansen, Dick Vreugdenhil, and Maarten Koornneef. The genetics of plant metabolism. *Nature Genetics*, 38(7):842–849, Jun 2006.
- [184] Ekta Khurana, Yao Fu, Jieming Chen, and Mark Gerstein. Interpretation of genomic variants using a unified biological network approach. *PLoS Computational Biology*, 9(3):e1002886, Mar 2013.
- [185] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O’Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, Feb 2014.
- [186] Laura S. Kremer, Daniel M. Bader, Christian Mertes, Robert Kopajtich, Garwin Pichler, Arcangela Iuso, Tobias B. Haack, Elisabeth Graf, Thomas Schwarzmayr, Caterina Terrile, and et al. Genetic diagnosis of mendelian disorders via rna sequencing. *Nature Communications*, 8:15824, Jun 2017.
- [187] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nat Protoc*, 4(8):1073–1081, Jun 2009.
- [188] Martina Kutmon, Anders Riutta, Nuno Nunes, Kristina Hanspers, Egon L. Willighagen, Anwesha Bohler, Jonathan Mélius, Andra Waagmeester, Sravanthi R. Sinha, Ryan Miller, and et al. Wikipathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*, 44(D1):D488–D494, Oct 2015.
- [189] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, and et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

BIBLIOGRAPHY

- [190] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1):D980–D985, Nov 2013.
- [191] Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, and et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868, Nov 2015.
- [192] Ilkka Lappalainen, Jeff Almeida-King, Vasudev Kumanduri, Alexander Senf, John Dylan Spalding, Saif ur Rehman, Gary Saunders, Jag Kandasamy, Mario Caccamo, Rasko Leinonen, and et al. The european genome-phenome archive of human data consented for biomedical research. *Nature Genetics*, 47(7):692–695, Jun 2015.
- [193] P. Leder and M. W. Nirenberg. Rna codewords and protein synthesis, iii. on the nucleotide sequence of a cysteine and a leucine rna codeword. *Proceedings of the National Academy of Sciences*, 52(6):1521–1529, Dec 1964.
- [194] Insuk Lee, Ben Lehner, Catriona Crombie, Wendy Wong, Andrew G Fraser, and Edward M Marcotte. A single gene network accurately predicts phenotypic effects of gene perturbation in *caenorhabditis elegans*. *Nature Genetics*, 40(2):181–188, Jan 2008.
- [195] R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tarraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, and et al. The european nucleotide archive. *Nucleic Acids Research*, 39(Database):D28–D31, Oct 2010.
- [196] Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O’Donnell

- Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, and et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, Aug 2016.
- [197] Stefan H Lelieveld, Margot R F Reijnders, Rolph Pfundt, Helger G Yntema, Erik-Jan Kamsteeg, Petra de Vries, Bert B A de Vries, Marjolein H Willemsen, Tjitske Kleefstra, Katharina Löhner, and et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nature Neuroscience*, 19(9):1194–1196, Aug 2016.
- [198] Monica Leu, Keith Humphreys, Ida Surakka, Emil Rehnberg, Juha Muilu, Päivi Rosenström, Peter Almgren, Juha Jääskeläinen, Richard P Lifton, Kirsten Ohm Kyvik, and et al. Nordidb: a nordic pool and portal for genome-wide control data. *European Journal of Human Genetics*, 18(12):1322–1326, Jul 2010.
- [199] B. Li, V. G. Krishnan, M. E. Mort, F. Xin, K. K. Kamati, D. N. Cooper, S. D. Mooney, and P. Radivojac. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, 25(21):2744–2750, Sep 2009.
- [200] J. Li. Genetical genomics: combining genetics with gene expression analysis. *Human Molecular Genetics*, 14(suppl-2):R163–R169, Oct 2005.
- [201] 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):e53–e53, Jan 2012.
- [202] Quan Li and Kai Wang. Intervar: Clinical interpretation of genetic variants by the 2015 acmg-amp guidelines. *The American Journal of Human Genetics*, 100(2):267–280, Feb 2017.

BIBLIOGRAPHY

- [203] Y. Li, R. Breitling, L. B. Snoek, K. J. van der Velde, M. A. Swertz, J. Riksen, R. C. Jansen, and J. E. Kammenga. Global genetic robustness of the alternative splicing machinery in *caenorhabditis elegans*. *Genetics*, 186(1):405–410, Jul 2010.
- [204] Yang Li, Rainer Breitling, and Ritsert C. Jansen. Generalizing genetical genomics: getting added value from environmental perturbation. *Trends in Genetics*, 24(10):518–524, Oct 2008.
- [205] Yang Li, Olga Alda Álvarez, Evert W. Gutteling, Marcel Tijsterman, Jingyuan Fu, Joost A. G. Riksen, Esther Hazendonk, Pjotr Prins, Ronald H. A. Plasterk, Ritsert C. Jansen, and et al. Mapping determinants of gene expression plasticity by genetical genomics in *c. elegans*. *PLoS Genetics*, 2(12):e222, 2006.
- [206] Graham J. Lieschke and Peter D. Currie. Animal models of human disease: zebrafish swim into view. *Nature Reviews Genetics*, 8(5):353–367, May 2007.
- [207] Nita A Limdi and David L Veenstra. Warfarin pharmacogenetics. *Pharmacotherapy*, 28(9):1084–1097, Sep 2008.
- [208] Steven M Lipkin, Laura S Rozek, Gad Rennert, Wei Yang, Peng-Chieh Chen, Joseph Hacia, Nathan Hunt, Brian Shin, Steve Fodor, Mark Kokoris, and et al. The *mlh1* d132h variant is associated with susceptibility to sporadic colorectal cancer. *Nature Genetics*, 36(7):694–699, Jun 2004.
- [209] Jan-Eric Litton. We must urgently clarify data-sharing rules. *Nature*, 541(7638):437–437, Jan 2017.
- [210] Xiaoming Liu, Xueqiu Jian, and Eric Boerwinkle. dbnsfp: A lightweight database of human nonsynonymous snps and their functional predictions. *Human Mutation*, 32(8):894–899, Jul 2011.

- [211] Xiaoming Liu, Simon White, Bo Peng, Andrew D Johnson, Jennifer A Brody, Alexander H Li, Zhuoyi Huang, Andrew Carroll, Peng Wei, Richard Gibbs, and et al. Wgsa: an annotation pipeline for human genome sequencing studies. *Journal of Medical Genetics*, 53(2):111–112, Sep 2015.
- [212] Katja Lohmann and Christine Klein. Next generation sequencing and the future of genetic diagnosis. *Neurotherapeutics*, Jul 2014.
- [213] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, and et al. The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580–585, May 2013.
- [214] Rachel Lyne, Richard Smith, Kim Rutherford, Matthew Wakeling, Andrew Varley, Francois Guillier, Hilde Janssens, Wenyan Ji, Peter McLaren, Philip North, and et al. Flymine: an integrated database for drosophila and anopheles genomics. *Genome Biology*, 8(7):R129, 2007.
- [215] D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, and et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, Feb 2012.
- [216] Matthew D Mailman, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, and et al. The ncbi dbgap database of genotypes and phenotypes. *Nature Genetics*, 39(10), Sep 2007.
- [217] R. Malik, S. Bevan, M. A. Nalls, E. G. Holliday, W. J. Devan, Y.-C. Cheng, C. A. Ibrahim-Verbaas, B. F. J. Verhaaren, J. C. Bis, A. Y. Joon, and et al. Multilocus genetic risk score associates with ischemic stroke in case-control and prospective cohort studies. *Stroke*, 45(2):394–402, Jan 2014.

BIBLIOGRAPHY

- [218] Frank Manola and Eric Miller, editors. *RDF Primer*. W3C Recommendation. World Wide Web Consortium, February 2004.
- [219] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, and et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, Oct 2009.
- [220] Arjun K. Manrai, Birgit H. Funke, Heidi L. Rehm, Morten S. Olesen, Bradley A. Maron, Peter Szolovits, David M. Margulies, Joseph Loscalzo, and Isaac S. Kohane. Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, 375(7):655–665, Aug 2016.
- [221] Francisco Martínez, Alfonso Caro-Llopis, Mónica Roselló, Silvestre Oltra, Sonia Mayo, Sandra Monfort, and Carmen Orellana. High diagnostic yield of syndromic intellectual disability by targeted next-generation sequencing. *Journal of Medical Genetics*, page jmedgenet–2016–103964, Sep 2016.
- [222] Cheryl A. Mather, Sean D. Mooney, Stephen J. Salipante, Sheena Scroggins, David Wu, Colin C. Pritchard, and Brian H. Shirts. Cadd score has limited clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel. *Genetics in Medicine*, May 2016.
- [223] Davis J McCarthy, Peter Humburg, Alexander Kanapin, Manuel A Rivas, Kyle Gaulton, Jean-Baptiste Cazier, and Peter Donnelly. Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, 6(3):26, 2014.
- [224] Mark I. McCarthy, Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. Genome-wide association studies for complex traits:

- consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, May 2008.
- [225] Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42(2):109–142, 1980.
- [226] K. L. McGary, T. J. Park, J. O. Woods, H. J. Cha, J. B. Wallingford, and E. M. Marcotte. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences*, 107(14):6544–6549, Mar 2010.
- [227] Patrick T. McGrath, Matthew V. Rockman, Manuel Zimmer, Heeun Jang, Evan Z. Macosko, Leonid Kruglyak, and Cornelia I. Bargmann. Quantitative mapping of a digenic behavioral trait implicates globin variation in *c. elegans* sensory behaviors. *Neuron*, 61(5):692–699, Mar 2009.
- [228] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham. Deriving the consequences of genomic variants with the ensembl api and snp effect predictor. *Bioinformatics*, 26(16):2069–2070, Aug 2010.
- [229] Heather M McLaughlin, Ozge Ceyhan-Birsoy, Kurt D Christensen, Isaac S Kohane, Joel Krier, William J Lane, Denise Lautenbach, Matthew S Lebo, Kalotina Machini, and et al. A systematic approach to the reporting of medically relevant findings from whole genome sequencing. *BMC Medical Genetics*, 15(1), Dec 2014.
- [230] Gregor Mendel. Versuche über pflanzen-hybriden. *Verh. Naturforsch. Ver. Brünn*, 4:3–47, 1866.
- [231] Marcus J. Miller, Adam D. Kennedy, Andrea D. Eckhart, Lindsay C. Burrage, Jacob E. Wulff, Luke A.D. Miller, Michael V. Milburn, John A. Ryals, Arthur L. Beaudet, Qin Sun, and et al.

