

---

**MutationDistiller – User-driven  
identification of disease mutations**

---

Inaugural-Dissertation  
to obtain the academic degree of  
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of  
Biology, Chemistry, Pharmacy  
of Freie Universität Berlin

by  
Daniela Rebekka Hombach

2019



Angefertigt zwischen Oktober 2014 und Juni 2019 unter der Leitung von Prof. Dr. Dominik Seelow an der Charité – Universitätsmedizin Berlin und im Berliner Institut für Gesundheitsforschung (BIH).

**Erster Gutachter:**

Prof. Dr. Dominik Seelow  
Berliner Institut für Gesundheitsforschung  
Charité – Universitätsmedizin Berlin  
Charitéplatz 1, 10117 Berlin

**Zweiter Gutachter:**

Prof. Dr. Sigmar Stricker  
Fachbereich Biologie, Biochemie, Pharmazie  
Institut für Biochemie und Chemie  
Freie Universität Berlin  
Takustr. 3, 14195 Berlin

Disputation am: 16.09.2019



# Acknowledgements

MutationDistiller and this thesis would not have been possible without the help and support of a whole lot of awesome people. I would like to express my special thanks to everyone who supported my work academically:

First and foremost, I would like to express my great appreciation and gratefulness to Prof. Dr. Dominik Seelow. MutationDistiller would never have happened without your support, encouragement, helpful input, programming aid, debugging, and ice cream! Thank you for giving me the freedom but also the helpful framework to make this happen.

I extend my special thanks to Prof. Dr. Markus Schülke for supporting me throughout the last five years. Thank you for your input to MutationDistiller, for taking me on board and for encouraging me all the way.

To Prof. Dr. Sigmar Stricker, thank you so much for agreeing to evaluate this thesis.

A big thank you to my colleagues in the Translational Genomics group, especially to Jana-Marie Schwarz and Sebastian Köhler for inspiring discussions on scientific or not-so scientific topics, for input to this thesis, for fun times and of course for unicorns and nerd guns.

To all the clinicians and researchers in the Neuropaediatrics group and in the Medical Genomics department at Charité - thank you so much for your input! Special thanks to Nadja Ehmke, Ellen Knierim, Gudrun Schottmann, Christine Oeien and Björn Fischer-Zirnsak.

I would also like to thank my colleagues in the Neuropaediatrics group for their moral and emotional support. Special thanks to Franziska Seiffert, Esther Gill and Ioana Polydorou for lots of laughs. You rock, babes!

Thank you to the Studienstiftung des Deutschen Volkes for generously funding not only my PhD thesis but also my previous studies. Special thanks to Dr. Peter Antes and Dr. Vivien Petras for your support and willingness to read reports and extension applications!

It takes a village to raise a child, or two children in this case. Thank you to everyone who is or has been helping with the kids and supported my work in this way:

To the Charité Familienbüro for the emergency childcare scheme KidsMobil, and special thanks to Marion and Sofija for looking after the kids numerous times.

Great thanks to all my friends and family who helped mind the kids whenever it was needed, especially to Diana, Maddie, Max, Leoni, and Lina!

I would not have been able to pursue this project or my studies without the never-ending support of my loved ones:

Thank you to my in-laws in Australia for taking us in twice during maternity leave and looking after the kids while I was writing this thesis! Thank you for loads of tea, ginger beer and chips on the beach, which were all absolutely essential to this project.

Ein unglaublich großes Danke an meine Eltern und Geschwister für all die Unterstützung, die ich immer von euch bekomme. Ich weiß, dass ich mich immer auf euch verlassen kann – ohne euren Rückhalt wäre das alles nicht machbar. Ich bin sehr dankbar, dass ich euch habe!

An meine Kinder: Danke, dass ihr da seid und mir immer wieder mit euren Ideen und Wünschen zeigt, dass es auch neben einer Doktorarbeit noch ein wunderbares Leben gibt.

Last but definitely not least: Thank you, James, for your love, support and patience with me and my work. Without you and your constant help, none of this would have been feasible. You're not only my best friend but also my greatest support. Damn!

# MutationDistiller – User-driven identification of disease mutations

Inaugural-Dissertation to obtain the academic degree of  
Doctor rerum naturalium (Dr. rer. nat)

**Daniela Hombach**

## **Abstract**

In rare genetic diseases, a single genetic alteration can be enough to cause a severe disorder. Recent advances in genetic research have introduced exome or genome sequencing into clinical care. However, each sequencing run delivers a myriad of candidate variants that have to be sifted through in the hunt for the causative mutation - a major data challenge, for which researchers and clinicians have to rely on computer tools.

With *MutationDistiller*, we have developed a freely available online tool to analyse whole exome sequencing data in a user-driven fashion. The tool aims at clinicians and researchers without bioinformatic experience who are working with real patient data, and allows them to distil the most likely causative variants from the sea of candidates. By uploading the patient's genetic information and adding information on the symptoms, they can combine genotype and phenotype to find the culprit. MutationDistiller allows a wide range of phenotype data, such as HPO, OMIM and Orphanet entries, gene panels, expression data, Gene Ontology terms, and affected pathways. In the output, the program provides an ordered list of candidate alterations matching the user-defined criteria. In addition, crucial data on the alteration and the affected gene can be reviewed at a glance.

This thesis describes the program, its background and usage, and compares it to current state-of-the-art tools. When assessing the tool, we found that it matches or out-competes similar software and is able to find the causative variant in a majority of cases. Moreover, its user-friendliness makes it a handy tool for clinicians and researchers, as is reflected by its usage: MutationDistiller routinely sees over 1,000 cases per month and has been used in over 14,000 cases at the time of writing. Thus, MutationDistiller has already found its way into the clinic.

The tool, comprehensive documentation and example cases are freely available at <https://www.mutationdistiller.org/>

# MutationDistiller – User-driven identification of disease mutations

Inaugural-Dissertation to obtain the academic degree of  
Doctor rerum naturalium (Dr. rer. nat)

**Daniela Hombach**

## Zusammenfassung

Im Fall von monogenen Krankheiten kann eine einzelne schädliche Mutation krankheitsauslösend sein. Fortschritte in der genetischen Forschung haben dazu beigetragen, dass Genom- oder Exomsequenzierungen zur Detektion krankheitsverursachender Mutationen in der Klinik einen Platz gefunden haben. Bei jeder Sequenzierung fallen jedoch Abertausende von Varianten an, die gefiltert und eingeordnet werden müssen. Für diese datentechnische Herausforderung müssen sich Forscher\*innen und Kliniker\*innen auf Computerprogramme verlassen.

Diese Arbeit beschreibt *MutationDistiller*, ein frei verfügbares Web-Programm zur Analyse von Exomsequenzierungsdaten, das sich an Kliniker\*innen und Forscher\*innen ohne bioinformatische Fachkenntnis richtet. Das Programm ermöglicht nutzerorientierte Untersuchungen zur Auffindung der krankheitsverursachenden Mutation(en) aus einer Vielzahl von Kandidaten. MutationDistiller kombiniert dabei Genotyp und Phänotyp der Patient\*innen und erlaubt somit einen Fokus auf die Genveränderungen, die im konkret vorliegenden Fall am wahrscheinlichsten für die weitere Analyse von Interesse sind. Eine Vielzahl von Phänotypdaten werden akzeptiert, unter anderem HPO, OMIM und Orphanet-Einträge, Listen von Kandidatengenomen, Expressionsdaten, Daten der Gene Ontology oder auch zu betroffenen Signaltransduktionswegen. Die Ergebnisseite fasst die Daten in nutzerfreundlichen Tabellen zusammen und zeigt detaillierte Informationen zu allen Kandidatengenomen sowie Hyperlinks zu weiteren Ressourcen, um die Einschätzung der Relevanz der Ergebnisse zu vereinfachen.

Diese Arbeit beschreibt Aufbau, Hintergrund und Nutzung von MutationDistiller sowie einen Vergleich mit ähnlich gelagerten Programmen. MutationDistiller hat bereits den Weg in die Klinik gefunden und wurde bisher in über 14.000 Fällen angewendet. Das Programm, eine umfassende Dokumentation und Beispielfälle sind frei verfügbar unter <https://www.mutationdistiller.org/>



# Publications

Figures 1.3 and 7.2 and parts of this thesis, in particular regarding the comparison with other tools, have been published previously in

**Hombach D**, Schuelke M, Knierim E, Ehmke N, Schwarz JM, Fischer-Zirnsak B, Seelow S. MutationDistiller – user-driven identification of pathogenic DNA variants. NAR Web Server Issue. 2019. doi:10.1093/nar/gkz330

## Further publications:

Schwarz JM, **Hombach D**, Koehler S, Cooper DN, Schuelke M, Seelow D RegulationSpotter: annotation and interpretation of extratranscriptic variants. NAR Web Server Issue. 2019. doi:10.1093/nar/gkz327

**Hombach D\***, Schwarz JM\*, Knierim E, Schülke M, Seelow D, Köhler S. Phenotero: Annotate as you write. Clinical Genetics. 2018. doi: 10.1111/cge.13471.

(\*: authors contributed equally)

**Hombach D**, Schwarz JM, Robinson PN, Schülke M, Seelow D. A systematic, large-scale comparison of transcription factor binding site models. BMC Genomics. 2016. doi: 10.1186/s12864-016-2729-8.



# Contents

Acknowledgements	v
Abstract	vii
Zusammenfassung	viii
Publications	ix
List of Figures	xiv
List of Tables	xiv
<b>1 Introduction</b>	<b>1</b>
1.1 The human genome . . . . .	1
1.1.1 The genetic code . . . . .	1
1.1.2 From genotype to phenotype . . . . .	2
1.2 Studying DNA variation . . . . .	4
1.3 Genetic disease and inheritance . . . . .	5
1.3.1 Complex diseases . . . . .	5
1.3.2 Mendelian disorders . . . . .	6
1.3.3 Inheritance patterns . . . . .	8
1.4 Finding disease causes in Mendelian disorders . . . . .	10
1.4.1 Next Generation Sequencing and bioinformatics . . . . .	11
1.4.1.1 Sequencing . . . . .	11
1.4.1.2 Data processing . . . . .	13
1.4.1.3 Variant annotation . . . . .	14
1.4.1.4 Variant prioritisation: patient information . . . . .	16
1.4.1.5 Ontologies . . . . .	19
1.4.2 Current variant prioritisation tools . . . . .	21
1.5 MutationDistiller . . . . .	24
1.5.1 Technical information . . . . .	24
1.5.2 MutationTaster . . . . .	25
1.5.3 GeneDistiller . . . . .	26
1.5.4 Combining genotype and phenotype . . . . .	26
<b>2 MutationDistiller: Data integration</b>	<b>28</b>
2.1 Databases . . . . .	29
2.2 Data sources . . . . .	30
2.2.1 MutationTaster predictions . . . . .	30
2.2.2 Genetic data . . . . .	31
2.2.2.1 Ensembl . . . . .	31
2.2.2.2 Entrez Gene . . . . .	32
2.2.3 Variant databases . . . . .	32
2.2.3.1 1000 Genomes Project . . . . .	32
2.2.3.2 Exome Aggregation Consortium Browser (ExAC) . . . . .	33
2.2.3.3 dbSNP and ClinVar . . . . .	33

2.2.4	Phenotype repositories . . . . .	34
2.2.4.1	Online Mendelian Inheritance in Man . . . . .	34
2.2.4.2	Orphanet . . . . .	34
2.2.4.3	Mouse Genome Database . . . . .	35
2.2.4.4	Human Phenotype Ontology . . . . .	35
2.2.5	Gene and protein function . . . . .	36
2.2.5.1	Gene Ontology . . . . .	36
2.2.5.2	Expression data . . . . .	36
2.2.5.3	Metabolic and signalling pathways . . . . .	38
2.2.5.4	Gene panels . . . . .	39
2.2.6	Protein families . . . . .	41
2.2.7	Protein-protein interactions . . . . .	42
2.2.7.1	STRING . . . . .	42
2.2.8	Mitochondrial data . . . . .	42
2.3	MutationDistiller’s database . . . . .	43
2.3.1	Database structure . . . . .	43
2.3.1.1	Query Engine schema . . . . .	43
2.3.1.2	MutationDistiller database schemas . . . . .	44
2.3.1.3	MutationDistiller entity-relationship diagram . . . . .	46
<b>3</b>	<b>MutationDistiller: Query Engine</b>	<b>47</b>
3.1	File upload . . . . .	47
3.2	Query Engine workflow . . . . .	47
<b>4</b>	<b>MutationDistiller: Prioritisation</b>	<b>53</b>
4.1	Filtering, scoring and providing information . . . . .	53
4.1.1	Initialising . . . . .	53
4.1.2	Gene information . . . . .	54
4.1.2.1	MutationDistiller score . . . . .	56
4.1.2.2	Scoring HPO matches . . . . .	56
4.1.3	Output generation . . . . .	58
<b>5</b>	<b>MutationDistiller: User interface</b>	<b>63</b>
5.1	Input and output pages . . . . .	63
5.1.1	Landing page . . . . .	63
5.1.2	Query Engine user interface . . . . .	63
5.1.3	User modes . . . . .	65
5.1.4	Query interface . . . . .	65
5.1.5	Output page . . . . .	69
5.2	Manual and tutorial pages . . . . .	71
<b>6</b>	<b>Implementation and Tools</b>	<b>72</b>
6.1	Software development . . . . .	72
6.1.1	MutationDistiller . . . . .	72
6.1.2	Query Engine . . . . .	73
6.2	Manuscript . . . . .	73
6.3	Hardware . . . . .	73

<b>7</b>	<b>MutationDistiller: Optimisation and validation</b>	<b>74</b>
7.1	Determination of HPO weights . . . . .	74
7.1.1	Training Data . . . . .	74
7.1.2	HPO weight parameter selection . . . . .	75
7.2	Testing and validation . . . . .	77
7.2.1	Test set . . . . .	77
7.2.2	Validation . . . . .	77
7.2.3	Comparison with state-of-the-art prioritisation tools . . . . .	79
<b>8</b>	<b>Discussion</b>	<b>82</b>
8.1	Data selection process . . . . .	82
8.1.1	Integrated data types . . . . .	82
8.1.2	Testing and training data . . . . .	85
8.2	Scoring . . . . .	86
8.2.1	HPO score optimisation . . . . .	86
8.2.2	Mutation severity . . . . .	87
8.3	Phenotype data variety . . . . .	87
8.3.1	Detection of new disease genes . . . . .	87
8.3.2	Symptom annotations . . . . .	88
8.4	Comparison with state-of-the-art tools . . . . .	89
8.5	Outlook . . . . .	90
8.5.1	Family analyses . . . . .	90
8.5.2	Genome version . . . . .	91
8.5.3	WGS data . . . . .	91
8.5.4	Mitochondrial DNA . . . . .	92
8.5.5	gnomAD . . . . .	92
8.5.6	Classification bins . . . . .	93
8.5.7	Data management . . . . .	93
8.5.7.1	User data sustainability . . . . .	93
8.5.7.2	Data sources . . . . .	94
8.6	Clinical use . . . . .	94
	<b>Abbreviations</b>	<b>96</b>
<b>A</b>	<b>Appendix – HPO optimisation weights</b>	<b>98</b>
<b>B</b>	<b>Appendix – Expression tissue groups</b>	<b>104</b>
<b>C</b>	<b>Appendix – MutationDistiller database ERD</b>	<b>112</b>
	<b>Bibliography</b>	<b>114</b>
	<b>Statement of independent work</b>	<b>125</b>

## List of Figures

1.1	Overview of genotypes and resulting phenotypes . . . . .	9
1.2	Excerpt of the graphical representation of the HPO . . . . .	21
1.3	Current variant prioritisation tools . . . . .	23
2.1	WikiPathways Bone Morphogenic Protein (BMP) Signalling and Regulation pathway (WP1425) . . . . .	40
2.2	Overview of MutationDistiller’s Query Engine Schema . . . . .	45
2.3	MutationDistiller database ERD . . . . .	46
3.1	Simplified view of MutationDistiller’s Query Engine workflow, part 1 . . .	50
3.1	Simplified view of MutationDistiller’s Query Engine workflow, part 2 . . .	51
3.1	Simplified view of MutationDistiller’s Query Engine workflow, part 3 . . .	52
4.1	Simplified view of MutationDistiller’s prioritisation workflow, part 1 . . .	59
4.1	Simplified view of MutationDistiller’s prioritisation workflow, part 2 . . .	60
4.1	Simplified view of MutationDistiller’s prioritisation workflow, part 3 . . .	61
4.1	Simplified view of MutationDistiller’s prioritisation workflow, part 4 . . .	62
5.1	Screenshot of MutationDistiller’s landing page . . . . .	63
5.2	Screenshot of MutationDistiller’s Query Engine upload page . . . . .	64
5.3	Screenshot of MutationDistiller’s query interface . . . . .	66
5.4	Screenshot of MutationDistiller’s variant selection . . . . .	67
5.5	Screenshot of MutationDistiller’s result table . . . . .	70
5.6	Screenshot of MutationDistiller’s detailed view . . . . .	70
7.1	MutationDistiller rank distribution for the validation set . . . . .	79
7.2	Tool Comparison: Cumulative rank frequencies . . . . .	81

## List of Tables

2.1	MutationTaster data integrated into MutationDistiller . . . . .	30
2.2	MutationDistiller expression data sources . . . . .	37
4.1	MutationDistiller gene information . . . . .	55
4.2	Weight categories overview . . . . .	56
7.1	HPO optimisation weights . . . . .	75
7.2	HPO weight iterations . . . . .	76
7.3	Validation set ranks . . . . .	78

# 1 Introduction

## 1.1 The human genome

### 1.1.1 The genetic code

The blueprint for human traits – anything that makes us unique and determines large parts of how we look, think and behave – is stored in the human genome. Determining the nature of this information has kept scientists and philosophers busy for centuries. Ancient thinkers and philosophers such as Hippocrates, Epicurus and Aristotle developed theories on how traits are determined long before the existence of genes or genomes was even postulated. In more modern times, breeding experiments conducted by Gregor Mendel in the 19th century laid the groundwork for what would later be called genetic research. In the 20th century, molecular approaches slowly led to the realisation that *deoxyribonucleic acid*, or DNA, was the carrier of those traits.

After decades of experiments by numerous researchers, each adding to the growing mountain of knowledge, the structure of this large molecule was finally determined by James Watson and Francis Crick [1] with important contributions by Maurice Wilkins, Raymond Gosling and Rosalind Franklin [2]. Thanks to all these advances and following research, we now know that DNA consists of repetitions and repetitions of nucleotides. These nucleotides, or *bases*, exist in the four varieties Adenine (A), Guanine (G), Thymine (T) and Cytosine (C) and arrange themselves in base pairs: A pairs with T, and G with C. Together, they form a double helix which twists and turns around itself, becoming coiled and tightly packed and organised into 2 sets of 23 chromosomes. These sets of chromosomes play an important role in inheritance, as each individual receives one set from their mother and one from their father.

The packing mechanism allows the approximately 2m long helix to be squeezed into the nucleus of cells which are 1,000,000 times smaller. In addition to this nucleic DNA, a small portion of the genome is present in mitochondria. In total, all chromosomes and the mitochondrial genome encode about 23,000 protein-coding genes, the molecular units of heredity, which are stored in about 6 billion base pairs.

The contents of the human genome – the sequence of the nucleic acid base pairs – can be read, much like a book. This technique of reading the genome or single genes is referred to as *DNA sequencing* and has enabled deep insights into the properties, structure and organisation of the DNA. The first draft of the human genome – the first complete

sequence – was achieved in a dramatic head-to-head race in 2001 by the Human Genome Project and Celera, a private venture established by Craig Venter [3, 4].

This step marked the onset of a new era in genetic research: Knowing the contents of the human genome sparked a whole new approach to the science of genetics and inheritance, enabling us to assess the mechanisms behind genetic diseases. As the methodologies for DNA sequencing improved, the costs dropped dramatically, enabling the inception of large-scale genomic projects. As a consequence, thousands of full or partial human genomes have been sequenced to date. Nowadays, even though the details of the genome are not entirely understood, we have a good insight into the variety and variability of human genomes and what consequences genetic changes can have for an organism.

## 1.1.2 From genotype to phenotype

The entirety of an individual's traits, which are largely determined by the genome as a blueprint, are often referred to as their *phenotype*. The genetic blueprint has to be translated into function: In a multi-step process, genes have to be read, or *transcribed*, into ribonucleic acid (RNA) and from there *translated* into proteins. Those genes that carry traits are usually termed *protein-coding genes* (as opposed to other types of genes which do not encode proteins but take on regulatory functions). Due to their importance for disease and this thesis, I will focus on protein-coding genes in this thesis and use the term *genes* for protein-coding genes unless indicated otherwise.

The processes of transcription – often termed *gene expression* – and translation are the basis of molecular functions. They are complex and well-regulated procedures which have been studied intensively and deserve their own theses. In the following, I will therefore limit myself to a short, simplified introduction to transcription and translation in humans with a focus on disease relevance.

### Transcription

In the first step of gene expression, a gene has to be transcribed to generate a messenger RNA (mRNA) molecule. RNA is a molecule quite similar to DNA but comes with three major differences: First, it is single-stranded (and therefore doesn't take on the shape of a double helix). Second, instead of a deoxyribose sugar it contains a ribose sugar; and third, instead of the base thymine (T) it contains a slightly different base called uracil (U).

Transcription begins in the promoter region of a gene, more precisely at the *transcription start site (TSS)*. There, *transcription factors*, proteins that regulate the process of transcription, bind to ensure that gene expression takes place exactly when and where



it is needed. The DNA splits open to generate a single strand that serves as a template. From this template DNA, the mRNA can be created by pairing complementary bases to it – A to T, G to C. The resulting premature mRNA then has to be processed further to form mature RNA. One main processing step is termed *splicing*:

The premature mRNA is full of sections which are not present in the mature mRNA, termed *introns*. They are removed, or *spliced out*, leading to an mRNA containing only *exons*. Splicing happens at specific sites termed *splice sites*. For each premature mRNA molecule, there are several ways for it to be spliced, leading to varying gene products or transcripts. This means that one single gene can actually generate a number of different mRNAs (and, in consequence, proteins). These different versions are often referred to as *transcripts*.

Non-protein-coding genes are transcribed and undergo maturation steps, but the next step, translation into protein, does not occur for them. Instead, they take on their important functions, e.g. for tRNAs the transfer of amino acids.

### **Translation**

In the next step, the remaining protein-coding mRNA has to be translated into a protein sequence. The mRNA is ‘read’ in 3-letter ‘words’ called *codons*: Each codon consists of three mRNA bases and encodes for one specific amino acid. In addition, four codons have the regulatory function to denote the start and end of the translation process. The start is determined by the start codon AUG – which plays a double role as it also encodes for the amino acid methionine – whereas three different codons serve as stop codons: UAA, UAG, and UGA. The span between start and stop codon is referred to as the *open reading frame*, *ORF*. Regions located within the mRNA but before the start site and after the stop codon do not become translated and are referred to as *untranslated regions*, *UTRs*. Even though they are not part of the final protein, these regions are still important because they take on regulatory functions.

### **Disease relevance**

Genetic alterations influencing the processes of transcription and translation can be the cause of genetic disease. For example, variants in the promoter region or the TSS can lead to too low or high transcriptional rates, which has been found to influence susceptibility or survival rates in cancers as well as other diseases [5–7]. Moreover, splice site aberrations can lead to altered proteins which are not able to fulfil their normal function. It was recognised some time ago that splice site aberrations are relevant to cancer [8], and since then further examples have appeared at a steady rate (e.g. [9, 10]). In addition, faulty splicing has been found to be involved in other genetic disorders such as the hereditary eye condition retinitis pigmentosa [11, 12]. Thus, a lot of evidence is accumulating which indicates that alterations affecting the processes of translation and transcription are relevant to the development of genetic disorders.

## 1.2 Studying DNA variation

Human individuality and variability are represented in the genomic sequence: Every individual carries a multitude of genetic variants – smaller or bigger alterations in the genetic sequence between individuals – which often have no effect, but are also the underlying cause for hair colour, height, size, or weight. On average, 1 in every 1000 base pairs is a genetic variant, which across the entire genome amount to millions of differences between any two individuals on the planet (except for identical twins): The exact number of variants is hard to gauge and varies widely depending on the study, methods, reference group, and other factors. However, every single sequencing run of an entire human genome detects on average 3 to 4 million alterations [13]. These genetic variants are, in most cases, harmless – like a different spelling of a word. However, sometimes, differences in the human genome can be harmful and cause disease. Detecting these harmful alterations and their implications on human health are two major goals in medicine and research, and they can be addressed by genomic sequencing.

Ever since the basic structure of the genome was determined, and possibly even earlier, scientists dreamed of reading its content to uncover the secrets hidden in it. The first method to achieve this at a mid-throughput level was developed in 1977 by Fred Sanger and colleagues [14] and is based on chain-termination during *in vitro* DNA replication:

In its early days, *Sanger sequencing* was conducted using a modified DNA polymerase called *Sequenase*. For *Sequenase* sequencing, the DNA strand to be analysed is combined with essential components for DNA replication: DNA primer, DNA polymerase and nucleotides. In addition, chain-terminating dideoxynucleotides (ddNTPs) are added to the mix. As these ddNTPs lack an OH group required for binding two nucleotides together, they cause DNA polymerase to terminate elongation. In the classical approach, this reaction is carried out in four different reaction tubes, each of which only contains one of the four ddNTPs. The fragments from each of the four reaction tubes are then denatured and size-separated via gel electrophoresis. By reading the order of the DNA bands on the gel image, starting at the shortest fragment and ending with the longest, the sequence of the template DNA can be decoded.

The introduction of *polymerase chain reactions*, *PCR*, has since simplified and automated Sanger sequencing. For instance, the use of dye-labelled ddNTPs, in which each of the four ddNTPs emits a different colour signal, allows researchers to conduct sequencing in a single reaction. This method has become the main approach in automated sequencing.

Modern Sanger sequencing can be used for sequences of up to 900 base pairs and has been playing a big role in genetic research. It has the advantage of being an accurate method and was the most frequently used DNA sequencing method for about four decades.

However, in more recent years, the advent of so-called *Next Generation Sequencing*, *NGS* techniques has revolutionised the field of genetics. These techniques allow for the cheap and fast determination of the entire human exome – the protein-coding part of the genome – or genome.

## 1.3 Genetic disease and inheritance

Genetic diseases are disorders which are at least in part caused by disease-causing variants in the DNA sequence. While the large majority of genetic alterations are completely harmless, some of them can cause disease or increase the likelihood for the development of disorders. Frequent harmless alterations are termed *polymorphisms* and are naturally occurring variants of which every individual harbours many. If an alteration is known to cause a disease, it can be described by various terms: Well-accepted descriptions are *harmful variant* or *disease-causing alteration*, as well as *mutation*, or combinations of these terms. To distinguish between harmless and harmful alterations, I attempt to make it clear in the context by adding explanations such as harmless, deleterious, or disease-causing.

Deleterious genetic alterations can lead to or influence the likelihood of developing a disease in various ways. In some cases, one single harmful variant can be enough to cause a disease, in others the disease mechanisms are more complex. In the following sections I will provide an overview of different genetic disease mechanisms, with a focus on Mendelian disorders due to their relevance for this thesis.

### 1.3.1 Complex diseases

In so-called *complex diseases*, a combination of several DNA variants increases the likelihood of an individual to develop a certain disease. Examples are widespread and include many civilisation disorders such as cancer, diabetes or cardiovascular problems; diseases which affect a large number of patients at some point in their lifetime. While complex disorders are not exclusively genetic – environmental factors such as diet and lifestyle also play a big role – certain variants are known to increase the likelihood of suffering from a complex disease. For the development of cancer or other complex diseases, one deleterious variant is usually not enough to trigger the onset of the disease. Instead, a

number of variants increase the probability for disease development in an incremental way, with each variant's effect contributing with differing effect sizes.

In familial cancers, the effect size of a variant is strongest: an individual is more likely to suffer from a certain type of cancer at some point in their life if close relatives have been affected by it and have passed on the alteration. As these variants are rare and have a clearly deleterious effect by destroying the protein function, they are often referred to as disease mutations even though they do not directly cause disease. Instead, environmental factors or additional mutations do still play a role in hereditary cancers.

An example of a familial cancer that made headlines in recent years was the case of Hollywood actress Angelina Jolie, who decided to undergo double mastectomy in 2013 and removal of her ovaries in 2015 after finding out that she carries a mutation in the *BRCA1* gene. The protein produced by this gene plays a crucial role in DNA repair. Hence, Jolie's *BRCA1* mutation, combined with a strong family history of breast and ovarian cancer, was estimated to increase her susceptibility to breast cancer by over 80% and to ovarian cancer by 50%. Angelina Jolie's example has led to an increase in *BRCA1* testing [15], which in turn caused debate on the risk of unjustified genetic testing [16, 17].

However, for many cancer patients neither a familial predisposition nor a lifestyle link can be established. Frequently in those cases, mutagens or even copying errors during cell replication lead to somatic mutations. These alterations only occur in a subset of cells rather than the entire body. Most somatic mutations are harmless, but in some cases they can lead to the development of certain cancers. In fact, this mechanism, which can best be summed up as 'bad luck', has recently been found to be a leading cause of non-familial cancers [18].

Complex diseases show great heterogeneity in severity, age of onset, influence of genetic and environmental factors. Hence, it is a real challenge to establish a link between a genetic variant and the onset of a complex disease – each factor only contributes with such a small effect size that it is extremely hard to pinpoint where things went wrong for the patient.

### 1.3.2 Mendelian disorders

In contrast to complex diseases with their myriad of contributing factors, in some cases a single damaging variant in a single gene can be enough to cause a severe genetic disorder. These disorders are termed *monogenic*, *rare*, or *Mendelian disorders* and affect a large number of individuals worldwide. Even though each single disease is rare, in 2015 over

7,000 such disorders were known and approximately 300 new diseases were estimated to be added each year [19]. Hence, the total number of affected patients lies in the millions. While dedicated statistics on affected individuals are difficult to come by, 6 to 8%, or 25-36 million patients are estimated to suffer from a rare disease in the European Union at some point in their lives<sup>1</sup>. It has to be noted that this number includes non-genetic rare diseases as well. However, it demonstrates strongly that 'rare' diseases as a whole are far from rare.

Many Mendelian disorders manifest in early childhood. They often have severe consequences and pose major burdens on affected families. Examples of early-onset Mendelian disorders include Cystic Fibrosis, Sickle Cell Anemia, or Phenylketonuria; diseases which do not only drastically impact life-quality but often lead to premature death. As an exception to this pattern, a number of rare genetic disorders appear later in life, such as Huntington's Disease, which usually manifests between 35 and 44 years of age. Another prominent example of a Mendelian disorder is *Autosomal dominant polycystic kidney disease*, *ADPKD*, a life-threatening disease in which large kidney cysts eventually lead to kidney failure. With a frequency of approximately 1 in 1000, it is one of the most common genetic disorders.

Curing the underlying cause of monogenic diseases requires alteration of the genetic sequence, a procedure termed *gene therapy*. This technology does currently not exist for most disorders. The one example which made headlines recently is the drug Zolgensma, which received approval of the US Food and Drug Administration (FDA) in May 2019<sup>2</sup>. This drug addresses the genetic cause of *spinal muscular atrophy*, mutations in the *SMA1* gene. Incidentally, it is also the most expensive drug ever admitted, at USD 2.1 million per treatment.

In all other cases where gene therapy is not (yet) an option, an early diagnosis can help doctors to treat symptoms and delay or halt some of the debilitating consequences. Many countries, including Germany, have introduced newborn screenings to test for a range of monogenic disorders, enabling diagnosis and potential treatment before the baby's first teeth appear. The case of *Phenylketonuria (PKU)* serves as an example of the importance of early diagnosis.

PKU is a congenital metabolic disease resulting in a decreased metabolism of the amino acid phenylalanine. It was first discovered by the Norwegian doctor Asbjørn Følling in 1934 [20]. When untreated, this disorder leads to intellectual disability, seizures, mental disorders and behavioural issues. It can be treated by maintaining a strict diet avoiding

<sup>1</sup>[https://ki.se/sites/default/files/council\\_recommendation\\_on\\_action\\_in\\_the\\_field\\_of\\_rare\\_diseases\\_0.pdf](https://ki.se/sites/default/files/council_recommendation_on_action_in_the_field_of_rare_diseases_0.pdf), accessed 29.12.2018

<sup>2</sup><https://www.fda.gov/vaccines-blood-biologics/zolgensma>, accessed 17.06.2019

uptake of phenylalanine, and when treated in this way from an early age, babies born with PKU can grow up healthy and reach a normal life span. Due to the importance of an early treatment, in many countries newborns are routinely screened for PKU at a few days' age. In Germany, a nation-wide test for PKU was introduced in the late 1960s, identifying affected babies at an extremely young age and allowing for optimal treatment.

However, disease management is not the only argument for early diagnosis: When a baby with a congenital disease is born, this has a strong impact on the affected families. The birth – and sometimes early death – of a baby with a congenital disease poses a strain on the mental health of the parents, who often struggle with feelings of guilt and responsibility. A molecular diagnosis is of great importance for parents and patients alike and helps them to better come to terms with the situation [21, 22]. Moreover, it allows for the assessment of the disease risk for future children by observing the inheritance pattern of the disease and by offering prenatal tests to affected families.

### 1.3.3 Inheritance patterns

As genetic diseases, monogenic disorders can be inherited from generation to generation. In their voyage through the generations, they follow certain patterns which are governed by *Mendel's laws*. By counting traits in pea plants, the Moravian monk Gregor Mendel (1822-1884) determined the rules underlying inheritance. Mendel observed an organism's phenotype – characteristics visible to the outside, such as traits or behaviours – to draw conclusions on the underlying genotype – the genetic identity that determines a certain, observable trait. From his experiments, which were largely ignored by scientists for 30 years and rediscovered in the early 20th century, the various modes of inheritance could be derived. These rules are determined by the organisation of the human genome and allow for categorisation of the many different Mendelian disorders.

The human genome is arranged in 46 chromosomes: one maternal and one paternal set of 22 *autosomes* (non-sex-linked chromosomes) and the two *allosomes* (sex chromosomes, XX for females and XY for males). Hence, every human carries two copies of each autosomal gene, one of which is inherited from the mother, and one from the father. These two versions of a gene are termed *alleles*.

For sex chromosomes, the matter is slightly different: Males only have one copy of genes located on the X-chromosome, which they inherit from their mother. Females, on the other hand, do not carry a Y-chromosome at all. In addition, large parts of one of the X chromosome are inactivated at random in each cell in females. This mechanism termed *random X inactivation* offsets the higher genetic load in females.

A genetic trait – and hence a Mendelian disorder – can be inherited in different ways: If a single alteration is enough to cause it, it is inherited in a *dominant* fashion and the presence of one disease allele (*heterozygous genotype*) as well as two disease alleles (*homozygous genotype*) will lead to disease. This is often the case for mutations that increase the function of the protein, so-called *gain of function (GOF) mutations*. In this case, if the trait is fully penetrant, every individual who carries the disease allele will develop the disease. However, GOF variants with a strong effect are subject to a high selective pressure. Thus, affected individuals born with a GOF mutation often do not survive, which usually prevents these mutations from manifesting in family pedigrees. Instead, GOF variants tend to appear newly in an individual as so-called *de novo mutations* or in late-onset diseases such as ADPKD.

A special type of dominant inheritance occurs with *dominant-negative* mutations, which lead to a gene product with an antagonistic function to the healthy allele. An example is Marfan syndrome, which is caused by mutations in the *FBN1* gene.

In the opposite case, the case of *loss-of-function (LOF)* alterations, a protein's function is reduced or completely abolished. In this case, two defective alleles are required for the manifestation of a disease as the remaining healthy allele is often still able to maintain function – therefore, two disease-causing alterations have to be present. Heterozygous individuals who carry only one copy of the disease allele are usually healthy and termed *carriers*. This mode of inheritance is called *recessive* and the disease allele has to be present in a *homozygous* fashion for the disease to manifest. If the second disease mutation necessary for the manifestation of a recessive disorder is not identical to the first mutation (but, for instance, present at a different location in the same gene), the genotype is termed *compound heterozygous*. Dominant and recessive inheritance patterns can be linked with autosomes or allosomes, resulting in four main modes of inheritance: autosomal recessive, autosomal dominant, allosomal recessive, and allosomal dominant. Figure 1.1 displays a simplified overview of example phenotypes and resulting genotypes.

genotype		phenotype
heterozygous	ATTGCCGTGCAAG <b>C</b> CGTGCATAGTACGGTGACCTGAT ATTGCCGTGCAAGTCGTGCATAGTACGGTGACCTGAT	affected in dominant MoI
compound heterozygous	ATTGCCGTGCAAG <b>C</b> CGTGCATAGTACGGTGACCTGAT ATTGCCGTGCAAGTCGTGCATAGTACGGT <b>A</b> ACCTGAT	affected in dominant & recessive MoI
homozygous	ATTGCCGTGCAAG <b>C</b> CGTGCATAGTACGGTGACCTGAT ATTGCCGTGCAAG <b>C</b> CGTGCATAGTACGGTGACCTGAT	affected in dominant & recessive MoI

FIGURE 1.1: **Overview of genotypes and resulting phenotypes.** Displays potential genotypes and resulting phenotypes depending on mode of inheritance (MoI). Harmful alteration indicated in bold and red.

In autosomal modes of inheritance, both genders have the same probability of suffering from a genetic disease. For allosomal inheritance, however, some important differences exist between the genders: As males carry only one copy of each gene located on the X chromosome, they will be affected by X-linked recessive disorders when they inherit only one disease allele. Affected males receive their X chromosome carrying a disease mutation from their mother, who often is unaffected by the disease. Moreover, carrier females can express an X-linked recessive disorder in varying degrees due to the aforementioned random X-chromosome inactivation.

Another mode of inheritance plays a role in a subset of genetic disorders: mitochondrial inheritance. Mitochondria, the powerhouses of a cell, contain a small circular genome that encodes 13 protein-coding genes. Diseases linked with mitochondria are termed *mitochondriopathies*. Although the majority of mitochondriopathies are due to disease mutations in the nucleic DNA, mutations in genes located in the mitochondria can lead to disorders such as Leigh syndrome or mitochondrial myopathies. These disorders show a distinct inheritance pattern: Mitochondria are inherited almost exclusively in a maternal fashion, leading to a pattern that mimics autosomal inheritance as both genders can be affected equally. However, this picture can be warped by *heteroplasmy*, the presence of several mtDNAs in a single cell [23]: Human cells contain hundreds of mitochondria in which the individual mtDNA molecules can be slightly different, with only some of them being affected by a given alteration. Depending on how many of the inherited mtDNA molecules do not carry the disease allele, the offspring might or might not be affected by the disease. These processes are determined by chance during cell division and development [24, 25].

Monogenic disorders can be transmitted following all modes of inheritance introduced above. However, recessive autosomal inheritance is the most common mode. This is due to the fact that the selective pressure towards recessive alterations is not as strong as for dominant ones: Recessive traits can be passed on through healthy carriers for generations and only manifest in homozygous individuals. In consanguineous families, where an individual's ancestors are related, recessive disorders manifest more frequently as the disease-causing variant has a higher likelihood of being inherited both paternally and maternally.