- Untargeted metabolomic analysis for the clinical screening of in-born errors of metabolism. *Journal of Inherited Metabolic Disease*, 38(6):1029–1039, Apr 2015.
- [232] Neil A. Miller, Emily G. Farrow, Margaret Gibson, Laurel K. Willig, Greyson Twist, Byunggil Yoo, Tyler Marrs, Shane Corder, Lisa Krivohlavek, Adam Walter, and et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Medicine*, 7(1), Sep 2015.
- [233] Eleni Mina, Mark Thompson, Rajaram Kaliyaperumal, Jun Zhao, van Eelke der Horst, Zuotian Tatum, Kristina M. Hettne, Erik A. Schultes, Barend Mons, and Marco Roos. Nanopublications for exposing experimental data in the life-sciences: a huntington's disease case study. *Journal of Biomedical Semantics*, 6(1):5, 2015.
- [234] E. V. Minikel, S. M. Vallabh, M. Lek, K. Estrada, K. E. Samocha, J. F. Sathirapongsasuti, C. Y. McLean, J. Y. Tung, L. P. C. Yu, P. Gambetti, and et al. Quantifying prion disease penetrance using large population control cohorts. *Science Translational Medicine*, 8(322):322ra9–322ra9, Jan 2016.
- [235] FuGE Functional Genomics Experiment model. Available: <http://fuge.sourceforge.net>. URL, 2010.
- [236] H. J. Muller Morgan, Thomas Hunt; Alfred H. Sturtevant and C. B. Bridges. The mechanism of mendelian heredity. *Henry Holt, New York*, 1915.
- [237] C. J. Mungall and D. B. Emmert. A chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, 23(13):i337–i346, Jul 2007.

- [238] Amanda J Myers, J Raphael Gibbs, Jennifer A Webster, Kristen Rohrer, Alice Zhao, Lauren Marlowe, Mona Kaleem, Doris Leung, Leslie Bryden, Priti Nath, and et al. A survey of genetic human cortical gene expression. *Nature Genetics*, 39(12):1494–1499, Nov 2007.
- [239] Preethy Sasidharan Nair and Mauno Vihinen. Varibench: A benchmark database for variations. *Human Mutation*, 34(1):42–49, Oct 2012.
- [240] P. Natarajan, N. B. Gold, A. G. Bick, H. McLaughlin, P. Kraft, H. L. Rehm, G. M. Peloso, J. G. Wilson, A. Correa, J. G. Seidman, and et al. Aggregate penetrance of genomic variants for actionable disorders in european and african americans. *Science Translational Medicine*, 8(364):364ra151–364ra151, Nov 2016.
- [241] Abhishek Niroula, Siddhaling Urolagin, and Mauno Vihinen. Pnp2: Prediction method for fast and reliable identification of harmful variants. *PLoS ONE*, 10(2):e0117380, Feb 2015.
- [242] Association of Clinical Genetics Netherlands. Available: <http://vkgn.org/vakinformatie/richtlijnen-en-protocollen/>. [accessed 15 march 2016]. *URL*, 2016.
- [243] The Genome of the Netherlands Consortium. The genome of the netherlands: design, and project goals. *European Journal of Human Genetics*, 22(2):221–227, Feb 2014.
- [244] The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the dutch population. *Nature Genetics*, Jun 2014.
- [245] K. Okonechnikov, O. Golosova, and M. Fursov. Unipro ugene: a unified bioinformatics toolkit. *Bioinformatics*, 28(8):1166–1167, Feb 2012.

BIBLIOGRAPHY

- [246] Omixed. Available: <http://www.omixed.org/>. URL, 2010.
- [247] Ruby on Rails. Available: <http://www.rubyonrails.org>. URL, 2010.
- [248] G. Ostlund, T. Schmitt, K. Forslund, T. Kostler, D. N. Messina, S. Roopra, O. Frings, and E. L. L. Sonnhammer. Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research*, 38(Database):D196–D203, Nov 2009.
- [249] Kristian Ovaska, Marko Laakso, Saija Haapa-Paananen, Riku Louhimo, Ping Chen, Viljami Aittomäki, Erkka Valo, Javier Núñez-Fontarnau, Ville Rantanen, Sirkku Karinen, and et al. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Medicine*, 2(9):65, 2010.
- [250] Brian D O'Connor, Allen Day, Scott Cain, Olivier Arnaiz, Linda Sperling, and Lincoln D Stein. Gmodweb: a web framework for the generic model organism database. *Genome Biology*, 9(6):R102, 2008.
- [251] Srivatsan Padmanabhan, Arnab Mukhopadhyay, Sri Devi Narasimhan, Gregory Tesz, Michael P. Czech, and Heidi A. Tissenbaum. A pp2a regulatory subunit regulates c. elegans insulin/igf-1 signaling by modulating akt-1 phosphorylation. *Cell*, 136(5):939–951, Mar 2009.
- [252] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stumpflen, H.-W. Mewes, and et al. The mips mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–834, Nov 2004.
- [253] Umadevi Paila, Brad A. Chapman, Rory Kirchner, and Aaron R. Quinlan. Gemini: Integrative exploration of genetic variation and genome annotations. *PLoS Computational Biology*, 9(7):e1003153, Jul 2013.

- [254] Michael F. Palopoli, Matthew V. Rockman, Aye TinMaung, Camden Ramsay, Stephen Curwen, Andrea Aduna, Jason Laurita, and Leonid Kruglyak. Molecular basis of the copulatory plug polymorphism in *caenorhabditis elegans*. *Nature*, 454(7207):1019–1022, Jul 2008.
- [255] C. Pang, D. Hendriksen, M. Dijkstra, K. J. van der Velde, J. Kuiper, H. Hillege, and M. Swertz. Biobankconnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. *Journal of the American Medical Informatics Association*, Oct 2014.
- [256] Chao Pang, Annet Sollie, Anna Sijtsma, Dennis Hendriksen, Bart Charbon, Mark de Haan, Tommy de Boer, Fleur Kelpin, Jonathan Jetten, Joeri K. van der Velde, and et al. Sorta: a system for ontology-based re-coding and technical annotation of biomedical phenotype data. *Database*, 2015:bav089, 2015.
- [257] Brent S. Pedersen, Ryan M. Layer, and Aaron R. Quinlan. Vc-fanno: fast, flexible annotation of genetic variants. *Genome Biology*, 17(1), Jun 2016.
- [258] Jaume Pellicer, Michael F. Fay, and Ilia J. Leitch. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, 164(1):10–15, Sep 2010.
- [259] K. Peng, W. Xu, J. Zheng, K. Huang, H. Wang, J. Tong, Z. Lin, J. Liu, W. Cheng, D. Fu, and et al. The disease and gene annotations (dga): an annotation resource for human disease. *Nucleic Acids Research*, 41(D1):D553–D560, Nov 2012.
- [260] Bjoern Peters, John Sidney, Phil Bourne, Huynh-Hoa Bui, Soeren Buus, Grace Doh, Ward Fleri, Mitch Kronenberg, Ralph Kubo, Ole Lund, and et al. The immune epitope database and analysis resource: From vision to blueprint. *PLoS Biology*, 3(3):e91, Mar 2005.

BIBLIOGRAPHY

- [261] Lauren A Peters, Jacqueline Perrigoue, Arthur Mortha, Alina Iuga, Won-min Song, Eric M Neiman, Sean R Llewellyn, Antonio Di Narzo, Brian A Kidd, Shannon E Telesco, and et al. A functional genomics predictive network model identifies regulators of inflammatory bowel disease. *Nature Genetics*, 49(10):1437–1449, Sep 2017.
- [262] Phillip H. Pham, William J. Shipman, Galina A. Erikson, Nicholas J. Schork, and Ali Torkamani. Scripps genome adviser: Annotation and distributed variant interpretation server. *PLOS ONE*, 10(2):e0116815, Feb 2015.
- [263] PAGE-OM The Phenotype and Genotype Object Model. Available: <http://www.pageom.org/>. URL, 2010.
- [264] Megan Phifer-Rixey and Michael W Nachman. Insights into mammalian biology from the wild house mouse *mus musculus*. *eLife*, 4, Apr 2015.
- [265] E M Phizicky and S Fields. Protein-protein interactions: methods for detection and analysis. *Microbiol Rev.*, 59:94–123, Mar 1995.
- [266] Hilikka Piirilä, Jouni Väliäho, and Mauno Vihinen. Immunodeficiency mutation databases (idbases). *Human Mutation*, 27(12):1200–1208, Dec 2006.
- [267] Eclipse Integrated Software Development platform. Available: <http://www.eclipse.org>. URL, 2010.
- [268] Sharon E. Plon, Diana M. Eccles, Douglas Easton, William D. Foulkes, Maurizio Genuardi, Marc S. Greenblatt, Frans B.L. Hogervorst, Nicoline Hoogerbrugge, Amanda B. Spurdle, and Sean V. Tavtigian. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Human Mutation*, 29(11):1282–1291, Nov 2008.

- [269] Martyn Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 2003.
- [270] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.
- [271] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, Oct 2009.
- [272] Nathan D Price, Andrew T Magis, John C Earls, Gustavo Glusman, Roie Levy, Christopher Lausted, Daniel T McDonald, Ulrike Kusebauch, Christopher L Moss, Yong Zhou, and et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nature Biotechnology*, 35(8):747–756, Jul 2017.
- [273] Pjotr Prins, Dominique Belhachemi, Steffen Möller, and Geert Smant. Scalable computing for evolutionary genomics. *Evolutionary Genomics*, page 529–545, 2012.
- [274] Pjotr Prins, Joep de Ligt, Artem Tarasov, Ritsert C Jansen, Edwin Cuppen, and Philip E Bourne. Toward effective software solutions for big biology. *Nature Biotechnology*, 33(7):686–687, Jul 2015.
- [275] The PubChem Project. Available: <http://pubchem.ncbi.nlm.nih.gov/>. URL, 2010.
- [276] S. B. Prusiner. Nobel lecture: Prions. *Proceedings of the National Academy of Sciences*, 95(23):13363–13383, Nov 1998.
- [277] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and et al.