## 1.4 Finding disease causes in Mendelian disorders

The classical approach to detecting the cause of genetic disorders consists of a complex procedure of various genetic tests: First, candidate regions are determined via *linkage analysis*, a method to find genetic markers which are inherited together with the disease



phenotype (or *co-segregate*) in an affected family. Genetic markers are genes or genetic sequences whose chromosomal location is known and which can thus be used to find out where a disease gene is located.

Second, the physician compiles a list of candidate genes located within those regions that are most likely to be linked with the disorder based on what is known about their function. In a third step, the coding sequence of these candidates is then sequenced in the patients, and, provided the discovery of potential disease mutations, in their relatives and controls from the same population. Finally, this array is then usually concluded by functional investigations, or – the gold standard – an animal model to determine the molecular relevance of a putative pathological alteration.

While this approach has been the standard for decades now, it is both time-consuming and expensive due to the multi-step set-up. In addition, it is only an option in large families or in cases where many families are afflicted as a number of affected and unaffected members are needed for linkage analysis.

In recent years however, thanks to the advent of so-called *Next Generation Sequencing*, *NGS* techniques, new analysis methods have taken over. While the availability of affected or unaffected relatives helps in elucidating disease causes, it is not a prerequisite for NGS.

Falling sequencing costs and recent advances in NGS methods not only sparked large-scale research projects such as GenomicsEngland's 100,000 Genomes project [26], but also the identification of connections between genes or mutations and disease. This has led to a wealth of knowledge – but also to a large amount of data which has to be sifted and analysed; a task in which we depend largely on computers.

## 1.4.1 Next Generation Sequencing and bioinformatics

### 1.4.1.1 Sequencing

DNA sequencing allows us to read the contents of the genome in order to find 'spelling errors', i.e. mutations relevant to genetic diseases. Depending on the disease in question, the availability of candidate genes, and healthy or affected relatives to be sequenced in parallel, several NGS sequencing strategies are possible:

*Panel Sequencing* refers to the assessment of a number of candidate genes known to be involved in certain diseases. These target genes are enriched in the sequencing process by capturing and isolating them, a step which requires heavy optimisation. Panel sequencing is used if a patient's symptoms point towards a specific disease or a group of diseases,

and it has been widely used in the past (also in combination with Sanger sequencing). While well-established panels are still widely used, due to dropping sequencing costs, improvements in data analyses and the complicated optimisation of the enrichment step, more comprehensive sequencing methods are now taking over [27, 28].

*Target-enriched sequencing* allows the sequencing of large genomic target areas. Thus, researchers can decide to target a subsection of the genome or a subset of genes. Selected regions are hybridised to target-specific probes, which can then be isolated, amplified and sequenced. One form of target-enriched sequencing is *Whole Exome Sequencing*, *WES*, in which the entire exome, the protein-coding part of the human genome, is analysed. The exome consists of only a small percentage of the genome – roughly 1% – but is considered to contain most of the known disease mutations [29]. In exome sequencing, the target regions have to be captured and enriched. Because it is currently not possible to evenly capture all target regions, WES has an inherent level of uncertainty. Depending on which exome version is used – for each reference genome as described in section 1.4.1.2, various exome versions exist – sequencing results may differ. Nevertheless, its advantages outweigh the costs. The first diagnosis of a Mendelian disorder using WES was achieved in 2010 [30] (also see section 1.2) and nowadays, WES is frequently used and currently considered the most cost-effective method of genetic analysis in clinical and research settings [27].

However, *Whole Genome Sequencing (WGS)*, the analysis of the entire genome, is steadily gaining ground as it is the most comprehensive sequencing approach, does not require any enrichment step and is hence considered to be more powerful than WES in variant detection [28, 31]. While WGS costs have been prohibitive in the past, falling costs and technological advances have led to an increased usage of WGS [32]. The first human genome ever sequenced – the Human Genome Project – was billed at USD 500 million to USD 1 billion. Nowadays, in a range of recent studies, WGS costs were found to lie between USD 1,906 and USD 24,810 per test, in comparison to USD 555 and USD 5,169 for WES studies [33] in different countries. As data analysis methods improve, increased demand is expected to drop the costs even further, leading to the growing importance of WGS.

Nowadays, most large-scale projects are conducted using NGS methods. Despite these advances and changes, the 'old-school-method', Sanger sequencing, still remains used for smaller projects and the validation of NGS results.

In NGS, it is no longer the sequencing step that is the main limiting factor, but the data processing: As each sequencing run generates millions of reads and tens of thousands (WES) or millions (WGS) of variants, it is a major struggle to make sense of this mountain of data. Determining the disease-causing variant in rare diseases is often described

with the metaphor of finding a needle in a haystack. Bioinformatic methods are indispensable in this task. In the following sections, I will give an insight into the various steps necessary to reach a meaningful understanding of the genetic variation found in humans .

#### 1.4.1.2 Data processing

##### Alignment

For Mendelian disorders, the goal of sequencing is to find the causal mutation(s). In order to achieve this, the raw fragment reads determined by NGS have to be aligned to a human *reference genome*. These reference genomes are compiled from the sequences of different humans and maintained by the Genome Reference Consortium (GRC). Until genome version GRCh37, the version before the latest, there has been an attempt to list the more common variant as the reference allele in cases of polymorphism. The current version, GRCh38, was published in 2013 and offers alternate sequences for genomic regions known to be highly variable.

Many secondary sources and applications still use the GRCh37. The human reference genome can be accessed at different sites, such as Ensembl [34] or the UCSC Genome Browser [35].

Various algorithms exist for alignment of NGS fragments to a reference genome, with new alternatives being developed constantly. The choice of algorithm depends on factors such as run-time, accuracy, and – last but not least – the researchers' familiarity with a certain tool. Frequently used algorithms in clinical research settings are BWA [36], one of the oldest – but still most common – options, and Bowtie [37].

##### Variant Calling

After mapping, the resulting data has to be scanned for variations, i.e. deviations from the reference sequence. This step allows researchers to identify various types of genetic alterations: *Single Nucleotide Variants*, *SNVs*, are changes of a single base pair – at a certain position in the reference genome, the base A might be present, whereas the patient's sequenced genome shows a G. They are also the most common type of genetic variation. Other, more complex types, are *insertions or deletions*, *InDels*, where one or more bases are inserted additionally to or deleted from the genomic sequence. Structural variants are larger alterations which span 1000 bases or more and can hardly be detected using WES. They include *inversions*, the flipping of a genomic sequence, *translocations*, its shifting to a different location, and *copy number variants*, *CNVs*, which are defined as DNA segments of one kilobase or larger that are present at a variable copy number when compared against a reference genome [38]. Technically, CNVs are large InDels, but due

to their size are treated separately from those smaller scale alterations. The detection of large structural alterations is where WGS excels, as it is capable of detecting even large structural aberrations spanning many genes, independent of whether the break point is located in an exon or not.

Various algorithms tailored to the analysis of different variant types exist. As SNVs and small InDels are the most common variants and the easiest to detect, these are usually the first ones to be investigated and thus the most relevant for my thesis; I will hence examine and elucidate them in more detail in the following sections.

During the variant calling step, it can also be determined whether an individual's genotype is homo- or heterozygous at a given location, a process termed *genotyping*. When attempting to determine the disease-causing mutation(s), the genotype provides valuable information that helps to reduce the number of candidates.

Short variants (single nucleotide variants, insertions and deletions) are usually exchanged in *variant call format (VCF)* files. A VCF file lists the genomic location – the chromosome and base position – for each variant found in an individual, combined with additional information<sup>3</sup> such as allele counts, reading depths, quality scores or, if available, the genotype. Subsequent analyses – the search for the disease causing mutation – are then carried out on the VCF files.

### 1.4.1.3 Variant annotation

Variability leads to a large number of variants detected in a patient: WES analyses usually yield tens of thousands of variants, whereas for WGS, this number lies in the millions – on average 3 to 4 million variants are found in a single WGS run [13]. Most of these variants are completely harmless, and to distinguish the harmful from the harmless is not a trivial task. Therefore, in a first step, these variants have to be annotated with information on their disease potential. As manual curation of these vast amounts of data is not an option, computer tools have to be employed for this task. These programs assess the disease-causing potential of a candidate variant by linking it with known biological data: Depending on where exactly an alteration is located and how it might alter the gene product, its effect can vary widely from being completely harmless to having a devastating impact on a patient's life.

The protein-coding part of the genome has been studied quite extensively in the last decades, so the scientific community has a lot of information about the potential impact of a genetic alteration within a coding sequence. For instance, it is easily conceivable

---

<sup>3</sup><https://samtools.github.io/hts-specs/vcfv4.2.pdf>, accessed 28.12.2018

that mutations leading to a premature stop codon (see section 1.1.2), thus truncating the protein, are extremely harmful. Other variants can lead to the exchange of an amino acid, which can be harmful or harmless, depending on what this change means for the function of the protein. In other cases, even though there is an alteration in the genetic code, there is no amino acid exchange. These *synonymous* alterations are mostly considered harmless.

Even variants located outside of the coding sequence can have a severe impact on gene function: alterations near splice sites, for example, can lead to heavily altered proteins which might not be functional. Moreover, alterations in untranslated regions can be of disease relevance by having a regulatory impact. Thus, the exact location, and the effect of a mutation on the protein, serve as indicators for the likelihood of a variant in question to cause disease.

A wide range of information on proteins, their function and structure can be found in databases such as Swiss-Prot [39], a manually annotated and reviewed knowledge base of curated protein information, such as protein function and classification.

Other important databases store information on known variants: For example, a large number of variants are already known to be harmless or, in the opposite case, have been found previously to be involved in genetic disease. dbSNP [40] is the most comprehensive example of a repository of known SNVs and InDels in humans which contains harmless alterations as well as known disease-causing mutations. The 1000 Genomes (1000G) project [41] and ExAC [42] collect data from healthy individuals. ExAC, for instance, contains data from over 60,000 individuals who do not suffer from a rare early-onset disorder but who might be carriers of disease alleles.

In general, variants found in ExAC or 1000G are not likely to be involved in the development of severe, early-onset genetic diseases and can be excluded in many cases. For instance, a variant with a frequency of 1% in ExAC can usually be excluded when assessing a disease which appears in 1 in every 3 million cases. In addition, a genome-wide version of ExAC, gnomAD [43], is now available. In contrast to variant databases with a focus on harmless variants, disease mutation sources such as ClinVar [44] or the commercial platform HGMD [45] store data on known disease-causing variants.

However, all variant databases have to be treated with care as they might not be suitable to answer every specific question. For instance, while frequent polymorphisms found in ExAC can be excluded from further analysis, the data also includes not-so-rare recessive disease-implicated alterations (e.g. cystic fibrosis mutations): As healthy individuals might be heterozygous carriers of a recessive disease mutation, one would expect to find these variants in the database, even though they are of clear disease relevance in

homozygous patients. In addition, a patient's 'private' variants – alterations which are harmless but not (yet) listed in any of the databases – cause problems. This is especially problematic for populations that are not covered in the databases (until recently most non-caucasian populations). Thus, it is necessary to not only rely on variant databases but to take additional information into account. For example, the location of a variant within the gene and its effect on the protein product play a role, together with information on evolutionary conservation. More insight into the data sources described above can be found in chapter 2.2.

Sophisticated computer programs are able to pull the information provided in a range of databases together and deliver an estimate for each variant in a VCF file. One of these programs, MutationTaster<sup>4</sup>, was developed in our research group [46, 47] and will be explained in further detail below (see section 2.2).

Some examples of other tools capable of annotating candidate variants or of predicting their disease-causing potential are Poly-Phen2 [48] and SIFT (Sorting Intolerant from Tolerant, [49]), which analyse variants based on sequence homology and the physical properties of amino acids, and can both only annotate coding non-synonymous SNVs. VAAST2 [50], on the other hand, combines the predictions of a number of programs. It offers a greater range of capabilities and can score coding and non-coding variants. Another combination tool is CADD, which integrates a range of annotations into one metric by contrasting variants that survived natural selection with simulated mutations [51].

Different tools draw their conclusions by different means and hence may come to contradicting results. It is known that the capabilities of various tools and their concordance vary widely [52, 53]. To account for this and to allow users to compare results between different software, in recent years a number of tools combining the output of several tools have been developed such as CADD [54], the Variant Effect Predictor (VEP, [55]), and the aforementioned VAAST2 [50].

#### **1.4.1.4 Variant prioritisation: patient information**

As NGS projects deliver large numbers of variants, even a list of previously filtered potentially harmful annotated variants is daunting and has to be committed to further scrutiny. In order to find the causal mutation, the variants have to be prioritised based on additional information: A clinician can determine variants located in disease-relevant genes for a specific case by including patient- or case-specific data such as symptoms, the expected gene function, or candidate genes from gene panels. In combination with

---

<sup>4</sup><http://www.mutationtaster.org>

a variant's effect on the gene product, this information enables powerful filtering or prioritisation of candidate alterations. The more accurate – hence the more personalised – the descriptions are, the more likely it is to detect the real culprit.

This combination of genotype and phenotype allows alterations labelled as 'harmful' that do not fit with the disease in question to be excluded. At the same time, alterations considered less severe by the computer program but which perfectly match the patient's phenotype could become more relevant. In this way, the rather broad categorisations into 'harmful' and 'harmless' become more tailored, allowing a step towards personalised medicine.

### **Phenotype**

A patient suffering from a genetic disease exhibits specific symptoms, which can be identified and classified by their clinician. In the clinic, the entirety of all symptoms observed in a patient is often referred to as their *phenotype*. Taking the phenotype into account can greatly facilitate diagnosis by establishing a link to known disorders, similar diseases, and by suggesting candidate genes.

Phenotyping – the identification of a patient's symptoms and their systematic and thorough documentation and communication – is not a trivial process, especially in diseases or syndromes which can show a high degree of heterogeneity [56, 57] and exhibit multiple symptoms resulting from just one mutation. Therefore, correct phenotyping is crucial for successful diagnostics. Major efforts have been made to categorise symptoms by using controlled vocabulary, ensuring that every expert uses the same terms to describe a given symptom. In a second step, these descriptions have been put into context by organising them in systematic collections, so-called *ontologies* (see section 1.4.1.5 for details).

The Human Phenotype Ontology (HPO) [58], for example, is a systematic collection of disease symptoms observed in patients suffering from (mostly monogenic) genetic diseases, and their connections. The information stored in ontologies such as the HPO not only helps to streamline the complex process of determining the patient's phenotype but also enables computational applications: HPO data can be used by computer programs to calculate and quantify the relationship between symptoms and their relevance.

The phenotypes organised in the HPO have been connected to OMIM (Online Mendelian Inheritance in Man, [59]) and Orphanet [60], large-scale collections of known genetic disorders, their symptoms, known disease genes, and related research. This allows for a systematic, computational assessment of genetic variants, symptoms and their phenotypic relevance.

### **Gene panels**

In many cases, lists of candidate genes, so-called gene panels, are already known for

a certain group of disorders. One well-studied example is the Kingsmore panel [61], a collection of genes known to be involved in rare recessive genetic disorders which manifest in early childhood. It has been revised and curated many times and is used for routine diagnostics worldwide. This is a major aid when assessing the relevance of candidate variants for a particular case. These panels can either be used for enrichment in targeted sequencing (see section 1.2) or as virtual panels to reduce the number of candidates in WES or WGS sequencing projects. Many clinics use in-house panels for various diseases and disease groups. In an attempt to generate a reliable knowledge-base from this wild growth of gene panels, Genomics England's PanelApp<sup>5</sup> is an initiative to generate expert curated gene panels for the scientific community. It stores expert reviewed virtual gene panels for over 200 human disease groups. For example, the PanelApp *Familial dysautonomia* panel currently contains 22 expert-reviewed genes of relevance for the disease, 14 of which are quoted with a high confidence ('green' genes).

Virtual panels such as the ones provided by PanelApp connect the benefits of panel sequencing – lower number of candidates, easier analyses – with the advantages of exome sequencing. Therefore, the restriction to virtual panels in cases with a clear phenotype is a convenient way of improving data analysis. However, this approach cannot detect new disease genes as it only works for genes that are already known to be involved in a certain disease or set of symptoms.

### Gene function

In cases where no mutation in any known disease gene can be found, clinicians have to take other traits into account: Knowledge about function, expression and interactions of a gene or gene product can help to close in on the disease cause. In contrast to relying on known disease-gene links, this approach has the advantage of opening the door to discovering hitherto unknown disease genes. There are a number of data sources that offer varied insights into the functions of and connections between genes and gene products. The Gene Ontology (GO) [62] stores machine readable knowledge on the function of genes and gene products in an ontology. GO data is often compiled from many organisms such as mice or zebra fish and humans. The GO addresses the functionality of genes and their products in a computer readable manner and stores the relations between them. By including this knowledge into the investigation of the disease gene, the search can be restricted to relevant genes, such as the relevant protein class (e.g. ion channel) for a given case.

Another approach is to look at the involvement of genes in molecular or signalling pathways: Resources such as Kyoto Encyclopedia of Genes and Genomes (KEGG [63]), Reactome [64] and WikiPathways [65] store this information. Particularly in cases where

---

<sup>5</sup><https://panelapp.genomicsengland.co.uk/>



laboratory data indicate a defect in a metabolic pathway, incorporating this information into the search for the disease cause can help to identify candidate genes and variants.

A similar approach is feasible in patients who are suffering from a disease limited to certain organs or organ groups. For instance, if a patient is plagued by a genetic disease which manifests itself in the skin, the disease gene might be expected to be expressed in the skin. To collect a list of candidate genes, the inclusion of gene expression data might be a valid option in this case. A large number of experiments determining the expression patterns of genes are conducted in laboratories around the world. Findings from various groups or projects are stored in ExpressionAtlas [66], a manually curated open science resource offering access to data on gene and protein expression.

#### 1.4.1.5 Ontologies

Ontologies are a powerful way of defining the basic concepts of a research domain, as well as the relationships between those concepts [67]. They serve as a valuable reference for researchers and clinicians to search and exchange (biological) data, and they allow information from heterogeneous methods and sources to be merged. In the case of rare diseases and human genetics, two major ontologies shed light on genes and their functions: The aforementioned *Gene Ontology (GO)* and the *Human Phenotype Ontology (HPO)* are valuable resources for variant prioritisation as they allow the evaluation of a candidate gene's relevance for a given disease or group of symptoms.

Originally, the term ontology was (and is still being) used in the field of philosophy. It comes from ancient Greek and describes the *study of existence and being*. In information technology, it has been given a slightly different meaning:

Probably the best known definition of modern-day ontologies was coined in 1995 by Thomas Gruber [68], who identifies an ontology as an *explicit, formal specification of a shared conceptualisation*. This short phrase sums up the core concepts of ontologies: First, the descriptions have to be precise and clear (*explicit*). Second, they store specifications in a machine-readable way (*formal*), and third, there is a *shared* understanding of an abstract concept which is represented by the ontology's *conceptualisation*. This conceptualisation is described in *classes, attributes and relationships* capturing the relevant distinctions in an abstract way while still being as clear as possible about the meaning of the terms. Together, they form a level of data model abstraction and display knowledge about individual terms, their properties, and their relationships between each other. To allow for computational usage, ontologies are specified in standardised languages enabling abstraction from the structures.

Ontologies can be depicted as a graph: Each term (or class) is a node and the relationships between them are edges of the graph. There is usually a loose hierarchy with descendant terms being a more specialised description of their ancestors. However, a term can have more than one ancestor, organising the ontology in a structure termed *directed acyclic graph (DAG)*. Most frequently, terms are equipped with a unique identifier which allows them to be stored and managed independently from a lexical, human-readable description, thus reducing errors. Moreover, this step allows the content stored in a term to be changed or updated later on without altering the ID.

As an example for a DAG, figure 1.2 shows the HPO term *Aplasia/Hypoplasia of the brainstem (HP:0007362)* and its first ancestors and descendants.

The relationships between the terms can be described in different ways, such as *is\_a*, *part\_of*, or *is\_opposite\_of*. In biological ontologies, they are often reduced to *is\_a*, simple class-subclass or ancestor-descendant relationships: In figure 1.2, Hypoplasia *is\_a* subclass of Aplasia/Hypoplasia of the brainstem, which in turn *is\_a* subclass of Abnormality of brainstem morphology. The further we traverse down the graph, the more specific a description becomes. As more and more types of relationships are being added, for instance in the GO (e.g. *negatively\_regulates*) and the HPO (e.g. *is\_opposite\_of*), the graph may contain circles and is hence not a clear DAG anymore. However, for computational methods, *is\_a* and *part\_of* are still the most commonly used relationships.

The structure of ontologies is particularly useful for determining the importance or specificity of a given term:

Based on information theory, Philip Resnik introduced an information-based measure for semantic similarity in the 1990s [69], which is nowadays an accepted method to compare semantic similarities in ontologies and taxonomies. In information theory, the information content of a concept is higher the less abstract it is. This can be expressed mathematically. The information content IC of a concept  $c$  (e.g. a node in the ontology) is measured as:

$$IC(c) = -\log p(c)$$

where  $p(c)$  is the probability of finding  $c$  in a given domain. In an ontology, this probability is usually expressed by the fraction of annotated terms for a concept – which in turn can be expressed as a term’s specificity.

The graph structure of an ontology allows us to determine the *descendants* and *ancestors* that are directly connected to a term. These appear below or above the term in the graph,

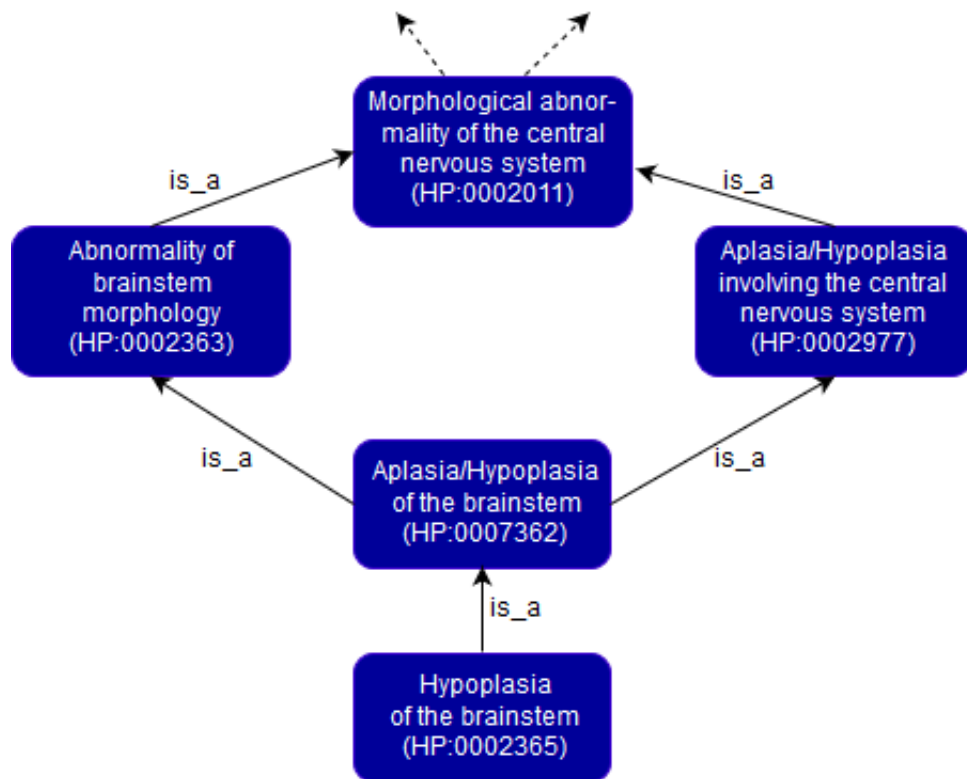


FIGURE 1.2: **Excerpt of the graphical representation of the HPO.** This figure shows the HPO term *Aplasia/Hypoplasia of the brainstem* (HP:0007362) and its relationships to its first ancestors and descendants. HP:0007362 has two different direct ancestors, HP:0002363 and HP:0002977, which it is connected to via an *is\_a* relationship.

and indicate a higher or lower degree of specificity. This concept will be elaborated further in section 4.1.2.1.

## 1.4.2 Current variant prioritisation tools

As computational data analysis is absolutely indispensable to make sense of the masses of results obtained by NGS methods, it comes as no surprise that a large number of computer tools aimed at different user groups have been developed over the last years. I will limit myself to phenotype-driven software for variant prioritisation in WES projects, as these are the most relevant for this work. In this section, I will give an overview of recent programs, compare their capabilities and describe advantages and shortcomings of the various approaches.

As described above (see sections 1.4.1.3 and 1.4.1.4), variant prioritisation generally consists of two steps: The candidate variants are annotated and filtered by severity, and the list of remaining candidates has to be prioritised based on additional information such as the patient's phenotype. Current tools usually combine these steps and offer a

comprehensive analysis of WES data. However, the various software options differ in many ways and cater to different needs. A lot of early solutions were aimed at bioinformaticians and provided their output in scores without any interpretation. Nowadays, however, more and more clinicians prefer to analyse their own data [70], for which they are the main experts, and this calls for software that they can use easily and readily. As extensive computational training is not compatible with a busy working life in the clinic, these users require different software: Their focus is on easy and intuitive tools which allow them to work with their patient's data in a convenient way.

Another distinction between the available tools is the types of data they accept. Earlier tools such as eXtasy [71], Phen-Gen [72], or the Exomiser [73] are largely based on the HPO to characterise the patient's phenotype. In a range of other recent computer programs, there are often more data entry options. Phevor [74] and PhenIX [75] for instance, allow multiple ontologies, whereas ANNOVAR [76] can take various disease-related terms as input. Other software, such as OVA [77], BiERapp [78] or QueryOR [79] are web-based frameworks which allow retroactively refined analyses but are not available without registration (QueryOR, BiERapp). Figure 1.3 displays an overview of recent and widely used web-based tools for the phenotype-based prioritisation of candidate variants.

State-of-the-art variant prioritisation tools are capable of analysing a wide range of data and cover many different cases. However, many of them have still not found their way into routine clinical applications as they are often too complex for clinical use or do not provide enough information for users to draw meaningful conclusions from their predictions. In addition, most of them can only accept non-synonymous SNVs and are limited to nuclear DNA. Another major hurdle for clinical use is file size restrictions, as many tools cannot handle complete VCF files.

A recent paper by Shyr *et al.* [70] stated the importance of usability and easy access for the success of sequencing projects. However, NGS analysis software is still often developed by bioinformaticians without taking clinicians and geneticists on board. Thus, a number of tools are only available as command-line scripts or source code which has to be compiled and installed locally, which is of little use for many clinical applications. A recent example is TAPER [80], a variant prioritisation tool which was published as source code in 2016. Even the installation of software itself can cause an obstacle for clinicians who work on different computers and usually do not have the administrative rights to install software.

The output of many tools can pose another difficulty for daily clinical use: In order to make sense of the results and to be able to draw further conclusions – which might have implications on the treatment of the patient – clinicians and geneticists need comprehensive information. However, most tools – even recent ones – deliver their results in

	PhenIX	Phen-Gen	eXtasy	Exomiser/ PHIVE	OVA	QueryOR	WANNORVAR	BiERapp	Mutation Distiller
latest update	2014	2014	2013	2018	2015	2017	2017	2016	2018
HPO									
OMIM									
Orphanet									
Pathways									
PanelApp									
gene list / region									
expression									
no registration									
gene info provided									
demo/tutorial									

FIGURE 1.3: **Current variant prioritisation tools.** Depicted is an overview of the features of current web-based variant prioritisation tools. Published in Hombach D *et al.* MutationDistiller – user-driven identification of pathogenic DNA variants. NAR Web Server Issue. 2019. doi:10.1093/nar/gkz330

flat tables or files containing a number of scores rather than offering further data on the biological context and the disease relevance of a specific data point, thus limiting their use for clinical applications [70].

Moreover, many tools do not provide hyperlinks to external resources, thus forcing the user to manually search the Internet for further information on their data. While this might seem trivial, it is time-consuming and can prevent users from adopting a given software.

## 1.5 MutationDistiller

To close the gap between the clinic and bioinformatics, and to provide the means for personalised NGS analysis, in the course of my PhD project, my colleagues and I have developed MutationDistiller (<https://www.mutationdistiller.org>), a variant prioritisation tool for use in clinical cases. It was developed in close collaboration with clinicians and human geneticists, taking their needs and requirements on board. As a consequence, the program is freely usable online and does not require any software installation. We aim to make usage as convenient as possible by offering a set of default user modes aimed to fulfil the needs of different user groups. Thus, we aim to provide software which can be used by clinicians, researchers and geneticists without extensive knowledge of bioinformatics. MutationDistiller has already found its way into the clinic and has seen over 14,000 analyses to date.

### 1.5.1 Technical information

MutationDistiller's programmatic structure follows the classical three-tier-structure<sup>6</sup> where the different functions fulfilled by a software – presentation, application processing, and data management – are separated in three layers:

**Presentation tier or User Interface (UI):** This Front End is the interface the user sees. Access is often provided via a web browser.

**Logic tier:** This layer is also termed *application server tier* or *middle tier*. It coordinates the application and contains mechanisms to run the user commands and to return results, thus connecting the other two layers.

**Data tier:** This Back End is where data are stored and retrieved; thus it usually contains a database. It passes the information to the logic tier, from where it will be returned to the user.

The advantage of the three-tier architecture is that the user only needs to have a web browser installed, without the need for additional software on their computer. In addition, for the developers, the three-tier architecture usually has advantages as it is less labour-intensive than distributed software: The entire control over software and data stays with the developers – in case of updates, developers can push these changes to the web-version in one step, without having to worry about distributed versions. This is one of the main reasons why MutationDistiller is available as a web-version only.

---

<sup>6</sup>[https://en.wikipedia.org/wiki/Multitier\\_architecture](https://en.wikipedia.org/wiki/Multitier_architecture), accessed 28.12.2018

MutationDistiller was written in the programming language Perl 5<sup>7</sup>. Dedicated Perl modules, which are discrete software components, contain all the functions necessary to fulfil MutationDistiller’s function. In these modules, functions are sorted and grouped according to their purpose. The modules reference and call each other, access the database and connect with the Front End.

MutationDistiller combines the powers of two tools previously generated in our research group: MutationTaster (<http://www.mutationtaster.org>, [46]), a variant effect predictor, and GeneDistiller (<http://www.genedistiller.org>, [81]), a gene ranking tool – hence the name MutationDistiller. In the next two sections, I will briefly introduce those tools and their capabilities.

## 1.5.2 MutationTaster

The variant effect predictor MutationTaster started in 2008 [47] and is now freely available online in its second version [46] at <http://www.mutationtaster.org>. It is able to predict whether a variant is most likely harmful or harmless. While users can also manually enter individual alterations, in the NGS age the tool is mainly working on entire VCF files. MutationTaster conducts *in silico* tests and employs a Naïve Bayes classifier to distinguish deleterious mutations from harmless variants: Each variant is sorted into either ‘harmless’ or ‘harmful’. The tool can handle coding and non-coding alterations, SNVs as well as short InDels. Moreover, it is not limited to protein-coding regions but can also annotate alterations located in introns and the untranslated regions.

For each variant, MutationTaster has four different prediction options: *Disease causing* indicates that the tool’s Naïve Bayes classifier found enough evidence to consider a given variant to be harmful. Variants causing frame-shifts and leading to nonsense-mediated decay, or that are listed as ‘pathogenic’ in ClinVar, are labelled *disease causing automatic* whereas variants known to be harmless from databases such as 1000G or ExAC are labelled *polymorphism automatic*. Finally, the *polymorphism* label denotes variants that the classifiers considers harmless.

When users upload their data to MutationDistiller, the program sends the information to MutationTaster to determine the pathogenicity of the submitted variants. In a second step, the variants (or the genes those variants are located in) are then ranked and excluded according to the user’s phenotype entries.

---

<sup>7</sup><https://www.perl.org/about.html>, accessed 27.12.2018

### 1.5.3 GeneDistiller

The gene-ranking part of MutationDistiller is based on GeneDistiller (<http://www.genedistiller.org>) which was published by our research group in 2008 [81]. GeneDistiller allows lists of candidate genes to be ranked according to how well they match a myriad of user-defined criteria. GeneDistiller takes regions from linkage intervals or simple gene lists as input and is also able to conduct whole genome or mitochondrial genome analyses. Users can filter for and highlight genes fulfilling a number of criteria such as cellular localisation, expression levels or phenotypes. Moreover, the tool prioritises the gene lists according to user-defined criteria and weights. Users can also compare their target genes to genes that show similar expression patterns or interactions. However, we found that the multitude of options together with a crowded user interface can overwhelm users and make analyses with GeneDistiller cumbersome. Hence, in MutationDistiller, we have not only added new resources but also adapted the tool to the requirements of NGS projects, and trimmed the user interface as well as the underlying algorithm to provide a user-friendly tool.

### 1.5.4 Combining genotype and phenotype

By connecting the powers of MutationTaster and GeneDistiller, MutationDistiller combines a patient's genotype with their phenotype. For the genotype, variants from panel sequencing, WES, or even WGS studies can be uploaded in VCF format. The phenotype can be entered in a multitude of ways: MutationDistiller accepts common ontology terms such as HPO symptoms or GO terms, diagnoses as OMIM and Orphanet entries, identifiers for molecular pathways (WikiPathways, Reactome), and expression data (ExpressionAtlas). Candidate genes can be entered manually or as panels via Genomic England's PanelApp.

In the output, MutationDistiller displays information about a variant and the gene it is located in on one page: In a summary table, the top variants are listed together with crucial information such as gene symbols, known diseases caused by mutations in this gene, and genotype occurrences in 1000G and ExAC, as well as coverage and compound heterozygosity. Further data on each gene is listed below, offering a comprehensive overview of each candidate gene and variant and their relevance for the specific case.

If the causative variant cannot be determined by an initial search, MutationDistiller offers the option to refine the query by adding or removing terms which have come up in the meantime. Thus, the program allows users to customise the hunt for the culprit in an iterative way.



In the following chapters, I provide an overview of the integrated data sources and structure of the database. Moreover, further information on the program is given, such as technical data, the scoring mechanism and a comprehensive description of the input and output options. I will explain the development and validation steps and describe use cases for MutationDistiller. Finally, I will give an outlook of future developments.

## 2 MutationDistiller: Data integration

In order to accomplish large data-driven projects, such as the prioritisation of the myriad of candidate variants generated in WES or WGS projects, huge amounts of data from various sources have to be brought into context. This process of assembling data from heterogeneous backgrounds and sources in one framework is termed *data integration*. Two main ways of integrating data exist:

**Uniform Data Access** or **Virtual Integration** keeps the data in their various source systems and provides access to them directly during the data query process (usually via the Internet). As such, in each query, the data are gathered together and the output is only saved for a short amount of time. A main advantage of this approach is that no additional hardware needs to be provided for storage of the information. Moreover, there is no delay in the uptake of data updates from the source system. However, this comes with a loss of control: No version management is feasible and no control over the data structure is given. In addition, updates can cause severe problems, especially if the database structure becomes altered. Bandwidth limitations can also cause issues, in particular with large data sets. Server failures or, even worse, the potential abandonment of servers may let data queries run dry.

**Physical Integration** or **Common Data Integration** on the other hand refers to the creation of a new system which stores a copy of the data from the source systems. The data can be stored and managed independently in a Data Warehouse. One example of this approach is the Ensembl Genome Browser [34] (see section 2.2). A disadvantage of this solution is that a system to store and handle the source data has to be provided. Moreover, updates to the source data will have to be manually kept up with.

However, physical integration comes with a number of advantages: First, it allows for flexible data management and the combination of data from heterogeneous sources and in different formats. In addition, the function of the data system is independent from the source system, generating better stability. Data updates can be planned and organised while the data can be checked for validity more easily than is the case for externally stored data. Finally, for the usage of the software, physical integration offers a major advantage as well: The run-time of the program will be reduced in comparison to virtual integration as locally stored data can be queried much more quickly. Due to these advantages, we decided to physically integrate the data sources used by MutationDistiller. In the following sections, I will give an overview of databases in general, as well as the sources used by MutationDistiller and their integration into the tool.

## 2.1 Databases

In this data-driven day and age, databases are a popular way of storing and managing data. Databases are persistent repositories stored in a computer system; meaning that the data is supposed to be available even after the software application using or creating it is closed. A single database table can be compared to a simple spreadsheet table. The real strength of modern data repositories is provided by *relational database management systems*, *RDBMSs*. These are software programs specifically designed to hold the data of related repositories. RDBMSs store relational databases, which contain their data in collections of *relations*, or *database tables*. A relation is defined as set of *tuples* – or database entries – belonging to a given data domain. A database's tables can therefore be seen as permanently stored relations. In these tables, columns represent properties (or *attributes*) while rows hold the values for these properties.

All related tables are held together in one or several *database schemas*, which contain not only the tables but also the connections between them.

RDBMSs offer a number of advantages: First of all, they are optimised for large amounts of data, and therefore handle them with great speed (*scalability*). Moreover, several users can access the database simultaneously through standardised interfaces, making queries secure and convenient. Finally, the information stored in different tables is usually related, making it quick and easy to cross-reference data. *Transaction control* ensures that a database query that accesses different tables but belongs to one logical task will either be concluded in its entirety or not at all, without allowing simultaneous write-access. An example of transaction control is the transfer of money from one account to another: Taking money out from one account without putting it safely into the destination account would not make sense (and probably have the bank lose their customers within no time). Thus, the transaction will only be concluded all-together – or not at all. In addition, transaction control is pivotal for multi-user tables by ensuring that only one user can modify the same data at a given time.

To facilitate one of the main advantages of databases, fast and convenient access to the data, database indices exist. An *index*, like the index of this thesis, is an ordered list of the values of one or more attributes stored in a relation, and allows each entry to be found faster without having to search the entire relation. This is particularly important for large tables as it can speed up the search process considerably. Indices can combine multiple different attributes and each table can hold many different indices.

## 2.2 Data sources

### 2.2.1 MutationTaster predictions

To assess the pathogenicity of a variant, MutationDistiller relies on the predictions generated by a functional prediction tool developed in our group, MutationTaster [46, 47] (see 1.5.2 for more information). MutationTaster employs a Naïve Bayes classifier to predict a variant’s likelihood to be harmful. To do this, it relies on a number of data sources itself. A list of MutationTaster’s data sources can be found in table 2.1. Some of the data integrated into MutationTaster are also available from MutationDistiller directly and will be described in further detail below.