BIBLIOGRAPHY

- Plink: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, Sep 2007.
- [278] D. Quang, Y. Chen, and X. Xie. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31(5):761–763, Oct 2014.
- [279] J Raes. Microbiome-based companion diagnostics: no longer science fiction? *Gut*, 65(6):896–897, Jan 2016.
- [280] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1):3, 2014.
- [281] T. Rausch, T. Zichner, A. Schlattl, A. M. Stutz, V. Benes, and J. O. Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, Sep 2012.
- [282] Tim F Rayner, Philippe Rocca-Serra, Paul T Spellman, Helen C Causton, Anna Farne, Ele Holloway, Rafael A Irizarry, Junmin Liu, Donald S Maier, Michael Miller, and et al. *BMC Bioinformatics*, 7(1):489, 2006.
- [283] K. C. Reddy, E. C. Andersen, L. Kruglyak, and D. H. Kim. A polymorphism in *npr-1* is a behavioral determinant of pathogen susceptibility in *c. elegans*. *Science*, 323(5912):382–384, Jan 2009.
- [284] Michael Reich, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo, and Jill P Mesirov. Genepattern 2.0. *Nature Genetics*, 38(5):500–501, May 2006.
- [285] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'95*,

- pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.
- [286] TAIR The Arabidopsis Information Resource. Available: <http://www.arabidopsis.org>. URL, 2010.
- [287] B. Reva, Y. Antipin, and C. Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research*, 39(17):e118–e118, Jul 2011.
- [288] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, and et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine*, 17(5):405–423, Mar 2015.
- [289] Graham R S Ritchie, Ian Dunham, Eleftheria Zeggini, and Paul Flicek. Functional annotation of noncoding sequence variants. *Nat Meth*, 11(3):294–296, Feb 2014.
- [290] Marylyn D. Ritchie, Emily R. Holzinger, Ruowang Li, Sarah A. Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, Jan 2015.
- [291] I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier. Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, Dec 2006.
- [292] PN Robinson and S Mundlos. The human phenotype ontology. *Clinical Genetics*, 77(6):525–534, Apr 2010.
- [293] M. V. Rockman, S. S. Skrovaneck, and L. Kruglyak. Selection at linked sites shapes heritable phenotypic variation in *c. elegans*. *Science*, 330(6002):372–376, Oct 2010.

BIBLIOGRAPHY

- [294] Miriam Rodriguez, L. Basten Snoek, Mario De Bono, and Jan E. Kammenga. Worms under stress: *C. elegans* stress response and its relevance to complex human disease and aging. *Trends in Genetics*, 29(6):367–374, Jun 2013.
- [295] Miriam Rodriguez, L. Basten Snoek, Joost A.G. Riksen, Roel P. Bevers, and Jan E. Kammenga. Genetic variation for stress-response hormesis in *c. elegans* lifespan. *Experimental Gerontology*, 47(8):581–587, Aug 2012.
- [296] Mark F. Rogers, Hashem A. Shihab, Matthew Mort, David N. Cooper, Tom R. Gaunt, and Colin Campbell. Fathmm-xf: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, Sep 2017.
- [297] Lisa Rosenbaum. Bridging the data-sharing divide — seeing the devil in the details, not the other camp. *New England Journal of Medicine*, Apr 2017.
- [298] G. Rustici, N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, J. Ison, M. Keays, and et al. Arrayexpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Research*, 41(D1):D987–D990, Nov 2012.
- [299] Lao H Saal, Carl Troein, Johan Vallon-Christersson, Sofia Gruberger, Åke Borg, and Carsten Peterson. *Genome Biology*, 3(8):software0003.1, 2002.
- [300] F. Sanger, S. Nicklen, and A.R. Coulson. Dna sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.*, 74(12):5463–5467, December 1977.
- [301] Susanna-Assunta Sansone, Philippe Rocca-Serra, Dawn Field, Eamonn Maguire, Chris Taylor, Oliver Hofmann, Hong Fang, Steffen Neumann, Weida Tong, Linda Amaral-Zettler, and

- et al. Toward interoperable bioscience data. *Nature Genetics*, 44(2):121–126, Jan 2012.
- [302] C. J. Saunders, N. A. Miller, S. E. Soden, D. L. Dinwiddie, A. Noll, N. A. Alnadi, N. Andraws, M. L. Patterson, L. A. Krivohlavek, J. Fellis, and et al. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Science Translational Medicine*, 4(154):154ra135–154ra135, Oct 2012.
- [303] Iris Schrijver, Nazneen Aziz, Daniel H. Farkas, Manohar Furtado, Andrea Ferreira Gonzalez, Timothy C. Greiner, Wayne W. Grody, Tina Hambuch, Lisa Kalman, Jeffrey A. Kant, and et al. Opportunities and challenges associated with clinical diagnostic genome sequencing. *The Journal of Molecular Diagnostics*, 14(6):525–540, Nov 2012.
- [304] Jana Marie Schwarz, David N Cooper, Markus Schuelke, and Dominik Seelow. Mutationtaster2: mutation prediction for the deep-sequencing age. *Nature Methods*, 11(4):361–362, Mar 2014.
- [305] U Seligsohn. High gene frequency of factor xi (pta) deficiency in ashkenazi jews. *Blood*, 51(6):1223–1228, 1978.
- [306] Rabah M. Shawky. Reduced penetrance in human inherited disease. *Egyptian Journal of Medical Human Genetics*, 15(2):103–111, Apr 2014.
- [307] Daniel D. Shaye and Iva Greenwald. Ortholist: A compendium of *c. elegans* genes with human orthologs. *PLoS ONE*, 6(5):e20085, May 2011.
- [308] H. A. Shihab, M. F. Rogers, J. Gough, M. Mort, D. N. Cooper, I. N. M. Day, T. R. Gaunt, and C. Campbell. An integrative approach to predicting the functional effects of non-coding and

BIBLIOGRAPHY

- coding sequence variation. *Bioinformatics*, 31(10):1536–1543, Jan 2015.
- [309] Alejandro Sifrim, Dusan Popovic, Leon-Charles Tranchevent, Amin Ardeshtirdavani, Ryo Sakai, Peter Konings, Joris R Vermeesch, Jan Aerts, Bart De Moor, and Yves Moreau. ecstasy: variant prioritization by genomic data fusion. *Nature Methods*, 10(11):1083–1084, Sep 2013.
- [310] D. Smedley, A. Oellrich, S. Kohler, B. Ruef, M. Westerfield, P. Robinson, S. Lewis, and C. Mungall. Phenodigm: analyzing curated annotations to associate animal models with human diseases. *Database*, 2013(0):bat025–bat025, May 2013.
- [311] D. Smedley, M. A. Swertz, K. Wolstencroft, G. Proctor, M. Zuberakis, J. Bard, J. M. Hancock, and P. Schofield. Solutions for data integration in functional genomics: a critical assessment and case study. *Briefings in Bioinformatics*, 9(6):532–544, Jul 2008.
- [312] Damian Smedley, Syed Haider, Benoit Ballester, Richard Holland, Darin London, Gudmundur Thorisson, and Arek Kasprzyk. Biomart – biological queries made easy. *BMC Genomics*, 10(1):22, 2009.
- [313] Damian Smedley, Max Schubach, Julius O.B. Jacobsen, Sebastian Köhler, Tomasz Zemojtel, Malte Spielmann, Marten Jäger, Harry Hochheiser, Nicole L. Washington, Julie A. McMurry, and et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. *The American Journal of Human Genetics*, 99(3):595–606, Sep 2016.
- [314] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, and et al. The obo

- foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, Nov 2007.
- [315] L. B. Snoek, K. J. Van der Velde, D. Arends, Y. Li, A. Beyer, M. Elvin, J. Fisher, A. Hajnal, M. O. Hengartner, G. B. Poulin, and et al. Wormqtl—public archive and analysis web portal for natural variation data in caenorhabditis spp. *Nucleic Acids Research*, 41(D1):D738–D743, Nov 2012.
- [316] L. Basten Snoek, Inez R. Terpstra, René Dekter, Guido Van den Ackerveken, and Anton J. M. Peeters. Genetical genomics reveals large scale genotype-by-environment interactions in arabidopsis thaliana. *Frontiers in Genetics*, 3, 2013.
- [317] Nara Sobreira, François Schiettecatte, David Valle, and Ada Hamosh. Genematcher: A matching tool for connecting investigators with an interest in the same gene. *Human Mutation*, 36(10):928–930, Aug 2015.
- [318] Rachel Soemedi, Kamil J Cygan, Christy L Rhine, Jing Wang, Charlston Bulacan, John Yang, Pinar Bayrak-Toydemir, Jamie McDonald, and William G Fairbrother. Pathogenic variants that alter protein code often disrupt splicing. *Nature Genetics*, 49(6):848–855, Apr 2017.
- [319] B. D. Solomon, A.-D. Nguyen, K. A. Bear, and T. G. Wolfsberg. Clinical genomic database. *Proceedings of the National Academy of Sciences*, 110(24):9851–9855, May 2013.
- [320] Wei Song, Sabrina A. Gardner, Hayk Hovhannisyan, Amanda Natalizio, Katelyn S. Weymouth, Wenjie Chen, Ildiko Thibodeau, Ekaterina Bogdanova, Stanley Letovsky, Alecia Willis, and et al. Exploring the landscape of pathogenic genetic variation in the exac population database: insights of relevance to variant classification. *Genetics in Medicine*, 18(8):850–854, Dec 2015.

BIBLIOGRAPHY

- [321] L. D. Stein. The generic genome browser: A building block for a model organism system database. *Genome Research*, 12(10):1599–1610, Oct 2002.
- [322] Lincoln D. Stein. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Reviews Genetics*, 9(9):678–688, Sep 2008.
- [323] Peter D. Stenson, Matthew Mort, Edward V. Ball, Katy Shaw, Andrew D. Phillips, and David N. Cooper. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, 133(1):1–9, Sep 2013.
- [324] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, and et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, Feb 2007.
- [325] Barbara E Stranger, Alexandra C Nica, Matthew S Forrest, Antigone Dimas, Christine P Bird, Claude Beazley, Catherine E Ingle, Mark Dunning, Paul Flicek, Daphne Koller, and et al. Population genomics of human gene expression. *Nature Genetics*, 39(10):1217–1224, Sep 2007.
- [326] Charles M Strom, Beryl Crossley, Arlene Buller-Buerkle, Michael Jarvis, Franklin Quan, Mei Peng, Kasinathan Muralidharan, Victoria Pratt, Joy B Redman, and Weimin Sun. Cystic fibrosis testing 8 years on: Lessons learned from carrier screening and sequencing analysis. *Genetics in Medicine*, 13(2):166–172, Jan 2011.
- [327] M. A. Swertz, E. O. de Brock, S. A. F. T. van Hijum, A. de Jong, G. Buist, R. J. S. Baerends, J. Kok, O. P. Kuipers, and R. C.