Data source/tool	Description
ENCODE project [82]	Encyclopedia of DNA Elements; repository of functional elements of the human genome ( <a href="https://www.encodeproject.org/">https://www.encodeproject.org/</a> )
Ensembl [34]	Central, freely available data warehouse of genome data for various species, including human and mice ( <a href="https://www.ensembl.org/index.html">https://www.ensembl.org/index.html</a> )
1000G [41]	1000 Genomes Project; public catalogue of human variation and genotype data ( <a href="http://www.internationalgenome.org/">http://www.internationalgenome.org/</a> )
dbSNP [40]	Collection of simple genetic polymorphisms ( <a href="https://www.ncbi.nlm.nih.gov/projects/SNP/">https://www.ncbi.nlm.nih.gov/projects/SNP/</a> )
ClinVar [44]	Aggregated information on human variation and its connection to disease ( <a href="https://www.ncbi.nlm.nih.gov/clinvar/">https://www.ncbi.nlm.nih.gov/clinvar/</a> )
Entrez Gene [83]	Integrated gene information on a wide range of species ( <a href="https://www.ncbi.nlm.nih.gov/gene">https://www.ncbi.nlm.nih.gov/gene</a> )
ExAC [42]	Exome Aggregation Consortium. We include ExAC genotype counts and loss-of-function scores ( <a href="http://exac.broadinstitute.org/">http://exac.broadinstitute.org/</a> )
Grantham Matrix [84]	Formula for differences between amino acids
PhyloP [85], PhastCons [86]	Computer programs to predict the evolutionary conservation of a given nucleotide
UniProtKB [39]	Database of protein sequences with annotations ( <a href="http://www.uniprot.org/">www.uniprot.org/</a> )
HGMD public [45]	Human Gene Mutation Database; non-redundant collection of disease-relevant DNA alterations ( <a href="http://www.hgmd.cf.ac.uk/ac/index.php">http://www.hgmd.cf.ac.uk/ac/index.php</a> )
* bl2seq [87]	Tool for the alignment of DNA sequences
* MaxEntScan [88]	Human splice site prediction tool
* Polyadq [89]	Tool for the detection of human polyadenylation sites

TABLE 2.1: **MutationTaster data and tools integrated into MutationDistiller.** This table lists the various data sources and tools that are used by both MutationTaster and MutationDistiller. Tools are marked with an asterisk. Please note that in the original MutationTaster version, the splice site tool was nnsplICE [90], which was replaced by MaxEntScan in later versions. Details on MutationTaster data sources and integrated tools can be found in [46, 47].

## 2.2.2 Genetic data

MutationDistiller, similar to its inspiration GeneDistiller, uses a gene-centric approach to integrate the many different data sources that come together within the software. In order to combine and connect these sources, we use the gene level as a mediating point and map or link all information to Entrez gene IDs provided by the US National Center for Biotechnology Information (NCBI) [83] and/or gene identifiers from Ensembl (ENSG) [34]. These identifiers define each protein-coding gene individually and hence enable us to cross-reference data. In MutationDistiller, this allows us to provide prioritised lists of variants and their connected genes based on a myriad of options.

### 2.2.2.1 Ensembl

Ensembl [34] is a joint project between the European Bioinformatics Institute (EMBL-EBI) and the Wellcome Trust Sanger Institute which began in 1999 to automatically annotate genome data. In its Genome Browser<sup>1</sup>, it provides access to genome annotations for multiple vertebrate species. Since its online launch in 2000, it has grown into a central, open resource for genome information which is used by many researchers from various fields. Its core component is the Ensembl Genes database which currently provides genome data and annotations for 135 mainly vertebrate species. The information content varies between species, with data for humans and model organisms such as mice and zebrafish being the most extensive. The Genome Browser provides convenient access to the data and visualises gene information, genetic sequences and annotations on the web. For data annotation, Ensembl relies on an automated process in which annotations of transcripts are based on experimental evidence: The automated pipeline uses mRNAs and protein sequences from public databases such as the European Nucleotide Archive ENA at EMBL-EBI, UniProtKB, or NCBI RefSeq. Moreover, transcript annotation data may be derived from other sources such as the Havana/Vega set [91] and the Consensus Coding Sequence (CCDS project, [92]), a collaborative project providing an overview of protein-coding regions with identical annotations for humans and mice. The Genome Browser is particularly suited for single search requests. For large-scale queries, Ensembl data is available in various ways. Data from the database can be downloaded or queried dynamically for virtual data integration. In addition, BioMart [93] offers access to Ensembl data sets. Expression and protein data obtained from Ensembl are described in sections 2.2.5.2 and 2.2.6, respectively.

Ensembl data used in MutationDistiller is mainly accessed through MutationTaster's prediction results. In addition, MutationTaster and MutationDistiller use the protein

---

<sup>1</sup><https://www.ensembl.org/index.html>

repository UniProt/Swiss-Prot hosted by Ensembl. Currently, MutationTaster is based on Ensembl build 37.

### 2.2.2.2 Entrez Gene

Entrez Gene is a genome database hosted and run by the NCBI. It offers a web interface and download providing easy access to gene-specific information. Entrez Gene stores a large array of data and provides information for specific transcripts. For many transcripts, a direct mapping to Ensembl Transcripts is available. Wherever possible, this is used by both MutationTaster and MutationDistiller. MutationDistiller also uses additional data available from Entrez to provide further information on a gene, such as Entrez Synonyms, which accounts for the fact that a number of genes are known under different names or abbreviations, or Gene Positions, which determines the genomic locations of a gene. Moreover, Entrez genes are linked to NCBI GeneRIFs, myriads of tweet-like explanations on the function of a particular gene (max. 255 characters). They are associated with a specific Entrez Gene database entry and link to a scientific publication supporting GeneRIFs. We downloaded GeneRIFs and display them in MutationDistiller to allow users to get a quick insight into a gene's function and relevance. Moreover, we use the Entrez gene identifier to link between Ensembl and NCBI data for each gene.

## 2.2.3 Variant databases

Variant databases allow to assess the relevance of a candidate variant and to put it into context. Thanks to previous research and their occurrence in healthy individuals, many variants are already known to be harmless, whereas others have been found to be involved in genetic disease. Thus, by using the information stored in those databases, a large number of known harmless variants in a WES project can be excluded from the candidate list, while known disease alterations will have to be considered with greater care.

### 2.2.3.1 1000 Genomes Project

The 1000 Genomes Project (1000G, [41]) ran between 2008 and 2015 with the goal to find the majority of genetic variants with a frequency of at least 1% in the populations studied. Data generated by the 1000G project has been made available to research communities and is now coordinated by the Data Coordination Centre at EMBL-EBI. Each sample was planned to be sequenced to 4X genome coverage. While sequencing

at this depth cannot detect every single variant in each sample and is not sufficient to determine the exact genotype at each location, it can still discover most alterations even with low frequencies. In the project's last stage, data from numerous samples was combined to enable accurate assignment of the genotypes in each sample at all the variant sites detected in the project. 1000G samples were obtained from healthy individuals with no known congenital disorder. Thus, variants found in the 1000G database are expected to be harmless and to not be involved in the development of rare, Mendelian diseases. However, as described in the introduction, this has to be taken with care as carriers might have been included in the data collection. Therefore, it has to be noted that the database might still contain harmful alterations (e.g. ones involved in complex or late-onset diseases or heterozygous alterations for recessive disorders).

### **2.2.3.2 Exome Aggregation Consortium Browser (ExAC)**

The Exome Aggregation Consortium (ExAC) Browser [42] is a curated repository of exome sequencing data from various NGS projects worldwide. It provides data from over 60,000 unrelated individuals who were sequenced as part of population genetic as well as disease-specific studies. However, data from individuals affected by severe paediatric disease have been removed. Thus, for rare early-onset Mendelian disorders, variants found in a homozygous state in the ExAC database can usually be excluded from further analysis. As with 1000G data, however, individuals included in the samples might have been heterozygous carriers of disease mutations. A genome-wide version of ExAC, gnomAD [43] exists, which is currently being integrated into MutationDistiller.

### **2.2.3.3 dbSNP and ClinVar**

dbSNP [40] and ClinVar [44] are public repositories of genetic variation run by the NCBI. dbSNP, or the NCBI Short Genetic Variations (SNV) database, is a collection of known short genetic variants in various species. Despite its name, it is not restricted to single nucleotide variants but also includes other types of variation, such as short insertions and deletions, short tandem repeats (microsatellites) and polymorphisms consisting of multiple nucleotides (multinucleotide polymorphisms). It contains harmless polymorphisms as well as alterations corresponding to known phenotypes. As such, it provides an archive of genetic variation across and within a number of species and allows for comparisons. To distinguish between harmful and harmless alterations, dbSNP variants are assigned to levels of severity such as pathogenic, probable-pathogenic, probable-non-pathogenic and non-pathogenic. Variants with clinical information are compiled in the clinical database ClinVar and can be accessed and downloaded separately. For a subset

of ClinVar cases, phenotype information is provided as well, which we used for the development of our program. MutationDistiller displays and treats data from dbSNP and ClinVar independently.

## 2.2.4 Phenotype repositories

MutationDistiller aims at illuminating the molecular cause of Mendelian disorders by connecting the patient's genotype with their phenotype. This allows the user to filter out variants which do not fit the phenotype while having a closer look at alterations in genes which have previously been found to be linked with a matching phenotype. To enable this, MutationDistiller includes phenotype data from a range of different sources.

### 2.2.4.1 Online Mendelian Inheritance in Man

Online Mendelian Inheritance in Man (OMIM, [59]) is a comprehensive collection of human genes and diseases focusing on the relationship between genotype and phenotype. It contains data on over 15,000 genes and all known Mendelian disorders. Long before the online age, it started as Mendelian Inheritance in Man (MIM) in the early 1960s, generating a manual list of Mendelian phenotypes and disorders. The online version began in 1985 and was uploaded to the Internet to become freely available to the public in 1987. Today, it is hosted and authored at the McKusick-Nathans Institute of Genetic Medicine at the Johns Hopkins University School of Medicine and builds the basis for many downstream applications on the connections between genetics and human symptoms (see section 1.4.1.5). However, the full OMIM data is currently not freely available anymore.

### 2.2.4.2 Orphanet

Orphanet [60] is a repository of rare diseases founded in France by the INSERM (French National Institute for Health and Medical Research) in 1997. Since 2000, it has become a European undertaking and is now hosted by a Consortium of 40 countries worldwide. Amongst other services and tools, it provides an inventory of rare diseases<sup>2</sup>, connected with various resources such as OMIM to enable the systematic storage and assessment of known rare disorders in humans. Moreover, with Orphanet Rare Disease Ontology (ORDO)<sup>3</sup>, a structured vocabulary for rare disease linking relationships between genes and disorders is currently being developed to support computational analyses.

<sup>2</sup>[https://www.orpha.net/consor/cgi-bin/Disease\\_Genes.php?lng=EN](https://www.orpha.net/consor/cgi-bin/Disease_Genes.php?lng=EN), accessed 11.06.2019

<sup>3</sup>[http://www.orphadata.org/cgi-bin/inc/ordo\\_Orphanet.inc.php/](http://www.orphadata.org/cgi-bin/inc/ordo_Orphanet.inc.php/), accessed 11.06.2019



### 2.2.4.3 Mouse Genome Database

The Mouse Genome Database (MGD, [94]) is a community data resource providing a comprehensive knowledgebase on mouse genes, genetic markers and genomic features. In addition, their associations to phenotypes and other properties are given as well. MutationDistiller displays MGD phenotype data to allow users to find genes which are known to cause a particular phenotype in mice. Adding mouse phenotypes as an additional layer of information can be particularly useful for genes which have not yet or cannot be studied extensively in humans. We are using the link between MGD entries and human diseases<sup>4</sup> to provide users an opportunity to search for genes causing a certain mouse phenotype.

### 2.2.4.4 Human Phenotype Ontology

The main goal of the *Human Phenotype Ontology* (HPO, [58]) was to create a standardised, computer-legible vocabulary of human phenotypic abnormalities in order to allow large-scale computational assessment of human phenotypes. By giving an identifier to each term and denoting their relationship to other terms, it allows phenotype data to be structured and helps to describe a patient's symptoms as accurately as possible. Currently, the HPO contains over 11,000 terms. It is organised in five subontologies. The main subontology is *phenotypic abnormality* with its description of disease phenotypes. Additional subontologies describe different aspects of the phenotypic abnormalities: mode of inheritance, mortality/aging, frequency and clinical modifier.

To organise the phenotypes and connect them with known disease genes, the HPO draws on data from OMIM and other sources. Nearly all clinical OMIM descriptions have been mapped to HPO terms. In addition, all Orphanet entries have been annotated, together with over 60 recurrent syndromes from DECIPHER [95], a web-based source of plausibly pathogenic genomic variants from well-phenotyped rare genetic disorder patients. By organising the data in an ontological structure, the HPO enables computational usage of the vast knowledge stored in these heterogeneous data sets. Moreover, regularly updated phenotype to gene mappings are provided. Phenotype-gene annotations are conducted using OMIM as a mediator platform. As OMIM compiles all symptoms for a given disease – irrespective of whether this disease can have multiple genetic causes or display multiple subsets of symptoms – this leads to a degree of uncertainty as not necessarily all symptoms of a disease are connected with every gene in the OMIM list. In addition, a layer of insecurity is added as neither there is no distinction between symptoms that are mandatory and others that might possibly or even only rarely occur in a given

<sup>4</sup><http://www.informatics.jax.org/diseasePortal>, accessed 11.06.2019

disorder. Nevertheless, the HPO and its gene-phenotype annotations are a valuable and widely used resource in human genetics. We obtained the OMIM-to-gene annotations via Medgen, the human medical genetics interface from the NCBI<sup>5</sup> and incorporated them into MutationDistiller to allow users to find genes linked with a specific phenotype. Thus, by describing the patient as accurately as possible, clinicians are able to reduce the relevant data to the most fitting genes.

## 2.2.5 Gene and protein function

The function of genes can be described in a number of different ways, such as a gene's role within molecular pathways, its disease relevance or its expression patterns. MutationDistiller combines a range of data sources linked with the various dimensions of gene functions:

### 2.2.5.1 Gene Ontology

The Gene Ontology (GO, [62]) is an ontological representation of genes and their functions at the molecular, cellular and tissue system level. It has grown to contain over 40,000 concepts annotating gene functions based on over 100,000 scientific publications. The GO is organised in three sub-ontologies storing molecular function, cellular component, and biological process of genes and gene products. Depending on the main point of interest taken, genes or gene products can be described via one or more of the sub-ontologies. For example, the gene product *cytochrome c* can be seen as part of all three of the sub-ontologies: *Oxidoreductase activity* focuses on the molecular function, whereas *oxidative phosphorylation* refers to the biological process and *mitochondrial matrix* to the cellular component<sup>6</sup>. Using GO terms and their relationships in MutationDistiller allows us to find genes fitting a patient's phenotype via their function without being restricted to what is known about human genes. For example, for a patient suffering from an enlarged kidney, their clinician might be able to find candidate genes by filtering the WES data via GO term *GO:35564: regulation of kidney size*.

### 2.2.5.2 Expression data

Expression data can be helpful especially in cases where the disease is limited to certain tissues or organs. Limiting the search to genes known to be expressed in the tissues of interest might help to reduce the list of candidate variants, in particular if no disease is

<sup>5</sup>[ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/mim2gene\\_medgen](ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/mim2gene_medgen), accessed 11.06.2019

<sup>6</sup><http://www.geneontology.org/page/ontology-documentation>, accessed 11.06.2019

already known for the patient’s symptoms. *ExpressionAtlas* [66], hosted by EMBL-EBI, is an open repository giving access to results from gene expression studies worldwide. It provides expression data from various species and under varying biological conditions (e.g. different tissues, cell types, and diseases). Different experimental methods are included, such as RNAseq or microarray data. All experiments can be accessed and visualised online as well as downloaded. Currently, data from over 3,000 experiments are available, which have been curated and re-analysed with standardised methods to enable continuity. For MutationDistiller, we have downloaded the *Tab Separated Values (TSV)* files for a number of data sets that we consider to be interesting for clinicians and human geneticists: These are baseline experiments, i.e. samples that had not been submitted to any experimental treatment, came from healthy tissues and organs, and were obtained by different experimental means and at various developmental stages. A list of the experiments included in MutationDistiller can be found in table 2.2.

data source	experiment	development	accession number
ENCODE [82]	RNAseq	adult	E-MTAB-4344
FANTOM5 [96]	RNA-CAGE	adult, fetal	E-MTAB-3358
GTEX [97]	RNAseq	adult	E-MTAB-5214
HPA [98]	Protein Expression	adult	E-PROT-3
HPA [98]	RNAseq	adult	E-MTAB-2836
PRIDE [99]	Protein Expression	adult, fetal	E-PROT-1

TABLE 2.2: **MutationDistiller expression data sources.** Overview of expression data included into MutationDistiller. Expression data was obtained from ExpressionAtlas [66]. Accession number: ExpressionAtlas identifier.

Saving and displaying gene expression data is a complex task: Different data sources cannot be combined or compared directly as expression levels are highly specific from experiment to experiment and from tissue to tissue. Moreover, there is a big difference between a gene not being expressed in a given tissue and a lack of expression data (zero value vs. NA).

We thus had to design ways to store a wide range of individual expression data sources while enabling users to easily access them. To achieve this, we decided to regard expression levels relative to median gene expression: For each tissue, we calculated its median gene expression across all genes and for a given data source, denoting all genes expressed below this median as not-expressed in the tissue. In addition, we calculated whether a gene’s expression in a given tissue is high (i.e. lies within the 75th percentile) or very high (i.e. within the 90th percentile). We saved this information in our database for each gene and experiment separately. We define genes as not expressed in a certain tissue if their expression levels lie below the median for all genes across this tissue – if a user selects to display only expressed genes, any genes expressed below the tissue median are removed from the results list.

In addition, we decided to collect tissues together in biological groups as each of the various data sources offers a wide range of different tissue types. The ENCODE data (E-MTAB-4344), for instance, has relatively broad tissue categories (brain, liver, heart, etc.), whereas the FANTOM5 data (E-MTAB-3358) contains sub-tissues (e.g. brain: amygdala, brain meninx, occipital lobe, etc.). In order to make search and selection more user friendly, we gathered each data source's sub-tissues together to generate groups, which can then be selected in the MutationDistiller interface. All the FANTOM5 brain sub-tissues, for example, are now collected within the category 'brain'. We stored these groups and sub-groups in our database, separately for each data source.

In the user interface, we grouped these categories again into *organs* (brain, heart, liver, etc.), *tissues* (muscle, placenta, throat, etc.) and *systems* (reproductive, nervous, immune, etc.). In addition, we identify the different data sources by the experimental means with which they were generated (RNA-CAGE, RNA-Seq and protein expression) as well as the developmental stage of the tissue (adult or fetal). As mentioned above, MutationDistiller considers a gene to be not expressed in a certain tissue group (e.g. brain) if its expression is lower than the median of all sub-groups (e.g. amygdala, brain stem, medulla...) that contain data for that particular gene. For all groups that a gene is expressed in, the sub-groups are listed in the results as well. A compilation of the expression groups can be found in table B.1 in the appendix.

### 2.2.5.3 Metabolic and signalling pathways

In cases where a phenotypic characterisation of the patient does not lead to success, i.e. for hitherto unknown disease genes, the inclusion of information on gene function can be helpful. One intuitive way of describing the function of a gene is to refer to their role in molecular pathways: Within a single pathway, numerous events (such as DNA-binding), protein complexes, reactions (e.g. adenylation), translocations and regulatory events can be represented in a simplified graphical view, enabling enhanced understanding of complex concepts and networks. We have included molecular pathway data into MutationDistiller to allow users to tackle rare, unknown cases not obviously linked with known disease genes.

The *Kyoto Encyclopedia of Genes and Genomes (KEGG)* [63] is a database platform covering various aspects of biological and cellular functions linked with genes and gene products. In its KEGG Pathway collection, manually drawn pathway maps are provided. An identifier is allocated to each map, denoting meta-information about the pathway. While the web-services are accessible freely to the public at <http://www.kegg.jp/kegg/pathway.html>, data download is only available with a paid academic subscription. As

MutationDistiller can only include open data, we have incorporated the last free data download of the KEGG Pathway collection. This data is from 2011 and does hence not entail up-to-date information. We thus decided to display the data but to not provide KEGG as a searchable resource or include it into the score.

*Reactome* [64] is an open access, open source database of molecular pathways providing access to curated and peer-reviewed data. It provides visualisation of pathways and tools for their analysis, enabling, amongst others, research and genome analyses. In its database, it combines molecules and nucleic acids interacting in reactions into biological pathways. Reactome was founded in 2001 and is now headed by a group of researchers from institutes around the globe.