- Jansen. Molecular genetics information system (molgenis): alternatives in developing local experimental genomics databases. *Bioinformatics*, 20(13):2075–2083, Apr 2004.
- [328] Morris A Swertz, Martijn Dijkstra, Tomasz Adamusiak, Joeri K van der Velde, Alexandros Kanterakis, Erik T Roos, Joris Lops, Gudmundur A Thorisson, Danny Arends, George Byelas, and et al. The molgenis toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics*, 11(Suppl 12):S12, 2010.
- [329] Morris A. Swertz and Ritsert C. Jansen. Beyond standardization: dynamic software infrastructures for systems biology. *Nature Reviews Genetics*, 8(3):235–243, Feb 2007.
- [330] Morris A Swertz, K Joeri Velde, Bruno M Tesson, Richard A Scheltema, Danny Arends, Gonzalo Vera, Rudi Alberts, Martijn Dijkstra, Paul Schofield, Klaus Schughart, and et al. Xgap: a uniform and extensible data model and software platform for genotype and phenotype experiments. *Genome Biology*, 11(3):R27, 2010.
- [331] Shigeki Taniguchi, Toshihide Kimura, Tatsuhito Umeki, Yuka Kimura, Hideo Kimura, Isao Ishii, Norimichi Itoh, Yasuhito Naito, Hideyuki Yamamoto, and Ichiro Niki. Protein phosphorylation involved in the gene expression of the hydrogen sulphide producing enzyme cystathionine gamma-lyase in the pancreatic beta-cell. *Molecular and Cellular Endocrinology*, 350(1):31–38, Mar 2012.
- [332] Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of ngs alignment formats. *Bioinformatics*, 31(12):2032–2034, Feb 2015.
- [333] S V Tavtigian. Comprehensive statistical study of 452 brca1 missense substitutions with classification of eight recurrent substi-

BIBLIOGRAPHY

- tutions as neutral. *Journal of Medical Genetics*, 43(4):295–305, Sep 2005.
- [334] Sean V. Tavtigian, Marc S. Greenblatt, Fabienne Lesueur, and Graham B. Byrnes. In silico analysis of missense substitutions using sequence-alignment based methods. *Human Mutation*, 29(11):1327–1336, Nov 2008.
- [335] Chris F Taylor, Dawn Field, Susanna-Assunta Sansone, Jan Aerts, Rolf Apweiler, Michael Ashburner, Catherine A Ball, Pierre-Alain Binz, Molly Bogue, Tim Booth, and et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the mibbi project. *Nature Biotechnology*, 26(8):889–896, Aug 2008.
- [336] I. R. Terpstra, L. B. Snoek, J. J. B. Keurentjes, A. J. M. Peeters, and G. Van den Ackerveken. Regulatory network identification by genetical genomics: Signaling downstream of the arabidopsis receptor-like kinase erecta. *PLANT PHYSIOLOGY*, 154(3):1067–1078, Sep 2010.
- [337] Bryony A. Thompson, Marc S. Greenblatt, Maxime P. Vallee, Johanna C. Herkert, Chloe Tessereau, Erin L. Young, Ivan A. Adzhubey, Biao Li, Russell Bell, Bingjian Feng, and et al. Calibration of multiple in silico tools for predicting pathogenicity of mismatch repair gene missense substitutions. *Human Mutation*, 34(1):255–265, Jan 2013.
- [338] Bryony A Thompson, Amanda B Spurdle, John-Paul Plazzer, Marc S Greenblatt, Kiwamu Akagi, Fahd Al-Mulla, Bharati Bapat, Inge Bernstein, Gabriel Capellá, Johan T den Dunnen, and et al. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the insight locus-specific database. *Nature Genetics*, 46(2):107–115, Dec 2013.

- [339] G. A. Thorisson, O. Lancaster, R. C. Free, R. K. Hastings, P. Sarmah, D. Dash, S. K. Brahmachari, and A. J. Brookes. Hgibase2p: a central genetic association database. *Nucleic Acids Research*, 37(Database):D797–D802, Jan 2009.
- [340] Gudmundur A. Thorisson, Juha Muiilu, and Anthony J. Brookes. Genotype–phenotype databases: challenges and solutions for the post-genomic era. *Nature Reviews Genetics*, 10(1):9–18, Jan 2009.
- [341] H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, Apr 2012.
- [342] Oswaldo Trelles, Pjotr Prins, Marc Snir, and Ritsert C. Jansen. Big data, but are we ready? *Nature Reviews Genetics*, 12(3):224–224, Feb 2011.
- [343] Peter C. van den Akker, Marcel F. Jonkman, Trebor Rengaw, Leena Bruckner-Tuderman, Cristina Has, Johann W. Bauer, Alfred Klausegger, Giovanna Zambruno, Daniele Castiglia, Jemima E. Mellerio, and et al. The international dystrophic epidermolysis bullosa patient registry: An online database of dystrophic epidermolysis bullosa patients and their col7a1 mutations. *Human Mutation*, 32(10):1100–1107, Sep 2011.
- [344] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, and et al. From fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, page 11.10.1–11.10.33, Oct 2013.
- [345] K. Joeri van der Velde, Eddy N. de Boer, Cleo C. van Diemen, Birgit Sikkema-Raddatz, Kristin M. Abbott, Alain Knoppers, Lude

BIBLIOGRAPHY

- Franke, Rolf H. Sijmons, Tom J. de Koning, Cisca Wijmenga, and et al. Gavin: Gene-aware variant interpretation for medical sequencing. *Genome Biology*, 18(1), Jan 2017.
- [346] K. Joeri van der Velde, Mark de Haan, Konrad Zych, Danny Arends, L. Basten Snoek, Jan E. Kammenga, Ritsert C. Jansen, Morris A. Swertz, and Yang Li. Wormqtlhd—a web database for linking human disease to natural variation data in *c. elegans*. *Nucleic Acids Research*, 42(D1):D794–D801, Nov 2013.
- [347] K. Joeri van der Velde, Joël Kuiper, Bryony A. Thompson, John-Paul Plazzer, Gert van Valkenhoef, Mark de Haan, Jan D.H. Jongbloed, Cisca Wijmenga, Tom J. de Koning, Kristin M. Abbott, and et al. Evaluation of cadd scores in curated mismatch repair gene variants yields a model for clinical validation and prioritization. *Human Mutation*, 36(7):712–719, May 2015.
- [348] Cleo C. van Diemen, Wilhelmina S. Kerstjens-Frederikse, Klasien A. Bergman, Tom J. de Koning, Birgit Sikkema-Raddatz, Joeri K. van der Velde, Kristin M. Abbott, Johanna C. Herkert, Katharina Löhner, Patrick Rump, and et al. Rapid targeted genomics in critically ill newborns. *Pediatrics*, page e20162854, Sep 2017.
- [349] C. Lee. Ventola. Pharmacogenomics in clinical practice: Reality and expectations. *Pharmacy and Therapeutics*, 36(7):412–450, 2011.
- [350] A. Vinuela, L. B. Snoek, J. A. G. Riksen, and J. E. Kammenga. Genome-wide gene expression regulation as a function of genotype and age in *c. elegans*. *Genome Research*, 20(7):929–937, May 2010.
- [351] A. Vinuela, L. B. Snoek, J. A. G. Riksen, and J. E. Kammenga. Aging uncouples heritability and expression-qt1 in *caenorhabditis*

- elegans. *G3&no.58; Genes—Genomes—Genetics*, 2(5):597–605, May 2012.
- [352] Rita JM Volkens, L Snoek, Caspara J van Hellenberg Hubar, Renata Coopman, Wei Chen, Wentao Yang, Mark G Sterken, Hinrich Schulenburg, Bart P Braeckman, and Jan E Kammenga. Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild caenorhabditis elegans populations. *BMC Biol*, 11(1):93, 2013.
- [353] Yvonne Wallis, Stewart Payne, Ciaron McAnulty, Danielle Bodmer, Erik Sistermans, Kathryn Robertson, David Moore, Stephen Abbs, Zandra Deans, and Andrew Devereau. Practice guidelines for the evaluation of pathogenicity and the reporting of sequence variants in clinical molecular genetics. *Association for Clinical Genetic Science and the Dutch Society of Clinical Genetic Laboratory Specialists*, 2013.
- [354] Roddy Walsh, Kate L. Thomson, James S. Ware, Birgit H. Funke, Jessica Woodley, Karen J. McGuire, Francesco Mazzarotto, Edward Blair, Anneke Seller, Jenny C. Taylor, and et al. Reassessment of mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genetics in Medicine*, Aug 2016.
- [355] Klaudia Walter, Josine L. Min, Jie Huang, Lucy Crooks, Yasin Memari, Shane McCarthy, John R. B. Perry, ChangJiang Xu, Marta Futema, Daniel Lawson, and et al. The uk10k project identifies rare variants in health and disease. *Nature*, 526(7571):82–90, Sep 2015.
- [356] Jason Wang, Garrett Gotway, Juan M. Pascual, and Jason Y. Park. Diagnostic yield of clinical next-generation sequencing panels for epilepsy. *JAMA Neurol*, 71(5):650, May 2014.

BIBLIOGRAPHY

- [357] Jinlian Wang, Jun Liao, Jinglan Zhang, Wei-Yi Cheng, Jörg Hakenberg, Meng Ma, Bryn D. Webb, Rajasekar Ramasamudram-chakravarthi, Lisa Karger, Lakshmi Mehta, and et al. Clinlabgeneticist: a tool for clinical management of genetic variants from whole exome sequencing in clinical genetic laboratories. *Genome Medicine*, 7(1), Jul 2015.
- [358] Jintao Wang, Robert W. Williams, and Kenneth F. Manly. Webqtl: Web-based complex trait analysis. *Neuroinformatics*, 1(4):299–308, 2003.
- [359] Weiqi Wang and Eswar Krishnan. Big data and clinicians: A review on the state of the science. *JMIR Medical Informatics*, 2(1):e1, Jan 2014.
- [360] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, and et al. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(Web Server):W214–W220, Jun 2010.
- [361] J. D. WATSON and F. H. C. CRICK. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.
- [362] Marjan M. Weiss, Bert Van der Zwaag, Jan D. H. Jongbloed, Maartje J. Vogel, Hennie T. Brüggewirth, Ronald H. Lekanne Deprez, Olaf Mook, Claudia A. L. Ruivenkamp, Marjon A. van Slegtenhorst, Arthur van den Wijngaard, and et al. Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: A national collaborative study of dutch genome diagnostic laboratories. *Human Mutation*, 34(10):1313–1321, Aug 2013.
- [363] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and

- et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006, Dec 2013.
- [364] Harm-Jan Westra, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, and et al. Systematic identification of trans eqtls as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, Sep 2013.
- [365] Amy B Wilfert, Katherine R Chao, Madhurima Kaushal, Sanjay Jain, Sebastian Zöllner, David R Adams, and Donald F Conrad. Genome-wide significance testing of variation from single case exomes. *Nature Genetics*, Oct 2016.
- [366] Mark D. Wilkinson, Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, and et al. The fair guiding principles for scientific data management and stewardship. *Sci. Data*, 3:160018, Mar 2016.
- [367] Mark D Wilkinson, Luke McCarthy, Benjamin Vandervalk, David Withers, Edward Kawas, and Soroush Samadian. Sadi, share, and the in silico scientific method. *BMC Bioinformatics*, 11(Suppl 12):S7, 2010.
- [368] Mark D. Wilkinson, Ruben Verborgh, Luiz Olavo Bonino da Silva Santos, Tim Clark, Morris A. Swertz, Fleur D.L. Kelpin, Alasdair J.G. Gray, Erik A. Schultes, Erik M. van Mulligen, Paolo Ciccarese, and et al. Interoperability and fairness through a novel combination of web technologies. *PeerJ Computer Science*, 3:e110, Apr 2017.
- [369] B. J. Willcox, T. A. Donlon, Q. He, R. Chen, J. S. Grove, K. Yano, K. H. Masaki, D. C. Willcox, B. Rodriguez, and