*WikiPathways* [65] is a similar, open and collaborative project providing access to curated biological pathways. It is based on the MediaWiki software<sup>7</sup> employed by Wikipedia combined with a graphical pathway editing tool. For each pathway, a wiki page displays the current diagram and offers references, descriptions, download options, and supporting information. Pathways can be edited and updated by the community and changes be monitored to ensure quality of the entries. We have downloaded the *Gene Matrix Transposed (GMT)* files available for download on WikiPathways. These are lists of gene sets containing the pathways and the genes within these pathways.

As an example of the visual representation of pathways, figure 2.1 shows the Bone Morphogenic Protein (BMP) Signalling and Regulation pathway, a pathway of importance in embryogenesis and development.

#### 2.2.5.4 Gene panels

Especially for patients with well-characterised diseases or symptoms, gene panels are known to be a great tool in the hunt for disease alterations [100, 101]. Targeted gene panels contain lists of genes known to be linked with a certain disease or group of diseases. They can either be used to sequence the panel genes only or as virtual gene panels to filter the results of a WES or WGS for panel genes, bringing down the number of remaining candidates substantially to variants located in genes matching the case. MutationDistiller incorporates virtual gene panels from various sources:

The *Kingsmore panel* [61] is a collection of genes which have been found to be involved in rare recessive genetic disorders manifesting in early childhood. After having been reviewed and assessed in multiple ways, various versions of the panel exist which differ in small aspects. We have decided to include the Kingsmore panel version included

<sup>7</sup><https://www.mediawiki.org/wiki/MediaWiki>, accessed 11.06.2019

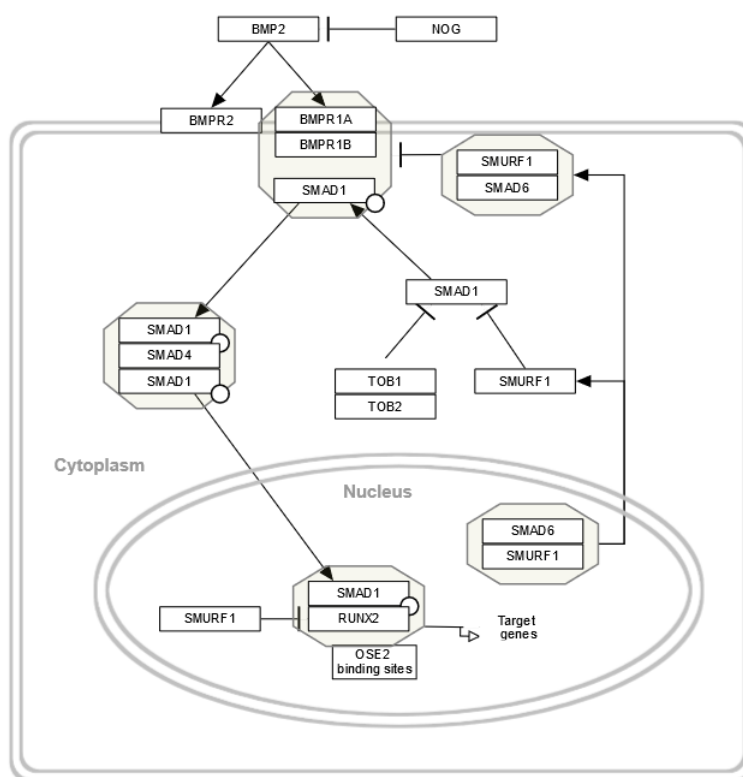


FIGURE 2.1: **WikiPathways Bone Morphogenic Protein (BMP) Signalling and Regulation pathway (WP1425)**. Displays the graphical representation of the BP pathway stored in WikiPathways. Provided by Waagmeester A, Pico A, Hanspers K, Osman BM *et al.* Accessed 17.09.2018.

in the heterozygote screening conducted by Pränatal-Medizin München as provided by Orphanet<sup>8</sup>. This version of the panel contains 550 genes for 258 different diseases.

Another prominent virtual gene panel which we integrated into MutationDistiller is the *HPO panel*. This panel contains all the genes which have been connected with any HPO term, thus any gene which has ever been found to be linked with a disease symptom. We obtained version 2 of this panel with 3061 genes from the Institute of Medical Genetics, Charité Berlin.

MutationDistiller also contains the *ACMG actionable genes* panel, which is a list of genes published by the American College of Medical Genetics and Genomics (ACMG). It compiles genes of medical interest: These are genes for which the knowledge of a mutation within it might be beneficial to the patient as action (prevention) can be taken. An example is the *APOB* gene, which is known to be involved in familial hypercholesterolemia. Knowledge of a mutation in this gene allows medical action to be taken, which could be

<sup>8</sup><https://www.orpha.net/data/dgs/DE/DgsID109383.pdf>, accessed 23.12.2018

life-saving. The most recent version of the ACMG actionable genes panel, ACMG SF v2.0 [102], is incorporated into MutationDistiller.

In addition to published panels, many clinics or sequencing services have generated their own gene panels from scientific literature. These can simply be used in MutationDistiller by copying a gene list into the respective entry field. To generate a reliable knowledge base for virtual gene panels commonly used in human genetics, Genomics England's PanelApp<sup>9</sup> offers expert curated gene panels from and for the scientific community. Currently, gene panels of varying size for over 2000 human conditions are available from their services, which we have downloaded and integrated into MutationDistiller. PanelApp is a community-driven approach that calls experts into action: Each virtual panel is reviewed and curated by clinicians or geneticists who are experts for a certain disease, gene or disease group. The panel genes are sorted into three categories depending on the confidence with which they have been added to the list. 'Green' genes are intended to be diagnostic-grade and according to their criteria require evidence from three or more unrelated families or from 2-3 unrelated families where there is strong additional functional data. All other genes which do not match these guidelines are rated as 'amber' or 'red' and should not be used in diagnostic settings, according to PanelApp's creators.

### 2.2.6 Protein families

To provide a user-friendly interface allowing clinicians and human geneticists to draw their own conclusions about a variant's relevance, MutationDistiller provides as much information as necessary and possible in one place. The inclusion of protein information might help users decide for themselves which variants to assess with further scrutiny and which ones to dismiss. MutationDistiller hence displays protein-related data from three major resources to assist users.

*PFAM* [103, 104], hosted by EMBL-EBI, is a database collecting protein families. It represents them by multiple sequence alignments and hidden Markov models (HMMs) to display similarities between proteins and to allow insights into protein functions. PFAM draws on data from the UniProt Reference Proteomes<sup>10</sup>. The information stored in PFAM is accessible online and can be downloaded as well.

*InterPro* [105] is another protein platform providing functional analyses of protein sequences. It offers an insight into the functions of a protein by storing predicted protein domains. Predictive models (signatures) provided by various member databases such as

<sup>9</sup><https://panelapp.genomicsengland.co.uk/>, accessed 11.06.2019

<sup>10</sup>[https://www.uniprot.org/help/reference\\_proteome](https://www.uniprot.org/help/reference_proteome), accessed 11.06.2019

*PFAM*, *PANTHER* [106] or *SMART* [107] are used to classify proteins. It is run by a consortium of protein databases from around the globe and hosted at EMBL-EBI.

## 2.2.7 Protein-protein interactions

### 2.2.7.1 STRING

STRING [108] is a database of protein-protein interactions developed by a consortium of European institutions. It contains experimental data as well as computational predictions. Currently, data on almost 10 million proteins from over 2000 organisms is included, which is available both online and can be downloaded. We have integrated human STRING data into MutationDistiller to display STRING interactions and to provide hyperlinks to relevant entries.

## 2.2.8 Mitochondrial data

While the mitochondrial genome with its 37 genes (including 13 protein-coding genes) only makes up a tiny fraction of the human genome, mitochondriopathies place a burden on a large number of patients [109]. Mitochondriopathies include diseases linked with proteins generated in the mitochondria directly and those that are shuttled into the organelle - the latter being the vast majority: Most of the more than 1,000 different mitochondrial proteins are encoded by nuclear DNA and have to be shuttled into the mitochondria to fulfil their function. MutationDistiller provides access to three different resources on mitochondrial data.

The *Maestro score* [110] is a broadly used scoring system to predict mitochondrial proteins encoded by nuclear DNA. It uses eight genomic data sets on targeting sequence prediction, protein domain enrichment, presence of cis-regulatory motifs, yeast homology, ancestry, tandem-mass spectrometry, co-expression, and transcriptional induction during mitochondrial biogenesis to determine the likelihood for a protein to be functional in mitochondria.

*MitoCarta* [111] is an inventory of mitochondrial proteins hosted by the Broad Institute<sup>11</sup>. It was generated by experimental means using mass spectrometry of mitochondria from fourteen different tissues. In addition, protein localization was assessed in large-scale GFP tagging/microscopy. The results were then integrated with other data sets, generating an inventory of 1158 human and mouse genes. Data are available online and can

---

<sup>11</sup><https://www.broadinstitute.org/scientific-community/science/programs/metabolic-disease-program/publications/mitocarta/mitocarta-in-0>, accessed 11.06.2018



be downloaded. We have incorporated MitoCarta data into MutationDistiller to allow users working in the field of mitochondrial pathologies to assess at a glance whether a certain gene is relevant in their case.

*Mitopred* [112] was a web server for the prediction of mitochondrial proteins encoded by nuclear DNA in eukaryotes. It based its predictions mainly on Pfam domain data comparing mitochondrial and non-mitochondrial locations. Data was available online and downloadable. While the service has since been discontinued, MutationDistiller still displays Mitopred data from the latest update (08/2016).

## 2.3 MutationDistiller's database

### 2.3.1 Database structure

All of MutationDistiller's data are stored in one database but organised in different *schemas* of related tables. These schemas – which are often distinct data entities but reference each other – mirror logical categories that the data can be sorted into. The data used by MutationDistiller can be divided into two main categories: Project-related data and general data, with general data falling into five schemas. In the following, I will describe the database structure and relationships between the different tables.

#### 2.3.1.1 Query Engine schema

The Query Engine (QE) was first developed for MutationTaster and has since been adapted for MutationDistiller. It reads the submitted VCF file line by line and saves the information in our database. During this process, it generates a number of project-specific tables and adds information to our variant tables. The QE database schema is visualised in figure 2.2 and the QE protocol is described in 3.

The submitted variants are saved in our variants table (*all\_vars*), which compiles all variants that have ever been uploaded to MutationDistiller and information related to them. Double entries, variants in the wrong format or with a coverage below a user-defined minimum threshold are discarded. Upon upload, the variants are checked for the correct version of the reference genome (currently GRCh37). In a next step, variants that are found in the variant databases 1000G or ExAC with a genotype count exceeding custom-set thresholds are filtered out as well. By default, variants that appear at least 10 times in a homozygous state in ExAC and 4 times in 1000G are discarded; however these settings can be changed by the user. In addition, the Query Engine upload page allows

users to restrict analysis to certain genomic regions and to only analyse homozygous alterations – if this is the case, all variants not matching these criteria will be filtered out. Only the variant itself is committed to the *all\_vars* table. There, it is assigned a variant number (*var\_number*), which serves as a primary key and at the same time allows other QE schema tables to access the information.

Some of the QE tables store and summarise data for all projects that have been uploaded to MutationDistiller thus far: *all\_projects* provides meta-data on all MutationDistiller projects such as email address (if provided), project name, or number of variants. *all\_results*, on the other hand, contains all MutationTaster results and background information (such as the results of the underlying tests conducted by MutationTaster) for all variants in the database. This helps to speed up run times and saves database resources as each variant has to be saved to the database and analysed only once – after committing it to the database it can be easily accessed later on.

In addition to the variant-related tables, separate tables are generated for each new project, allowing MutationDistiller to quickly and simply access the projects if a user enters the relevant ID: Input information, i.e. a project's variants, their coverage, and homozygosity state are saved in an *input\_ProjectID* table. In addition, the QE schema contains tables for each project that identifies areas that could be present in a compound heterozygous state. These compound heterozygosity tables (*comphet\_ID\_VARSEL*) are generated depending on the selected MutationTaster variant severity predictions: Each variant combination that a user analyses (e.g. severe variant settings for both variants versus severe first variant and benign second variant settings) generates a new compound heterozygosity table during the prioritisation (see sections 4.1.1 and 5.1.4 for details). Therefore, this table is technically also part of the prioritisation protocol. Nevertheless, in our database set up we decided to include it in the QE schema because - as a project-specific table with long-term storage - it belongs to the QE logically.

For further information on the QE, its processes and features, please refer to chapter 3, for the QE's user interface to 5.

### 2.3.1.2 MutationDistiller database schemas

#### Public

Some of the data are accessed by several of our applications and are stored in the schema type *public*. This schema was first established for the gene-ranking tool GeneDistiller [81] and has since grown to accommodate a range of tools developed in our research group. This schema contains all gene-related data, such as gene names and numbers and their position. In addition, external data sources referring to those genes are stored in

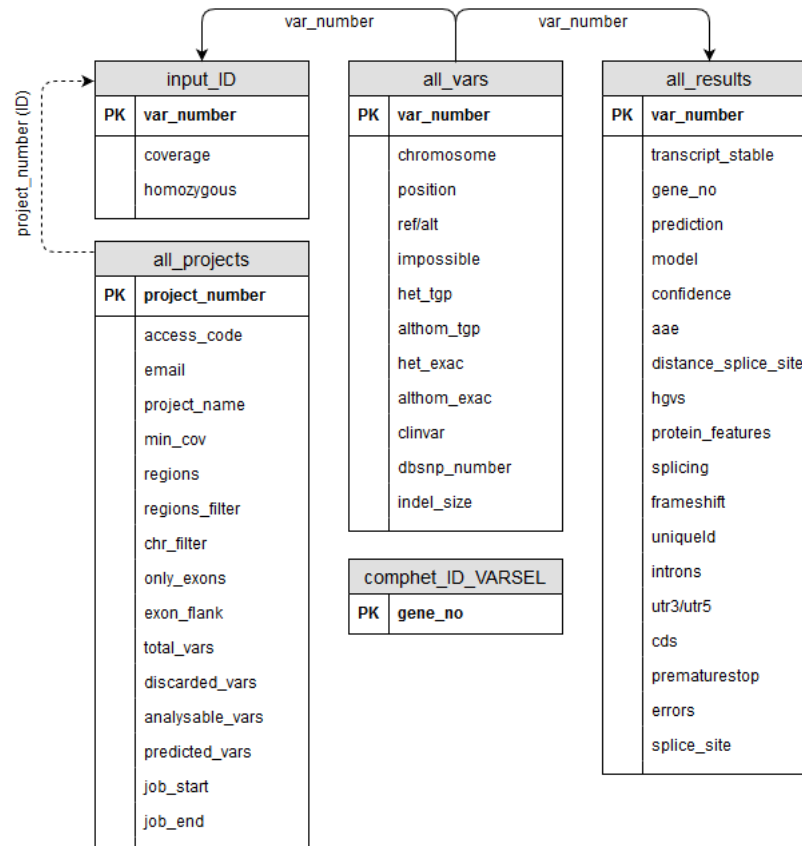


FIGURE 2.2: **Overview of MutationDistiller's Query Engine Schema.** Displays the database tables stored in the QE schema and their references, important attributes and primary keys (PK). Please note that for reasons of legibility, some attributes have been omitted in this depiction.

this schema as well, including OMIM and OrphaNet data, GO and MGD entries and pathway data.

### Ensembl

The Ensembl data used by MutationDistiller is stored in the schema type *ensembl37\_85*. This schema includes gene and transcript tables obtained from Ensembl, exon and transcript data, and links to *Entrez gene ids*, which enables us to connect Ensembl data to a multitude of other resources stored in the *public* schema. In addition, the expression-gene links are stored in this schema as well. It currently contains version 85 of GRCh 37 data stored in Ensembl, but other versions can be added and run in parallel.

### HPO

HPO data is stored in a separate schema, *hpo*. It contains HPO terms, their ancestors and descendants together with synonym and opposite terms. In addition, the relevant genes linked to each HPO entry are stored in this schema.

## Expression

The *expression* schema contains expression data obtained from ExpressionAtlas and connects to the *ensembl37\_85* schema.

## Build37

The *build37* schema of our database stores all genome version-specific data, such as the variant databases as well as the genotype counts for the Query Engine and user interface. This schema is therefore crucial for the variant effect predictions conducted by MutationTaster, which form the basis of MutationDistiller's sorting and filtering mechanism.

### 2.3.1.3 MutationDistiller entity-relationship diagram

The tables, schemas and relations of a database can be represented in an *entity-relationship diagram (ERD)*, which allows the database's structure to be visualised. Figure 2.3 displays such an ERD in a simplified version for the tables and schemas used by MutationDistiller. In this depiction, I have decided to omit a number of tables for reasons of legibility. A comprehensive ERD displaying all the tables and schemas employed by MutationDistiller can be found in the Appendix.

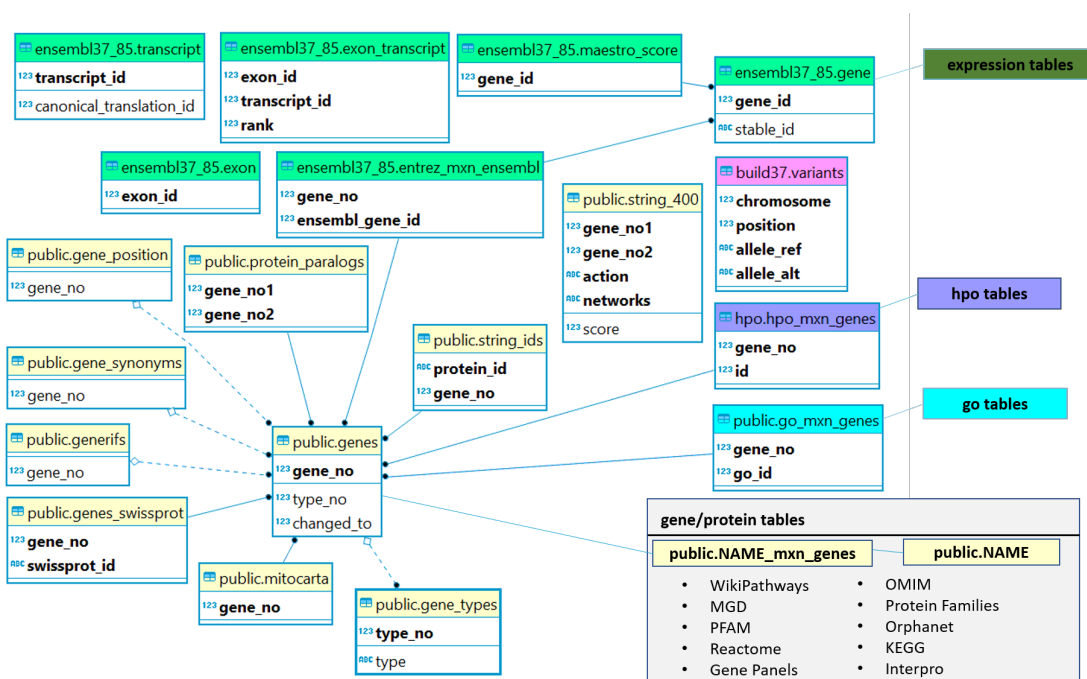


FIGURE 2.3: **MutationDistiller database ERD.** A simplified view of the database tables used by MutationDistiller and their connections, references and keys. The colours indicate different database schemas. Specific sub-schemas (HPO, GO, expression and protein/gene tables) were removed from this ERD to provide a legible diagram. Symbols denote data types: 123 - numeric; ABC - text. Please note that only selected columns are shown, as a comprehensive ERD is located in the Appendix.