BIBLIOGRAPHY

- J. D. Curb. Foxo3a genotype is strongly associated with human longevity. *Proceedings of the National Academy of Sciences*, 105(37):13987–13992, Sep 2008.
- [370] David Withers, Edward Kawas, Luke McCarthy, Benjamin Vandervalk, and Mark Wilkinson. Semantically-guided workflow construction in taverna: The sadi and biomoby plug-ins. *Leveraging Applications of Formal Methods, Verification, and Validation*, page 301–312, 2010.
- [371] Josefine S Witteveen, Marjolein H Willemsen, Thaís C D Dombroski, Nick H M van Bakel, Willy M Nillesen, Josephus A van Hulten, Eric J R Jansen, Dave Verkaik, Hermine E Veenstra-Knol, Conny M A van Ravenswaaij-Arts, and et al. Haploinsufficiency of mecp2-interacting transcriptional co-repressor sin3a causes mild intellectual disability by affecting the development of cortical integrity. *Nature Genetics*, 48(8):877–887, Jul 2016.
- [372] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, and et al. The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. *Nucleic Acids Research*, 41(W1):W557–W561, May 2013.
- [373] Katherine Wolstencroft, Olga Krebs, Jacky L. Snoep, Natalie J. Stanford, Finn Bacall, Martin Golebiewski, Rostyk Kuzyakiv, Quyen Nguyen, Stuart Owen, Stian Soiland-Reyes, and et al. Fairdomhub: a repository and collaboration environment for sharing systems biology research. *Nucleic Acids Research*, 45(D1):D404–D407, Nov 2016.
- [374] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, and et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11):1173–1186, Oct 2014.

- [375] Taverna Workbench. Available: <http://taverna.sourceforge.net>. URL, 2010.
- [376] B. S. Yandell, T. Mehta, S. Banerjee, D. Shriner, R. Venkataraman, J. Y. Moon, W. W. Neely, H. Wu, R. von Smith, and N. Yi. R/qtlbim: Qtl with bayesian interval mapping in experimental crosses. *Bioinformatics*, 23(5):641–643, Jan 2007.
- [377] J. Yang, C. Hu, H. Hu, R. Yu, Z. Xia, X. Ye, and J. Zhu. Qtlnetwork: mapping and visualizing genetic architecture of complex traits in experimental populations. *Bioinformatics*, 24(5):721–723, Jan 2008.
- [378] Yaping Yang, Donna M. Muzny, Jeffrey G. Reid, Matthew N. Bainbridge, Alecia Willis, Patricia A. Ward, Alicia Braxton, Joke Beuten, Fan Xia, Zhiyv Niu, and et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med*, 369(16):1502–1511, Oct 2013.
- [379] K. Yook, T. W. Harris, T. Bieri, A. Cabunoc, J. Chan, W. J. Chen, P. Davis, N. de la Cruz, A. Duong, R. Fang, and et al. Wormbase 2012: more genomes, more data, new website. *Nucleic Acids Research*, 40(D1):D735–D741, Nov 2011.
- [380] Dèlia Yubero, Núria Brandi, Aida Ormazabal, Àngels Garcia-Cazorla, Belén Pérez-Dueñas, Jaime Campistol, Antonia Ribes, Francesc Palau, Rafael Artuch, and Judith Armstrong. Targeted next generation sequencing in patients with inborn errors of metabolism. *PLoS ONE*, 11(5):e0156359, May 2016.
- [381] Javad Zahiri, Joseph Bozorgmehr, and Ali Masoudi-Nejad. Computational prediction of protein–protein interaction networks: Algorithms and resources. *Current Genomics*, 14(6):397–414, Sep 2013.

BIBLIOGRAPHY

- [382] Andrea Zatkova. An update on molecular genetics of alkaptonuria (aku). *Journal of Inherited Metabolic Disease*, 34(6):1127–1136, Jul 2011.
- [383] Barry R Zeeberg, Joseph Riss, David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Linehan, J Carl Barrett, and John N Weinstein. *BMC Bioinformatics*, 5(1):80, 2004.
- [384] H. Zeng, L. Luo, W. Zhang, J. Zhou, Z. Li, H. Liu, T. Zhu, X. Feng, and Y. Zhong. Plantqtl-ge: a database system for identifying candidate genes in rice and arabidopsis by gene expression and qtl information. *Nucleic Acids Research*, 35(Database):D879–D882, Jan 2007.
- [385] Feng Zhang and James R. Lupski. Non-coding genetic variants in human disease: Figure 1. *Human Molecular Genetics*, 24(R1):R102–R110, Jul 2015.
- [386] Hongyi Zhou, Mu Gao, and Jeffrey Skolnick. Entprise: An algorithm for predicting human disease-associated amino acid substitutions from sequence entropy and predicted protein structures. *PLOS ONE*, 11(3):e0150965, Mar 2016.
- [387] Jian Zhou and Olga G Troyanskaya. Predicting effects of non-coding variants with deep learning-based sequence model. *Nat Meth*, 12(10):931–934, Aug 2015.
- [388] Y. Zhou, Y. Liang, K. H. Lynch, J. J. Dennis, and D. S. Wishart. Phast: A fast phage search tool. *Nucleic Acids Research*, 39(suppl):W347–W352, Jun 2011.
- [389] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, and et al. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nat Genet*, 48(5):481–487, Mar 2016.

- [390] Mark Ziemann, Yotam Eren, and Assam El-Osta. Gene name errors are widespread in the scientific literature. *Genome Biology*, 17(1), Aug 2016.

BIBLIOGRAPHY

List of Tables

1.1	Glossary of key terms, pt. 1/2.	16
1.2	Glossary of key terms, pt. 2/2.	17
2.1	Features of XGAP database	40
2.2	Use cases of core data types.	43
2.3	Use cases of extended data types	45
2.4	Use cases of annotation data types	47
2.5	Use cases of the graphical user interface	54
2.6	Use cases of the application programming interface	55
2.7	XGAP participating consortia	61
4.1	Top 15 results for the disease enrichment	90
4.2	Result hits 16-58 for the disease enrichment	91
4.3	Result hits 59-100 for the disease enrichment	92
5.1	Number of InSiGHT variants reassigned	107
5.2	Explanations according to InSiGHT, pt. 1/2	108

LIST OF TABLES

5.3	Explanations according to InSiGHT, pt. 2/2	109
5.4	Variants of class 2 for which class 5 is the predicted . . .	111
5.5	The 24 variants predicted to be likely pathogenic	127
6.1	Origins of the benchmark data sets used	135
6.2	Stratification of data set into manifestations	136
6.3	Performance overview of all tested tools.	137
6.4	Estimate of impact in clinical diagnostics	142
6.5	Tools used to evaluate our benchmark variant set	151
7.1	rVCF format specification	165
7.2	Examples of VCF INFO field definitions	169
7.3	Variants that were missed by the GAVIN+	175
8.1	ClinVar review status and star rating	216
8.2	Brief review of omics data types	233
F.1	Other academic activities, pt. 1/2	324
F.2	Other academic activities, pt. 2/2	325

List of Figures

- 1.1 Translational science overview 23
- 1.2 Overview of thesis chapter progression 32

- 2.1 Extensible genotype and phenotype object model 42
- 2.2 Simple text file format 50
- 2.3 Graphical User Interfaces 52
- 2.4 Application programming interfaces 56
- 2.5 Customizing XGAP 59
- 2.6 Auto-generation of XGAP software 65

- 3.1 Screenshot of xQTL workbench with all features 71

- 4.1 Human and worm data integration 80
- 4.2 Cross-experiment search 82
- 4.3 WormQTL^{HD} poster 96

- 5.1 Probability that a CADD score belongs to a class 105
- 5.2 Flowchart describing the analysis 120

LIST OF FIGURES

5.3	Beanplot showing CADD score data and density	121
5.4	CADD scaled C-scores for MLH1 gene	122
5.5	CADD scaled C-scores for MSH2 gene	123
5.6	CADD scaled C-scores for MSH6 gene	124
5.7	CADD scaled C-scores for PMS2 gene	125
5.8	SnPEff effect vs. CADD score	126
6.1	Performance of GAVIN and other tools	138
6.2	Comparison of gene-specific thresholds	140
7.1	Overview of the framework for automation	162
7.2	Applications of the rVCF format	164
7.3	Example screenshot of a patient report	171
7.4	GAVIN+ false omission on pathogenic variants	173
7.5	Estimations of GAVIN+ false discovery rate	177
8.1	The MOLGENIS family of software	199
8.2	Graph of the Human Phenotype Ontology	203
8.3	Relation between calibration and number variants	215
8.4	Relation between calibration and review quality	217
8.5	GAVIN (r0.3) calibration plot for SCN5A.	219
8.6	GAVIN (r0.3) calibration plot for TTN.	220
8.7	GAVIN (r0.3) calibration plot for F11.	222
8.8	Interplay between data, tools and protocols	236

Appendices

Appendix A

Summary

All organisms have a genome made of DNA (deoxyribonucleic acid). The genome can be found in nearly every cell and is the blueprint for the growth, development, maintenance and repair of the body. It performs these functions by transcribing small pieces of DNA, the genes, from the genome and translating them to proteins. These proteins are the tiny workhorses of the body that break down food, give bones their strength, make muscles move, let brains think, and so on. There are many thousands of different genes and proteins with each their own task.

The genome is copied from cell to cell, and is inherited from generation to generation. The copying process is incredibly precise, but always makes a few little mistakes. These so-called mutations cause small differences between individuals, so that natural selection and thus evolution can take place. Unfortunately, mutations can also cause detrimental effects, such as genetic disorders. When the function of a gene

is disrupted by a mutation, a specific disorder can arise. For a lot of genes it is known which disorder they can cause, but for most genes we do not know what happens when they are disrupted.

In this thesis we research and develop various bioinformatics models, methods and systems to elucidate which genes and DNA differences can make people ill. To support the research into genes, we develop a database in chapter 2. This database is useful for collecting all kinds of biological data. Chapter 3 presents software to analyze these data as an extension to this database. This software can determine which region of the genome is responsible for diseases and other physical traits.