## 3 MutationDistiller: Query Engine

When a user submits variants stored in a VCF file to MutationDistiller, the tool commits the data to its database, sends the variants to MutationTaster for pathogenicity predictions, and records the received predictions in the database. All these steps are conducted by MutationDistiller's Query Engine (QE) and take place independently from the prioritisation process. In this chapter, I will explain the various steps undertaken by the QE. In the following chapter 4, I will then describe the prioritisation process which gets started when a user calls a project from the database, and in chapter 5 the interface of QE and main program.

### 3.1 File upload

When starting a new MutationDistiller project, the user submits their VCF file to our QE system. The QE was first developed for MutationTaster and has since been adapted for MutationDistiller. It consists of a number of Perl scripts, which are called via shell scripts to manage all projects. The submitted projects are scheduled using the freely available resource manager TORQUE<sup>1</sup>. The Perl scripts that make up the QE send jobs to TORQUE, which organises them according to file size: Smaller projects get processed faster than larger ones, and large projects may be split up and worked on in parallel. In addition, two customised Perl modules contain query-engine related functions: *QueryEngine.pm* governs general query engine tasks, and *SendMail.pm* sends notification emails from the query engine. These Perl modules are collections of Perl functions that were not written especially for MutationDistiller but are shared by all our programs.

From MutationDistiller's start page<sup>2</sup>, the QE can be accessed via a hyperlink which leads to a HTML page for file upload<sup>3</sup>. The QE interface is described in further detail in section 5.1.2.

### 3.2 Query Engine workflow

In order to store all the relevant data in MutationDistiller's database, the program goes through a dedicated routine which I describe in the following and which is depicted in figure 3.1 at the end of this chapter in a simplified flow chart.

<sup>1</sup><http://www.adaptivecomputing.com/products/open-%20source/torque/>, accessed 11.06.2019

<sup>2</sup><https://www.mutationdistiller.org/>

<sup>3</sup><https://www.mutationdistiller.org/MutationTaster/StartQueryEngine.html>

As a first step, the *MDQE\_start* script checks all submitted data for coherence and usability. For example, it ensures that a VCF file has been submitted, that the minimum coverage was entered as a valid number, or that the given candidate genes exist. If something is wrong, the program returns an error message, alerting the user to the problem and pointing them in the right direction. If all the submitted data are correct, MutationDistiller generates a running project number and a unique, random, six-digit access code. These two components are combined to form the case ID enabling users to access their project. The project number allows us to keep track of the cases seen by MutationDistiller thus far. The case ID ensures that only the user holding the ID can access the case. In addition, it allows the unique identification of a project: While several users might give their projects the same title, each ID is only allocated once. A typical case ID would be 123\_456789, where the digits in front of the underscore are the project number and the six digits after are the access code.

The start script then saves the user settings and proceeds to the next script. This script, *VCF2DB*, reads the VCF file line by line and extracts the data for each variant: chromosomal position, reference and alternative allele, coverage and genotype. It checks whether the variant's coverage is greater than the entered minimum coverage and discards variants that do not fulfil this criteria. In addition, it checks that the same variant is not yet located in the database (in table *all\_vars*) and only commits new variants to this table to ensure low run-times and to save data storage space. Furthermore, the first 40 SNVs are checked for correct annotation (i.e. reference allele matching genome version 37). If this check fails, the whole process will be aborted and an error message will be sent to the user via the *errors* script. Please note that for reasons of clarity, this step has been omitted in the graphical representation. Finally, all project-related variant information passing all the filters will be saved to table *input\_ID*.

In the next script, *Map2Transcripts*, the QE maps the submitted variants to transcripts to be able to send it to MutationTaster for variant effect predictions. This is achieved for each variant and for each possible transcript consecutively. To speed up the process and to decrease the amount of data that are returned from the database, only variants for which no MutationTaster result is stored in *all\_results* are queried. Variants are mapped to all protein-coding transcripts with which they overlap (same chromosome and variant start before/equal to transcript end and variant end after/equal to transcript start).

In the end, a temporary database table *transcript\_ID* containing variant-transcript pairs is created for the project. This list is then handed over to the *CreateTasterPackages* script, which splits it into a number of packages to be run in parallel in the next step – the MutationTaster analyses. The number and size of the packages depends on the size of the project: Large project will be sent in a great number of different packages. Splitting

up large projects in this way allows the upload to be sped up tremendously and ensures that under normal circumstances, a conventional WES run will be handled within a few minutes at most, and significantly faster if it contains a large amount of variants that have already been seen by MutationDistiller. The packages are then sent to MutationTaster by the *Query\_MT* script, which in turn saves all the results in the *all\_results* table. This script goes through each entry in *transcript\_ID*, runs the MutationTaster analysis and saves the results. It then deletes the variant-transcript pair from *transcript\_ID*.

The following script *CallMissingTests*, ensures that all sets and tests have been completed and saved by checking whether there are any remaining entries in *transcript\_ID*. If this is the case, it calls *CreateTasterPackages* again. Finally, *MDQE\_finish* compiles statistics, saves the total number of variants included in the project to the database and notifies the user of the completion of the project. In addition, it calls the script *CleanUp* to remove all temporary tables. The *errors* script mentioned above might be called at any point where a problem is encountered, such as incorrect variant formats or missing entries.

After the project has been successfully processed by the QE, the user will be redirected to an overview page with data on the analysed file. From there, they can directly access MutationDistiller's main page with the project identification number (ID) pre-filled. In addition, if an email address was entered, the user will receive one email notifying them of the submission and one after upload and analysis have been completed, together with a link to their project and the project ID.

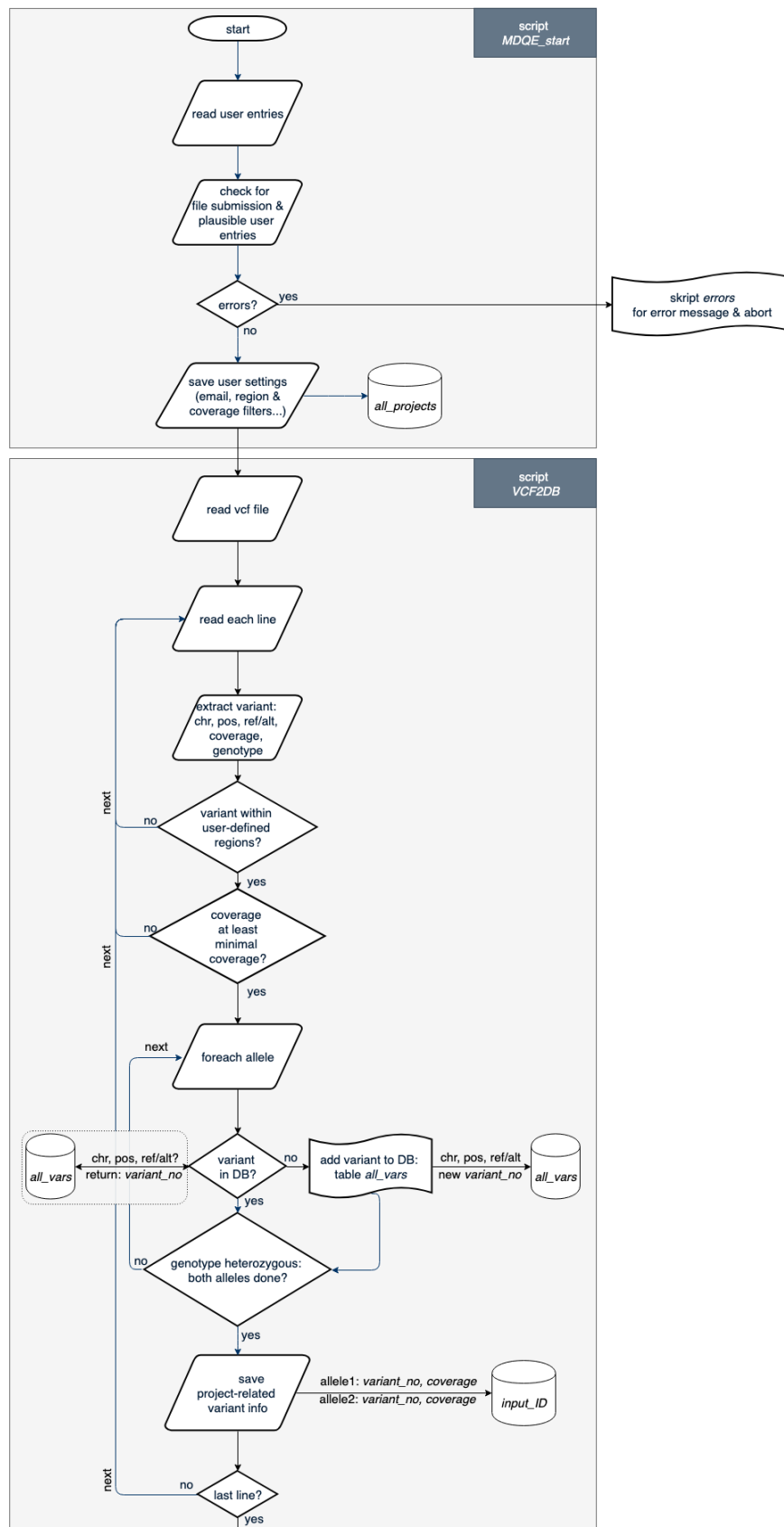


FIGURE 3.1: Simplified view of MutationDistiller's Query Engine workflow, part 1. chr: chromosome, pos: position, ref/alt: reference/alternative allele, MT: MutationTaster, DB: database, MDQE: MutationDistiller Query Engine.



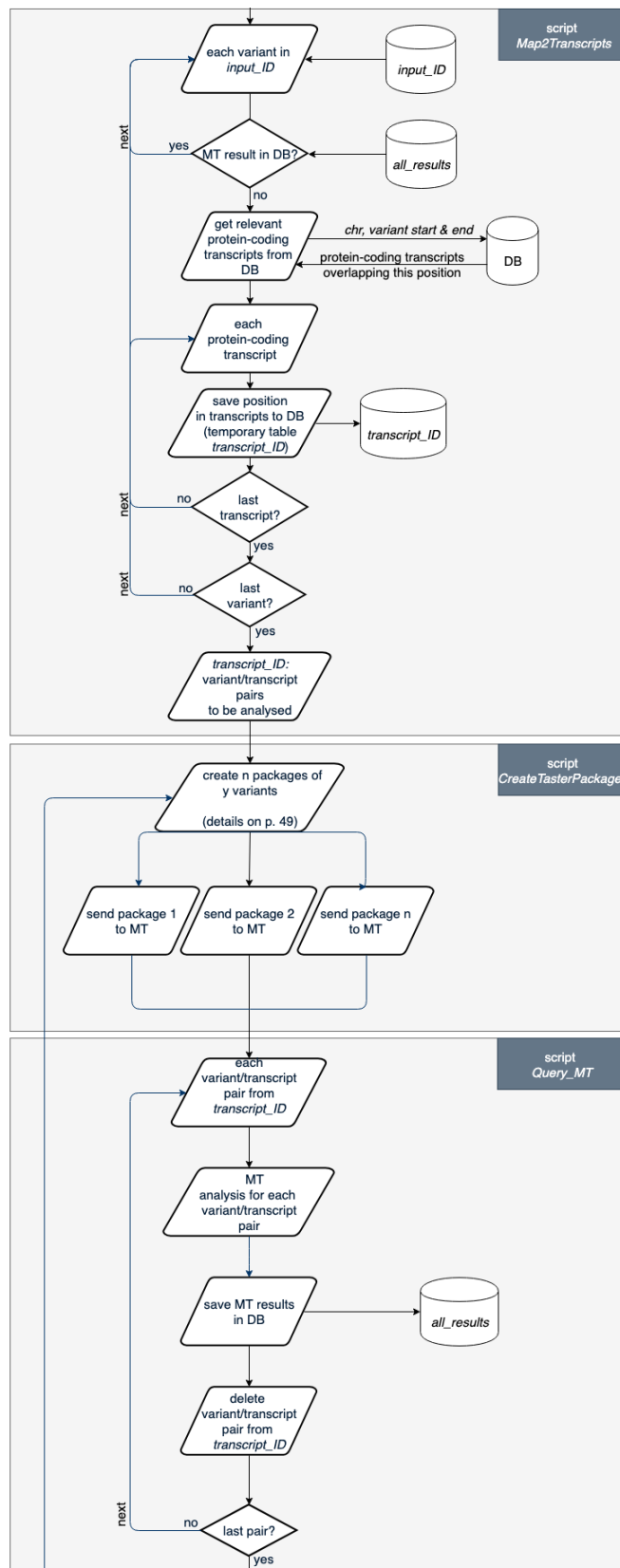


FIGURE 3.1: Simplified view of MutationDistiller’s Query Engine workflow, part 2. chr: chromosome, pos: position, ref/alt: reference/alternative allele, MT: MutationTaster, DB: database, MDQE: MutationDistiller Query Engine.

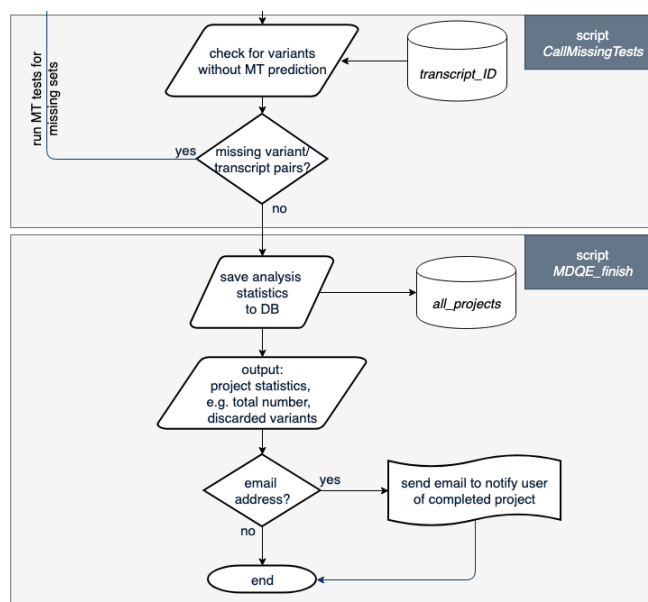


FIGURE 3.1: **Simplified view of MutationDistiller’s Query Engine workflow, part 3.** chr: chromosome, pos: position, ref/alt: reference/alternative allele, MT: MutationTaster, DB: database, MDQE: MutationDistiller Query Engine.

## 4 MutationDistiller: Prioritisation

### 4.1 Filtering, scoring and providing information

Once a user has uploaded a project to our QE and it has been committed to our database, the data can be accessed via the project's ID and security code. Users can call their projects and enter case-specific information, which MutationDistiller will use to determine the most likely candidates in a given case.

Upon data submission, a *Perl CGI* script *MDresults.cgi* is called which governs and runs the Perl functions necessary for MutationDistiller to run its course. The various functions called by this script are organised in a number of customised Perl modules: *Input* reads the input and passes it on to subsequent functionalities. *DBqueries* contains functions that retrieve data from the database, *Scoring* governs the scoring and weighting process, and *Output.pm* generates the output page. In addition, *Errors* generates error messages, *Settings.pm* governs MutationDistiller's settings, and *Debugging.pm* holds functions for internal debugging.

Other customised Perl modules are used by all of our programs and are also employed by MutationDistiller: *common.pm* holds general functions used by all our programs, *database.pm* governs database handling, and *HTML.pm* deals with HTML-related tasks.

Figure 4.1 at the end of this chapter provides a simplified overview of the data analysis steps undertaken by MutationDistiller in this sorting and prioritisation process, which I will explain in further detail below. Please note that I decided to split this figure across sections of varying size to avoid interrupting logical sub-queries or loops within the program.

#### 4.1.1 Initialising

Before going through its program routine, MutationDistiller reads the user entries and checks for problems and correct authentication: For instance, the script checks whether the entered case ID is correct and whether there are any misspelled entries that cannot be identified. At this stage, manual user entries such as HPO or WikiPathway identifiers get trimmed to remove superfluous spaces or other symbols. If the program cannot find any critical errors, it reads the selected variant classes (see also sections 5.1.4 and 3.2). For each variant selection combination, a new *comphet\_ID\_VARSEL* table is generated, if it does not exist already from a previous analysis with the exact same settings for

both variants (in cases of compound heterozygosity). This table contains all genes with at least two heterozygous variants fulfilling the variant-class criteria. For example, if a user chose a strict setting for the first variant (only ClinVar or nonsense-mutations) and a lenient setting for the second variant (all variants), the *comphet\_ID\_VARSEL* table for this analysis will include all genes that contain at least two variants fulfilling these criteria. At this step of the algorithm, MutationDistiller assesses whether the table exists for the current settings already. If not, it is generated now and committed to the database.

In the next step MutationDistiller queries its database for all project-specific variants (via the unique project ID) and pathogenicity predictions obtained by MutationTaster. As displayed in figure 4.1, this is achieved by combining information from various database tables:

The *comphet\_ID\_VARSEL* table will be linked – i.e. form a relation – with other project-specific (*Input\_ID*) and global tables (*all\_vars* and *all\_results*) to allow MutationDistiller to receive all variants and MT results relevant for this specific analysis. In this process, all variants that do not pass a region or candidate gene filter will not be called from the database. For example, if candidate genes are provided, MutationDistiller will only call variants located in those genes.

MutationDistiller’s variant selection filter (see 5.1.4 for details) allows users to focus their attention on a subgroup of variants depending on the predicted effect they have on the resulting protein. This filter is applied in the next step of MutationDistiller’s algorithm and removes all variants from further processing that do not fulfil the given criteria. For instance, if a user decided to only include variants listed as ‘pathogenic’ in the ClinVar database and/or that cause a nonsense mutation, all other alterations will be removed at this stage.

If the user selected to filter for recessive mode of inheritance, all alterations with a heterozygous genotype are removed unless their harbouring genes are found in the respective table for compound heterozygous variants. In this way, MutationDistiller trims the candidate list to only those variants that the user considered to be of interest for their specific case.

### 4.1.2 Gene information

MutationDistiller then initialises its scoring and prioritisation protocol: First, the program assesses each gene containing a variant that has not been excluded in the filtering steps, and queries gene-specific information for scoring and prioritising the candidates.

For all retrieved genes, the respective data is gathered from the database via gene identifiers provided by NCBI and Ensembl to access the data for every gene of interest. The sources for this data are described in section 2.3.

Next, MutationDistiller reads the user’s additional phenotype entries (such as HPO terms or GO identifiers) to retrieve additional gene information from the database and to select relevant data for scoring and display: On the input page, users can select specific areas or search domains to be excluded from display (e.g. OMIM reports). If any domain is neither needed for prioritisation nor set to be displayed, it is removed at this stage. For all other search domains, MutationDistiller checks whether a match score is to be allocated (e.g. HPO match) and adds these scores up to receive a final MutationDistiller score for each gene. MutationDistiller’s scoring system is explained in further detail in the following section 4.1.2.1. The different types of gene information included in MutationDistiller are listed in table 4.1.

In a second step, the program filters genes out that the user decided to exclude in addition to the region filters mentioned above. This can also be the case for genes that are not expressed in a tissue of interest, or for genes linked with a given HPO term that the user excluded from their search. For each gene not passing the filter, a flag (*dont\_show*) is set to remove it from the output – thus, although these genes will be called from the database, they will not be displayed to the user in the end.

type	description
gene type	describes the type of gene (e.g. protein coding)
reported mutations	known mutations located in the gene
pathways	KEGG pathways, WikiPathways, Reactome pathways
phenotype	HPO, OMIM, OrphaNet, MGD entries for the gene
generifs	short summary statements
gene function	Gene Ontology entries
transcripts	known Ensembl transcripts
interactions	STRING protein interactions
mitochondria	MitoCarta, Maestro, MitoPred entries
protein information	InterPro domains, NCBI paralogs, PFAM protein families
expression	ExpressionAtlas data

TABLE 4.1: **MutationDistiller gene information.** Lists information provided for each gene.