Organisms such as rat, worm or zebrafish allow us to perform research that would be impractical or unethical on humans. These studies deliver valuable biological insights, but it often remains unclear how those insights can help us to understand human disorders. In Chapter 4 we report the development of an interactive database that connects research into worms to the genetics of human disease. Despite the fact that worms do not look much like humans, they have thousands of genes that work exactly the same way as in humans. By looking at disorders, physical traits and genes in both organisms we discover new ways to use worms for research into human diseases.

Next to understanding the genes lies the challenge to determine the harmfulness (pathogenicity) of new mutations. Each individual carries many unique mutations no one else has. This makes it challenging to find the causal mutation for a patient with a genetic disorder. We can use our knowledge of the genome and evolution to predict how pathogenic new mutations are. In chapter 5 we report a smart new method which predict which mutations are harmless and which cause hereditary colon cancer. This works rather well and we make recommendations for the guideline that establishes diagnoses.

Encouraged by these results we expanded our scope from just a few to thousands of disease genes in chapter 6. For this, we use DNA from individuals that have no connection to severe disorders. We compared that to disease causing mutations that were found in patients. By

crunching the numbers for every gene, it is determined when a mutation is probably disease causing. The final result is a public website where the DNA of patients can be scanned quickly and accurately for probable pathogenic mutations.

Finally, chapter 7 describes how we developed a system for automated DNA analysis, including a protocol specific for genome diagnostics. This protocol uses our new method but also the latest knowledge on mutations and genes. Pathogenic mutations are not always responsible for the disease of a patient. That is why the DNA of family members is used to determine if the genetic pieces of the patient truly fit. The output is a new file format in which medically relevant information is formally expressed. This file can be converted to a clear report in which the most important information is found at the top.

One of the advantages is that we can apply this analysis without manual work to the genomes of thousands of healthy people. The results act as a control that tells us how often the software returns an accidental hit in each gene. By stating this information in the final report, medical experts can focus their attention on genes with the fewest accidental hits. This increases the speed and confidence at which a genetic diagnosis is established for the patient.

Appendix B

Samenvatting

Alle organismen hebben een genoom dat is opgebouwd uit DNA (desoxyribonucleïnezuur). Het genoom zit in bijna elke cel en is de blauwdruk voor de groei, ontwikkeling, onderhoud en herstel van het lichaam. Het vervult deze functies door kleine stukjes DNA, de genen, van het genoom af te schrijven en deze te vertalen naar eiwitten. Deze eiwitten zijn de werkpaardjes van het lichaam die voedsel afbreken, botten hun sterkte geven, spieren laten bewegen, hersenen laten denken, enzovoort. Er zijn vele duizenden verschillende genen en eiwitten met allemaal hun eigen taak.

Het genoom wordt gekopieerd van cel naar cel, en wordt overgeërfd van generatie op generatie. Het kopieerproces is ongelooflijk precies, maar maakt altijd wel een paar foutjes. Deze zogeheten mutaties zorgen voor kleine verschillen tussen individuen, waardoor natuurlijke selectie en dus evolutie kan plaatsvinden. Helaas kunnen mutaties ook nadelige effecten veroorzaken, zoals erfelijke ziektes. Wanneer de werk-

ing van een gen verstoord wordt door een mutatie kan een bepaalde ziekte optreden. Van een hoop genen is bekend welke ziekte ze kunnen veroorzaken, maar van de meeste genen weten we niet wat er gebeurt wanneer ze verstoord worden.

In dit proefschrift onderzoeken en ontwikkelen we verscheidene bioinformatica modellen, methoden en systemen om op te helderen welke genen en DNA verschillen mensen ziek kunnen maken. Om het onderzoek naar genen te ondersteunen, ontwikkelen we een database in hoofdstuk 2. Deze database is handig voor het verzamelen van allerlei biologische gegevens. Als uitbreiding op deze database presenteert hoofdstuk 3 software voor de analyse van deze gegevens. Deze software kan bepalen welk gebied van het genoom verantwoordelijk is voor ziektes en andere uiterlijke kenmerken.

Dieren zoals rat, worm en zebrafis stellen ons in staat om onderzoek te doen die onpraktisch of onethisch zou zijn op mensen. Deze onderzoeken leveren waardevolle biologische inzichten op, maar het blijft vaak onduidelijk hoe die inzichten ons kunnen helpen om ziektes in de mens te begrijpen. In hoofdstuk 4 rapporteren we de ontwikkeling van een interactieve database die het onderzoek naar wormen verbindt aan de genetica van menselijke ziektes. Ondanks het feit dat wormen niet echt op mensen lijken, hebben zij duizenden genen die precies hetzelfde werken als bij mensen. Door te kijken naar ziektes, uiterlijke kenmerken en genen in beide organismen ontdekken we nieuwe manieren om wormen te gebruiken voor onderzoek naar menselijke ziektes.

Naast het begrijpen van de genen ligt de uitdaging om de schadelijkheid (pathogeniteit) van nieuwe mutaties te bepalen. Ieder individu draagt vele unieke mutaties die niemand anders heeft. Dit maakt het een uitdaging om de schuldige mutatie te vinden bij een patient met een genetische ziekte. We kunnen onze kennis van het genoom en de evolutie inzetten om te voorspellen hoe pathogeen nieuwe mutaties zijn. In hoofdstuk 5 rapporteren we een slimme nieuwe methode die voorspelt welke mutaties ongevaarlijk zijn en welke erfelijke darmkanker veroorzaken. Dit blijkt vrij goed te kunnen en we doen aanbevelingen voor de

richtlijn die diagnoses stelt.

Aangemoedigd door deze resultaten zijn we onze speelruimte gaan uitbreiden van slechts een paar tot wel duizenden ziektegenen in hoofdstuk 6. Hiervoor gebruiken we DNA van mensen die geen verband hebben met ernstige ziektes. Dat vergelijken we met ziekteverwekkende mutaties die bij patienten gevonden zijn. Met het nodige rekenwerk voor ieder gen wordt bepaald wanneer een mutatie waarschijnlijk ziekteverwekkend is. Het resultaat is een openbare website waar het DNA van patienten snel en accuraat gescand kan worden op mogelijk pathogene mutaties.

Tenslotte beschrijft hoofdstuk 7 hoe we een systeem voor geautomatiseerde DNA analyse ontwikkeld hebben, inclusief een protocol specifiek voor genoom diagnostiek. Dit protocol gebruikt onze nieuwe methode maar ook de laatste kennis over mutaties en genen. Pathogene mutaties zijn niet altijd verantwoordelijk voor de ziekte van een patient. Daarom wordt het DNA van familieleden gebruikt om te bepalen of het genetische plaatje bij de patient ook echt klopt. De uitvoer is een nieuw bestandsformaat waarin medisch relevante informatie formeel wordt uitgedrukt. Dit bestand kan worden omgezet naar een overzichtelijk rapport waarin de meest belangrijke informatie bovenaan staat.

Eén van de voordelen is dat we deze analyse zonder handwerk kunnen toepassen op de genomen van duizenden gezonde mensen. De uitkomst hiervan dient als controle die ons verteld hoe vaak de software een toevalstreffer heeft in elk gen. Door deze informatie in het uiteindelijke rapport te vermelden kunnen medisch experts hun aandacht richten op genen met de minste toevalstreffers. Hiermee wordt de snelheid en zekerheid waarmee een genetische diagnose bij de patient wordt vastgesteld, verhoogd.

Appendix C

Acknowledgements

APPENDIX C. ACKNOWLEDGEMENTS

LOCUS ACKNWLDGMNTS 1380 bp DNA linear KJV 15-NOV-2017
DEFINITION Dear friends: living beings are the result of a complex network of thousands of interacting genes. Likewise, a PhD thesis is the result of the support and collaboration of many interacting individuals - you! Truly, you may consider yourself the DNA of this thesis. I would like to sincerely thank each and every one of you for your kindness, brilliance and great times that allowed me to complete this book. Dank jullie wel!

ACCESSION DNKWRD2017
VERSION DNKWRD2017.1
KEYWORDS Thanks; bedankt; gracias; merci; danke.
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 1380)
AUTHORS Van der Velde, K.J.
TITLE Acknowledgements
JOURNAL Translational software infrastructure for medical genetics
COMMENT Special thanks goes out to my promotor Prof. Morris Swertz, for providing me with many interesting projects, fruitful collaborations, and opportunities to try out all kinds of ideas. Dear Morris, your energy, ambition and always positive "can do" attitude has been a real driver and great inspiration. Cheers! I would also like to express my gratitude to my other promotores, Prof. Richard Sinke and Assistant Prof. Yang Li, as well as Prof. Cisca Wijmenga and Prof. Ritsert Jansen. Your clever insights, constructive feedback and incredible commitment has taken my work to the next level. Furthermore, Prof. Rolf Sijmons and Prof. Lude Franke have been important sources of inspiration and wild ideas for which I owe them my thanks. I am also grateful for the help from our editors Kate and Jackie. You turn what looks like English, into English. Last but not least, I want to thank my paranymphs Freerk and Bart, and everyone who contributed in their own way including friends, past and present colleagues, students and family. I hope I managed to include all your names in the 'DNA code' below. Lieve papa, mama, Lianne en Misha, dank voor jullie steun en interesse in mijn werk. Lieve Annelies, dank voor je verhelderende blik, creativiteit en het aanhoren van oefenpraatjes.

FEATURES Location/Qualifiers
source 1..1380
/organism="Homo sapiens"
/mol_type="acknowledgemental DNA"
/db_xref="taxon:9606"

```

mRNA          join(1..1380)
              /gene="YOU"
              /product="THESIS"

ORIGIN
1  grichardat ccmieketcc atpaulatac ahermanacg glioneltat ctateccacc
61  rosalieta ca ggtaarnett aga japketc terwincaac aacgdouweg aacgertjan
121 catkate tgc johncgacat gaghelenea caggerbent tagjonnegt atcmartijn
181 luukgtcgag agttacjana agctbirgit aaaseboacg agreneacgt ahenriette
241 gtcagjuhac tctgclinni machielatc tlennartga agccsander gctgludeaa
301 gtrijniett ctacellent aagggjtjosg gataatessa cat susanne jannekecat
361 cmennocgtg chenkaagac caagconnor aaccfloris ganniquecc aatasalome
421 gacchaosac atcarinatg tedgaraaca tattpeerta gpatrickga taannelies
481 tanicolien cctegcleoa xanderaaat aatroanaaa ccmntjegg copanoma
541 jacquescac tjonathang laurenttca ttaharment tataatgerc agrobinaaa
601 arjencagaa cgmichielc aniekaaaat tbastenato cciscaacta tpieterata
661 attcgeorge smartenaag klazienacg cruggeroga acharlieaa amariellea
721 jingyuanaa adavidgaac arobertacg cgtmariska catsophiaa gaacttleon
781 tsipkotggc aadespoina tmariotcgc gtommytcac marcjanaaa taaattdaan
841 jokettggca acttafleur ttomgtttcc tettejoris matthieuga gceddyagta
901 ctcpjotrga jackiegccc alaintgtct caandrewag apeteratgt asvenataat
961 amienteccc mariekeatc gtsidoaggt atggwesley ttaaaYangg kristinata
1021 gcwimatctc cingridaca acctc bart adannyagct johannekec cttgmorris
1081 ccgmartina gagtcgbote jipccctcct ttgtcthomg afreerkgt marloesatt
1141 ttcaactkimt dennistca markt atgag aactrobtat tadriaant cttaeditht
1201 tritserctc talextactc tcacerikat crolfetgta gtfrankgat tgacacahans
1261 tgcaaedwin cagccmisha awietzecca tliannecac ruditagaag konradaaca
1321 gaaca joela tmartineta ctjakobtaa tagaaelisa aaatkoenta jelkotatct
//

```

Appendix D

About the author

Kasper Joeri van der Velde was born on May 24th 1986 in Drachten (municipality of Smallingerland), The Netherlands. He completed his bachelors Bioinformatics in 2008, graduating on pathway visualizations of kinase activity at the Groningen Bioinformatics Centre in collaboration with the UMC Groningen dept. of Cell Biology. He continued to work on MOLGENIS software infrastructure for multi-omics research in Arabidopsis, mice and C. elegans at the Groningen Bioinformatics Centre. In 2012 he started as a PhD student at the UMC Groningen dept. of Genetics, working on new methods for downstream clinical analysis of next-generation sequenc-



APPENDIX D. ABOUT THE AUTHOR

ing data in the group of Morris Swertz. Amongst a unique crowd of clinicians, software developers, geneticists, parallel computing experts, wet-lab technicians and statisticians, he aims to discover new ways to revolutionize the speed, yield and applications of genome interpretation for medical genetics, powered by a wealth of untapped resources available in the public domain.

Appendix E

List of publications

1. Global genetic robustness of the alternative splicing machinery in *Caenorhabditis elegans*. Li Y, Breitling R, Snoek LB, **van der Velde KJ**, Swertz MA, Riksen J, Jansen RC, Kammenga JE. *Genetics*. 2010 Sep;186(1):405-10. doi: 10.1534/genetics.110.119677. Epub 2010 Jul 6.
2. OntoCAT – a simpler way to access ontology resources. Adamusiak T, Burdett T, **van der Velde KJ**, Abeygunawardena N, Antonakaki D, Parkinson H, and Swertz M. OntoCAT – a simpler way to access ontology resources. *Nature Precedings*. 2010. doi: 10.1038/npre.2010.4666.1
3. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. Swertz MA, Dijkstra M, Adamusiak T, **van der Velde JK**, Kanterakis A, Roos ET, Lops J, Thorisson GA, Arends D, Byelas G, Muilu J, Brookes AJ, de Brock EO, Jansen

APPENDIX E. LIST OF PUBLICATIONS

- RC, Parkinson H. BMC Bioinformatics. 2010 Dec 21;11 Suppl 12:S12. doi: 10.1186/1471-2105-11-S12-S12.
4. XGAP: a uniform and extensible data model and software platform for genotype and phenotype experiments. Swertz MA, **Velde KJ**, Tesson BM, Scheltema RA, Arends D, Vera G, Alberts R, Dijkstra M, Schofield P, Schughart K, Hancock JM, Smedley D, Wolstencroft K, Goble C, de Brock EO, Jones AR, Parkinson HE; Coordination of Mouse Informatics Resources (CASIMIR); Genotype-To-Phenotype (GEN2PHEN) Consortiums, Jansen RC. Genome Biol. 2010;11(3):R27. doi: 10.1186/gb-2010-11-3-r27. Epub 2010 Mar 9.
 5. OntoCAT—simple ontology search and integration in Java, R and REST/JavaScript. Adamusiak T, Burdett T, Kurbatova N, **Joeri van der Velde K**, Abeygunawardena N, Antonakaki D, Kapushesky M, Parkinson H, Swertz MA. BMC Bioinformatics. 2011 May 29;12:218. doi: 10.1186/1471-2105-12-218.
 6. Bioinformatics tools and database resources for systems genetics analysis in mice – a short review and an evaluation of future needs. Durrant C, Swertz MA, Alberts R, Arends D, Möller S, Mott R, Prins P, **van der Velde KJ**, Jansen RC, Schughart K. Brief Bioinform. 2012 Mar;13(2):135-42. doi: 10.1093/bib/bbr026. Epub 2011 Jul 8.
 7. Modifiers of mutant huntingtin aggregation - functional conservation of C. elegans-modifiers of polyglutamine aggregation. Teuling E, Bourgonje A, Veenje S, Thijssen K, de Boer J, **van der Velde J**, Swertz M, Nollen E. PLoS Curr. 2011 Aug 12;3:RRN1255. doi: 10.1371/currents.RRN1255.
 8. Observ-OM and Observ-TAB: Universal Syntax Solutions for the Integration, Search and Exchange of Phenotype And Genotype Information. Adamusiak T, Parkinson H, Muilu J, Roos E, **van**

-
- der Velde KJ**, Thorisson GA, Byrne M, Pang C, Gollapudi S, Ferretti V, Hillege H, Brookes AJ, Swertz MA. *Hum Mutat.* 2012 May;33(5):867-73. doi: 10.1002/humu.22070. Epub 2012 Apr 4.
9. xQTL workbench: a scalable web environment for multi-level QTL analysis. Arends D, **van der Velde KJ**, Prins P, Broman KW, Möller S, Jansen RC, Swertz MA. *Bioinformatics.* 2012 Apr 1;28(7):1042-4. doi: 10.1093/bioinformatics/bts049. Epub 2012 Feb 3.
 10. WormQTL—public archive and analysis web portal for natural variation data in *Caenorhabditis* spp. Snoek LB, **Van der Velde KJ**, Arends D, Li Y, Beyer A, Elvin M, Fisher J, Hajnal A, Hengartner MO, Poulin GB, Rodriguez M, Schmid T, Schrimpf S, Xue F, Jansen RC, Kammenga JE, Swertz MA. *Nucleic Acids Res.* 2013 Jan;41(Database issue):D738-43. doi: 10.1093/nar/gks1124. Epub 2012 Nov 24.
 11. An overview and online registry of microvillus inclusion disease patients and their MYO5B mutations. **van der Velde KJ**, Dhekne HS, Swertz MA, Sirigu S, Ropars V, Vinke PC, Rengaw T, van den Akker PC, Rings EH, Houdusse A, van Ijzendoorn SC. *Hum Mutat.* 2013 Dec;34(12):1597-605. doi: 10.1002/humu.22440. Epub 2013 Oct 16.
 12. Worm variation made accessible: Take your shopping cart to store, link, and investigate! Snoek LB, **Joeri van der Velde K**, Li Y, Jansen RC, Swertz MA, Kammenga JE. *Worm.* 2014 Jan 1;3(1):e28357. doi: 10.4161/worm.28357. Epub 2014 Mar 6.
 13. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Genome of the Netherlands Consortium (Laurent C Francioli, Androniki Menelaou, Sara L Pulit, Freerk van Dijk, ..more., **K Joeri van der Velde**, ..more.,

- Paul I W de Bakker, Morris A Swertz and Cisca Wijmenga). *Nat Genet.* 2014 Aug;46(8):818-25. doi: 10.1038/ng.3021. Epub 2014 Jun 29.
14. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. Deelen P, Bonder MJ, **van der Velde KJ**, Westra HJ, Winder E, Hendriksen D, Franke L, Swertz MA. *BMC Res Notes.* 2014 Dec 11;7:901. doi: 10.1186/1756-0500-7-901.
 15. WormQTL^{HD}—a web database for linking human disease to natural variation data in *C. elegans*. **van der Velde KJ**, de Haan M, Zych K, Arends D, Snoek LB, Kammenga JE, Jansen RC, Swertz MA, Li Y. *Nucleic Acids Res.* 2014 Jan;42(Database issue):D794-801. doi: 10.1093/nar/gkt1044. Epub 2013 Nov 11.
 16. BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. Pang C, Hendriksen D, Dijkstra M, **van der Velde KJ**, Kuiper J, Hillege HL, Swertz MA. *J Am Med Inform Assoc.* 2015 Jan;22(1):65-75. doi: 10.1136/amiajnl-2013-002577. Epub 2014 Oct 31.
 17. Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. Deelen P, Zhernakova DV, de Haan M, van der Sijde M, Bonder MJ, Karjalainen J, **van der Velde KJ**, Abbott KM, Fu J, Wijmenga C, Sinke RJ, Swertz MA, Franke L. *Genome Med.* 2015 Mar 27;7(1):30. doi: 10.1186/s13073-015-0152-4. eCollection 2015.
 18. Pheno2Geno - High-throughput generation of genetic markers and maps from molecular phenotypes for crosses between inbred strains. Zych K, Li Y, **van der Velde KJ**, Joosen RV, Ligterink W, Jansen RC, Arends D. *BMC Bioinformatics.* 2015 Feb 19;16:51. doi: 10.1186/s12859-015-0475-6.

-
19. Evaluation of CADD Scores in Curated Mismatch Repair Gene Variants Yields a Model for Clinical Validation and Prioritization. **van der Velde KJ**, Kuiper J, Thompson BA, Plazzer JP, van Valkenhoef G, de Haan M, Jongbloed JD, Wijmenga C, de Koning TJ, Abbott KM, Sinke R, Spurdle AB, Macrae F, Genuardi M, Sijmons RH, Swertz MA, InSiGHT Group. *Hum Mutat.* 2015 Jul;36(7):712-9. doi: 10.1002/humu.22798. Epub 2015 May 20.
 20. MOLGENIS/OMX for multi-omics and personalized medicine. Morris Swertz and **K. Joeri van der Velde**. *Clin Bioinforma.* 2015; 5(Suppl 1): S5. Published online 2015 May 22. doi: 10.1186/2043-9113-5-S1-S5
 21. SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data. Chao Pang, Annet Sollie, Anna Sijtsma, Dennis Hendriksen, Bart Charbon, Mark de Haan, Tommy de Boer, Fleur Kelpin, Jonathan Jetten, **Joeri K. van der Velde**, Nynke Smidt, Rolf Sijmons, Hans Hillege, and Morris A. Swertz. *Database (Oxford).* 2015; 2015: bav089. Published online 2015 Sep 17. doi: 10.1093/database/bav089
 22. MOLGENIS/connect: a system for semi-automatic integration of heterogeneous phenotype data with applications in biobanks. Pang C, van Enkevort D, de Haan M, Kelpin F, Jetten J, Hendriksen D, de Boer T, Charbon B, Winder E, **van der Velde KJ**, Doiron D, Fortier I, Hillege H, Swertz MA. *Bioinformatics.* 2016 Jul 15;32(14):2176-83. doi: 10.1093/bioinformatics/btw155. Epub 2016 Mar 21.
 23. GAVIN - Gene-Aware Variant INterpretation for medical sequencing. **K. Joeri van der Velde**, Eddy N. de Boer, Cleo C. van Diemen, Birgit Sikkema-Raddatz, Kristin M. Abbott, Alain Knoppers, Lude Franke, Rolf H. Sijmons, Tom J. de Koning, Cisca Wijmenga, Richard J. Sinke and Morris A. Swertz. *Genome Biology.* 2017, 18(1). doi:10.1186/s13059-016-1141-7

APPENDIX E. LIST OF PUBLICATIONS

24. reGenotyper: detecting mislabeled samples in genetic data. Konrad Zych, L. Basten Snoek, Mark Elvin, Miriam Rodriguez, **K. Joeri van der Velde**, Danny Arends, Harm-Jan Westra, Morris A. Swertz, Gino Poulin, Jan E. Kammenga, Rainer Breitling, Ritsert C. Jansen and Yang Li. PLOS ONE. 2017, e0171324(2) doi:10.1371/journal.pone.0171324
25. Rapid Targeted Genomics in Critically Ill Newborns. Cleo C. van Diemen, Wilhelmina S. Kerstjens-Frederikse, Klasien A. Bergman, Tom J. de Koning, Birgit Sikkema-Raddatz, **K. Joeri van der Velde**, Kristin M. Abbott, Johanna C. Herkert, Katharina Löhner, Patrick Rump, Martine T. Meems-Veldhuis, Pieter B.T. Neerincx, Jan D.H. Jongbloed, Conny M. van Ravenswaaij-Arts, Morris A. Swertz, Richard J. Sinke, Irene M. van Langen and Cisca Wijmenga. Pediatrics 2017. Published Online September 22, 2017. doi: 10.1542/peds.2016-2854

Appendix F

Other academic activities

APPENDIX F. OTHER ACADEMIC ACTIVITIES

Date	Activity
01-09-2012	Participated in the GSMS Project Management course
06-09-2012	Attended the 20th Annual GBB Symposium
12-09-2012	Attended the BioSHaRE Annual Meeting, Paris
27-09-2012	Gave an oral presentation at the LFN Symposium, Wageningen
08-11-2012	Participated in the PhD Introduction Event, Ezinge
19-11-2012	Attended the Connecting Biobanks Meeting, Utrecht
28-11-2012	Participated in the course 'Introduction to HL7/DCM'
03-12-2012	Participated in the Nordic BBMRI Meeting, Tartu
07-01-2013	Talked at the ENCODE Journal Club about machine learning
13-03-2013	Visited dermatology clinic at Instytut Matki i Dziecka, Warsaw
19-03-2013	Attended the CTMM TraIT Symposium, Utrecht
03-04-2013	Gave an oral presentation at the TarGet Conference
16-04-2013	Presented a poster at the NBIC Conference, Lunteren
16-04-2013	Reviewed a paper for J. of the Am. Medical Informatics Assoc.
14-05-2013	Received the BOSC 2013 Student Travel Award
17-06-2013	Participated in the SYSGENET MC Meeting, Prague
18-06-2013	Gave an oral presentation at the BOSC Conference, Berlin
25-06-2013	Visited the dept. of Genetics at the University of Leicester
01-07-2013	Supervised internship student Mark de Haan
11-07-2013	Reviewed a paper for Journal of Web Semantics
26-08-2013	Participated in the GOPHER/RUG PhD Day
11-09-2013	Gave an oral presentation at the BMB Meeting, Dusseldorf
12-09-2013	Presented a poster at the CTMM Annual Meeting, Utrecht
04-11-2013	Gave an oral presentation at the BioShare AM, Barcelona
21-11-2013	Attended the HandsOn Biobanks Conference, The Hague
12-12-2013	Visited several research groups at the EBI, Hinxton
01-01-2014	Supervised graduation student Pieter Dopheide
01-01-2014	Supervised graduation student Mark de Haan
28-01-2014	Talked at ADCB Journal Club about Deng <i>et al.</i> (Science)
23-02-2014	Attended the Joint RD-Connect Meeting, Heidelberg
20-05-2014	Gave an oral presentation at the HVP5 Conference, Paris
31-05-2014	Presented a poster and satellite talk at ESHG Conference, Milan
01-07-2014	Supervised graduation student Tommy de Boer
20-08-2014	Taught a course segment at UMCG Biobanking Summer School

Table F.1: Other academic activities, pt. 1/2.

Date	Activity
12-09-2014	Written a Jan Kornelis de Cock grant proposal
20-09-2014	Participated in the GOPHER/RUG PhD Day
28-10-2014	Taught course segment at VKGL/VKGN NGS diagn., Rotterdam
28-11-2014	Presented a poster at Connecting Biobanks Conference, Leiden
01-01-2015	Supervised internship student Marieke Bijlsma
19-01-2015	Talked at ADCB Journal Club about Leiserson <i>et al.</i> (Nat. Gen.)
02-02-2015	Reviewed a paper submission for ISMB/ECCB
31-03-2015	Participated in the course 'Introd. to Genetic Epidem. Research'
06-06-2015	Presented a poster at the ESHG Conference, Glasgow
11-06-2015	Attended the GSMS PhD Development Conference
25-06-2015	Talked at the Epigenome Journal Club about DIY analysis in R
01-07-2015	Supervised graduation student Marieke Bijlsma
10-07-2015	Presented a poster at the BOSC Conference, Dublin
22-09-2015	Taught course segment at VKGL/VKGN NGS diagn., Rotterdam
05-11-2015	Talked at ADCB Journal Club about Itan <i>et al.</i> (PNAS)
01-01-2016	Supervised graduation student Mariska Slofstra
01-01-2016	Supervised graduation student Thom Steenhuis
18-02-2016	Participated in the ProjectFactory course by TOC consultants
04-03-2016	Written a BBMRI voucher on variant data sharing
21-03-2016	Participated in the course 'Publishing in English gr.c'
14-04-2016	Talked at ADCB Journal Club about Zhu <i>et al.</i> (Nat. Gen.)
21-05-2016	Presented a poster and a satellite talk at ESHG, Barcelona
01-06-2016	Reviewed a paper for The American Journal of Human Genetics
03-09-2016	Gave an oral presentation at the ECCB Conference, The Hague
20-09-2016	Taught course segment at VKGL/VKGN NGS diagn., Rotterdam
11-10-2016	Taught a course segment at MPDI TopMaster course I
03-11-2016	Talked at ADCB Journal Club about Turpin <i>et al.</i> (Nat. Gen.)
19-12-2016	Reviewed a paper for Genome Biology
03-07-2017	Visited the SciLifeLab Clinical Genomics facility, Stockholm
07-09-2017	Started to supervise graduation student Sander van den Hoek
11-09-2017	Taught course segment at VKGL/VKGN NGS diagn., Rotterdam
02-10-2017	Participated in Bioschemas Adoption Meeting at EBI, Hinxton
09-10-2017	Started to supervise internship student Peer Ketelaars
16-11-2017	Gave an oral presentation at the NASPM Conference, Høvik

Table F.2: Other academic activities, pt. 2/2.

Back cover image explained

The back cover shows an altered version of the Arecibo message, send out on 16 November 1974 from the Arecibo Observatory in Puerto Rico. Key numbers were updated, the antenna dish graphic left out, and it was extended for medical genetics / bioinformatics. The original transmission was broadcasted with a power of 1,000 kW towards globular star cluster M13, which it will reach in roughly 25,000 years.

The Arecibo message as send out in 1974, with key numbers updated with data from 2016		The numbers 1-10 in binary notation, note that the '8' is offset to the right relative to the start-of-number indicator			
		Atomic numbers (i.e. proton number) of the elements H (1), C (6), N (7), O (8) and P (15), used in the DNA molecular diagram below			
		Deoxyribose (C ₅ H ₁₀ O)	Adenine (C ₅ H ₅ N ₅)	Thymine (C ₅ H ₇ N ₂ O ₂)	Deoxyribose (C ₅ H ₁₀ O)
		Phosphate (PO ₄)			Phosphate (PO ₄)
		Deoxyribose (C ₅ H ₁₀ O)	Cytosine (C ₄ H ₅ N ₃ O)	Guanine (C ₅ H ₇ N ₅ O)	Deoxyribose (C ₅ H ₁₀ O)
		Phosphate (PO ₄)	Number base pairs in the human genome, updated from the 1974 estimate of 4.3 billion to 3,077,073,773 (non-N bases in GRCh38.p9, 26/9/2016), in binary: 1011011101101000 0110101101101101		Phosphate (PO ₄)
		Double helix structure of a DNA molecule			Double helix structure of a DNA molecule
		Human height, as 14 (1110) x 126 mm (the transmission wavelength) equals 1764 mm (i.e. 1.76m)	Graphical representation of a human individual	Human population size, updated from ~4.3 billion in 1974 to 7,473,173,118 at 23/12/2016	
		Our solar system, with Earth standing out as our home planet			
	Addition to depict medical genetics / bioinformatics		Common variation in our genome, indicated by the juts on a strand of DNA	Pathogenic variants in our genome, indicated by a single (rare) jut on a strand of DNA	
		88M variants (1KG proj.)	50k path. var. (ClinVar 1/12/2016)		
		Atomic numbers for elements Al (13), Si (14), Cu (29), Sn (50) and Au (79), crucial for modern digital computing	Networked processors i.e. the internet	Human with a disorder (close to path. vars.)	
		Numeric inputs: 1 and 1 2 and 1 3 and 2	A central processing unit (CPU), calculating output from input using conductive traces	Numeric outputs from basic addition: 2, 3 and 5	
		One (one) input genome with new variants	CPU helps to analyze a genome	The output genome with one candidate variant selected	

Translational software infrastructure for medical genetics



Deep inside the core of our cells resides the deoxyribonucleic acid (DNA) molecule known as the genome. DNA encodes the information that allows life to grow, survive, diversify and evolve. Unfortunately, the same mechanisms that let us adapt to a changing environment can also cause genetic disorders. While we are able to diagnose a number of these disorders using modern technological advancements, much remains to be discovered and understood. This thesis presents software infrastructure for investigating the molecular etiology of genetic disease using data from model organisms, demonstrates how to translate findings from fundamental research into new software tools for genome diagnostics, and introduces a downstream genome analysis framework that assists the automation and validation of the latest tools for applied patient care.

ISBN: 978-94-034-0351-9