MutationDistiller uses all gene information to generate a score, which forms the base for its prioritisation. I describe the scoring system for the different search domains below, with a focus on the HPO score.

#### 4.1.2.1 MutationDistiller score

MutationDistiller’s score embodies how well a gene and its contained variants match the user-defined criteria. Depending on what sort of data was entered, it can be comprised of a number of sub-scores linked with various domains of interest such as ‘HPO’, ‘OMIM’ and ‘Reactome’. These search domains are allocated different weights based on biological and functional considerations and mirror the quality of a match.

Some domains are given a much greater weight than others, mirroring their biological or functional relevance. For instance, OMIM entries receive a high weight as we assume the existence of a secured diagnosis to be a strong indicator for a gene’s relevance. For pathway data, if a user entered information that scores several matches within one pathway, the weight gets adjusted to avoid weighing it too heavily. For example, the initial weight for matching a Reactome term is 5, but for subsequent matches this is lowered to 3. In addition, to ensure that known harmful variants are ranked highly, a ClinVar score is added to the final score if a variant is known to be disease causing (independent of user entries). A list of the various domains and their current weights can be found in table 4.2. However, please note that this information can change with any MutationDistiller update.

weight category	weight
HPO direct	5
HPO descendant	2
HPO ancestor	0.05
ClinVar	0.5
OMIM-ID	20
OMIM-title	1
OrphaNet	5
mode of inheritance (Mol)	5
homozygous genotype in recessive Mol	2
generifs	1
MGD phenotype	1
GO	1
* WikiPathways	5
* Reactome	5

TABLE 4.2: **Weight categories overview.** Displays weights assigned to the different categories used for scoring. Please note that for Reactome and WikiPathways, marked with an asterisk, consecutive matches are allocated a lower weight of 3.

#### 4.1.2.2 Scoring HPO matches

Due to their systematic nature, ontologies allow us to express and quantify the importance of a term in comparison to other terms (see section 1.4.1.5 for details). Therefore, we were able to develop a dynamic scoring system for HPO matches: *Direct matches* get scored depending on their *information content*, i.e. their relevance for the user. The

more precise a HPO term is, the fewer genes will be annotated with it. An example is the HPO-term *HP:0004940, Generalized arterial calcification*, which is linked with the specific OMIM entry *OMIM:208000* and only one gene (*ENPP1*). Thus, it is a very precise term which carries a high amount of information. If such a specific term is entered by a user and finds a match for a gene found in the submitted variants file, it is quite likely that a deleterious variant is relevant in the given case. MutationDistiller will honour this with a high HPO score:

As described in 1.4.1.5, information theory allows us to express a node's specificity as the fraction of annotated terms for it. In the HPO, this can be seen as the number of genes annotated with a given term. Its information content can thus be determined as follows:

$$IC(t) = -\log(g(t)/g),$$

where  $IC(t)$  is the information content of a specific HPO term,  $g(t)$  is the number of genes annotated with it and  $g$  is the total number of genes annotated with any HPO term (currently 3526). For performance reasons, we have encoded this as

$$IC(t) = \log(g/g(t)),$$

which is the same mathematically speaking. The result of this expression is higher, the lower  $g(t)$  is and thus accounts for term specificity.

As phenotyping is a highly subjective process, there is always a degree of uncertainty involved. Therefore, scoring not only direct matches but also related terms enables us to minimise losses due to this phenotyping uncertainty. We thus decided to score ancestors and descendants as well, but with a weight accounting for phenotyping gaps and errors:

$$HPOscore = IC(t) * weight.$$

The weight for scoring HPO terms was set and optimised on clinical data as described in section 7.1. If a HPO term is matched both directly and via an ancestor, only the direct match is counted, and if several descendant or ancestor terms match, only the highest score is counted.

It has to be noted that the HPO's coverage is not uniform across the entire ontology. Different areas within the HPO are covered with various degrees of depth due to the nature of its generation – parts that are better annotated and thus have better phenotyping

accuracy and depth are more reliable and detailed than others. Therefore, the different branches of the HPO cannot be compared easily with each other, rendering the distance between two terms meaningless when assessing the quality of a match. We therefore decided not to take the distance between two terms into account when determining its importance for scoring but instead focused on term specificity as described above.

### 4.1.3 Output generation

After scoring all relevant genes, the program sorts the results according to the score and prepares the output page, limited to the number of display genes specified by the user (default 10). We chose this default because in our test we found that MutationDistiller was capable of ranking the vast majority of disease-relevant genes (82.2%) within the first 10 ranks (see section 7.2.3 for details).

For each gene, MutationDistiller generates an entry in the summary table at the top of the page. This table contains all the variants located in the gene (after filtering for severity, position, and genotype), their MutationDistiller score, and basic gene or variant information. Below this, detailed information is listed for each gene: To allow clinicians and researchers to see background information on their patient's variants at a glance, MutationDistiller provides comprehensive information for each gene. This information is displayed even if it is not used for scoring or prioritising or in cases where the user only provides the variant data without any further selections or restrictions. MutationDistiller uses NCBI and Ensembl gene identifiers to access the gene-specific data. The layout and setup of the user interface including the output page is described in detail in the following chapter.



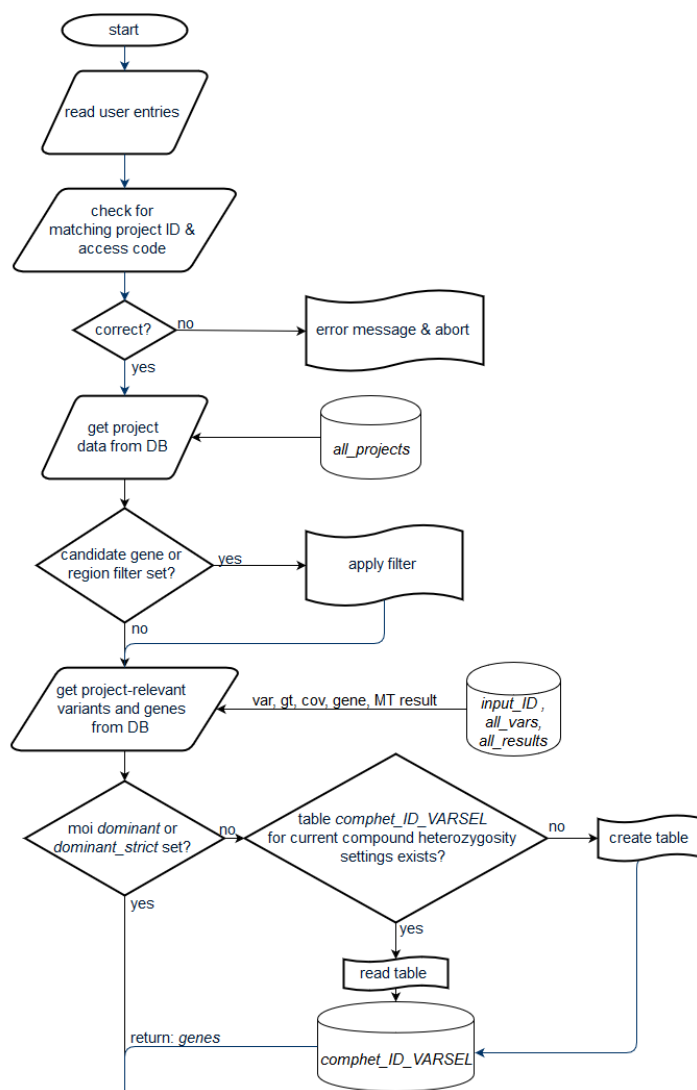


FIGURE 4.1: **Simplified view of MutationDistiller’s prioritisation workflow, part 1.** chr: chromosome, pos: position, ref/alt: reference/alternative allele, cov: coverage, DB: database, MT: MutationTaster.

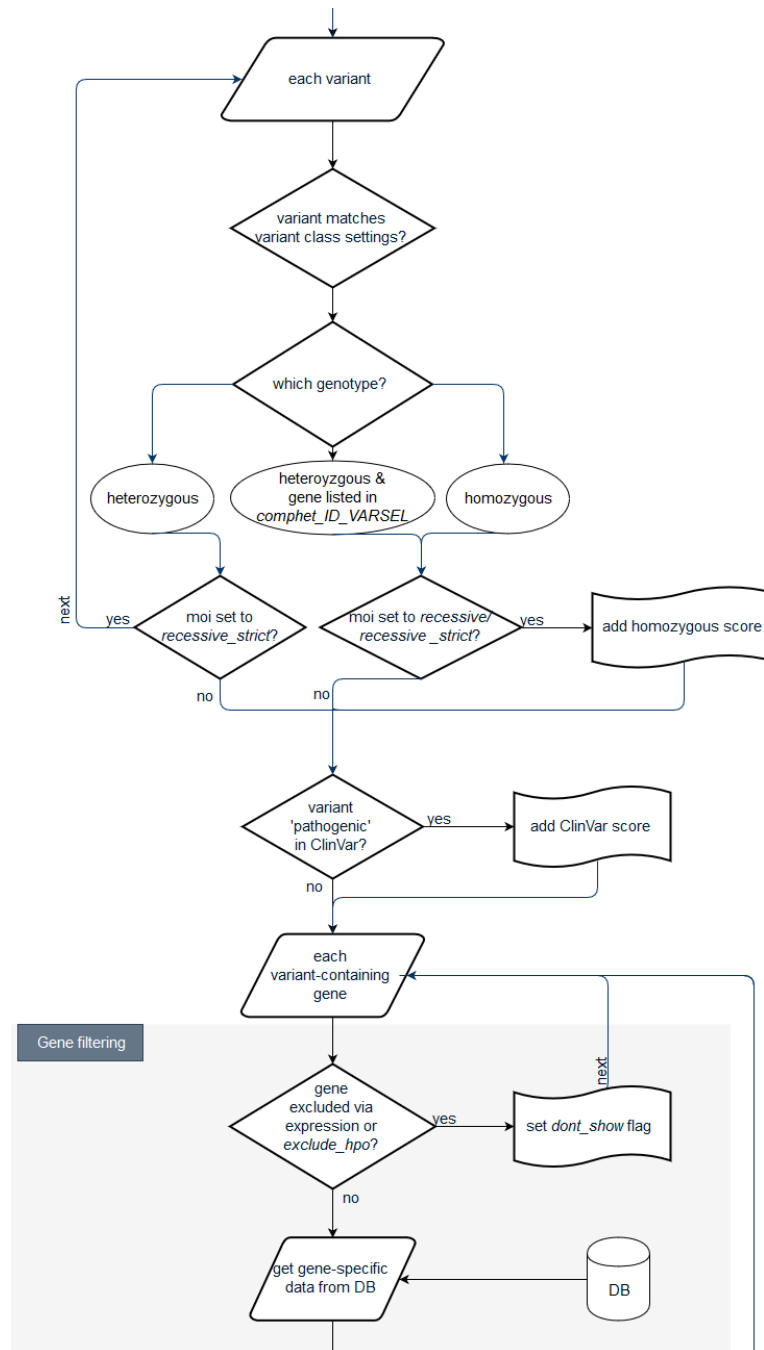


FIGURE 4.1: **Simplified view of MutationDistiller's prioritisation workflow, part 2.** chr: chromosome, pos: position, ref/alt: reference/alternative allele, cov: coverage, DB: database, MT: MutationTaster.

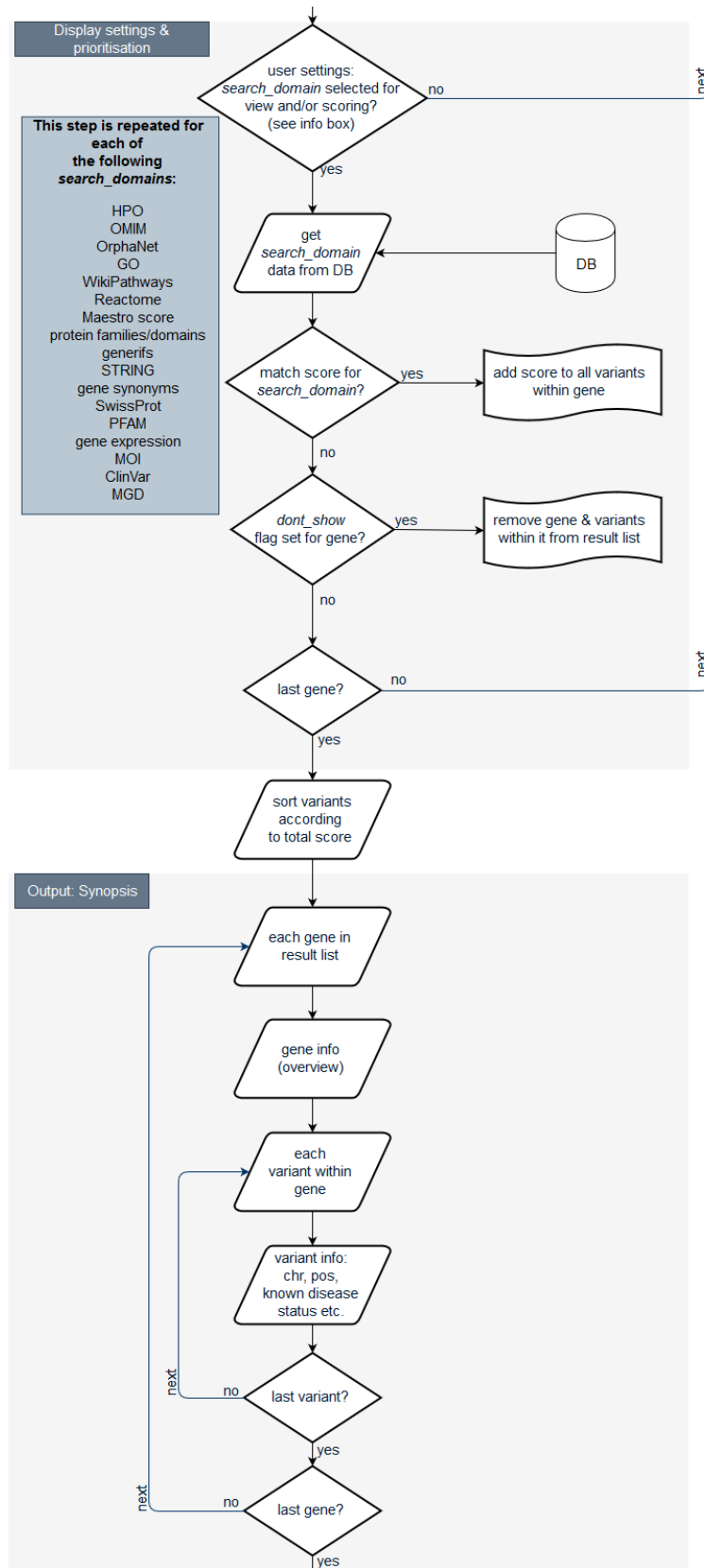


FIGURE 4.1: Simplified view of MutationDistiller’s prioritisation workflow, part 3. chr: chromosome, pos: position, ref/alt: reference/alternative allele, cov: coverage, DB: database, MT: MutationTaster.

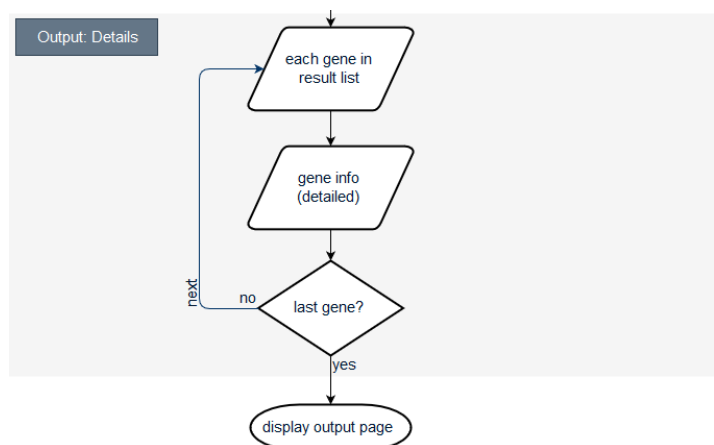


FIGURE 4.1: **Simplified view of MutationDistiller's prioritisation workflow, part 4.** chr: chromosome, pos: position, ref/alt: reference/alternative allele, cov: coverage, DB: database, MT: MutationTaster.

## 5 MutationDistiller: User interface

### 5.1 Input and output pages

#### 5.1.1 Landing page

The first screen users are presented with when opening MutationDistiller at <https://www.mutationdistiller.org> is our landing page that allows them to select whether they wish to a) upload a new VCF file or b) access a previous project through one of our user modes. We organised the landing page in a simple design consisting of clickable bricks to allow easy access to MutationDistiller. Figure 5.1 shows a screenshot of the landing page.

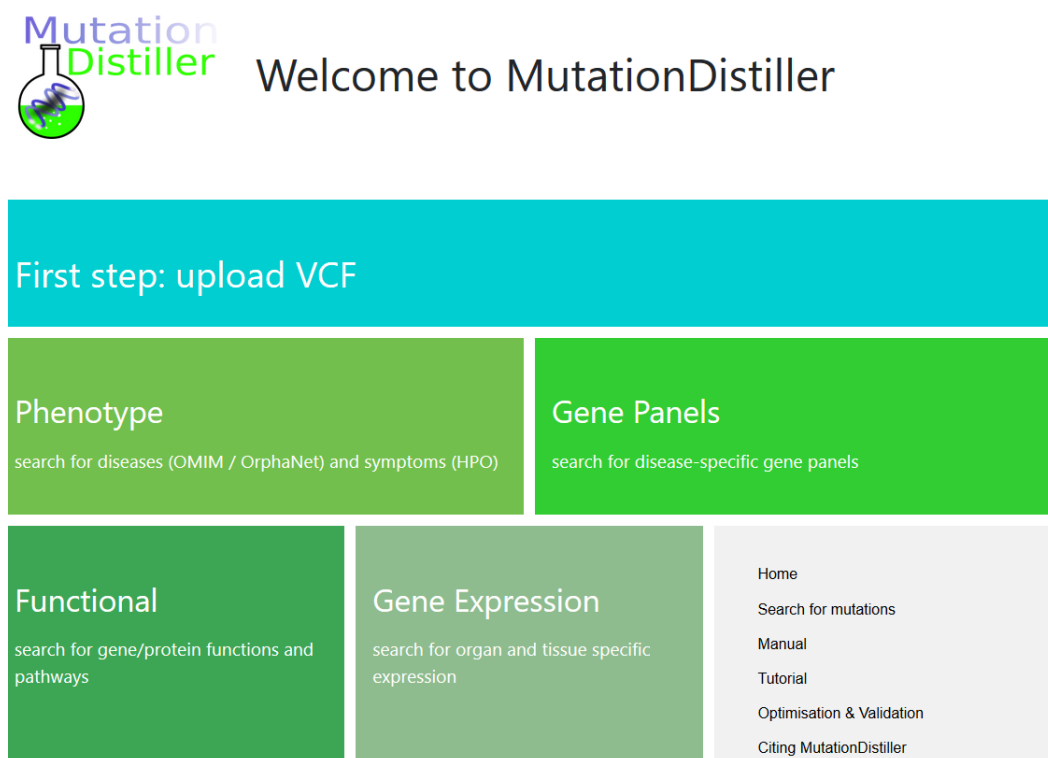


FIGURE 5.1: Screenshot of MutationDistiller’s landing page. This page is the first page users see when calling MutationDistiller and prompts users to either upload a file or to access a previous project.

#### 5.1.2 Query Engine user interface

When clicking on the file upload hyperlink, the user will be redirected to MutationDistiller’s QE (described in chapter 3). On the QE page, users can upload their project’s

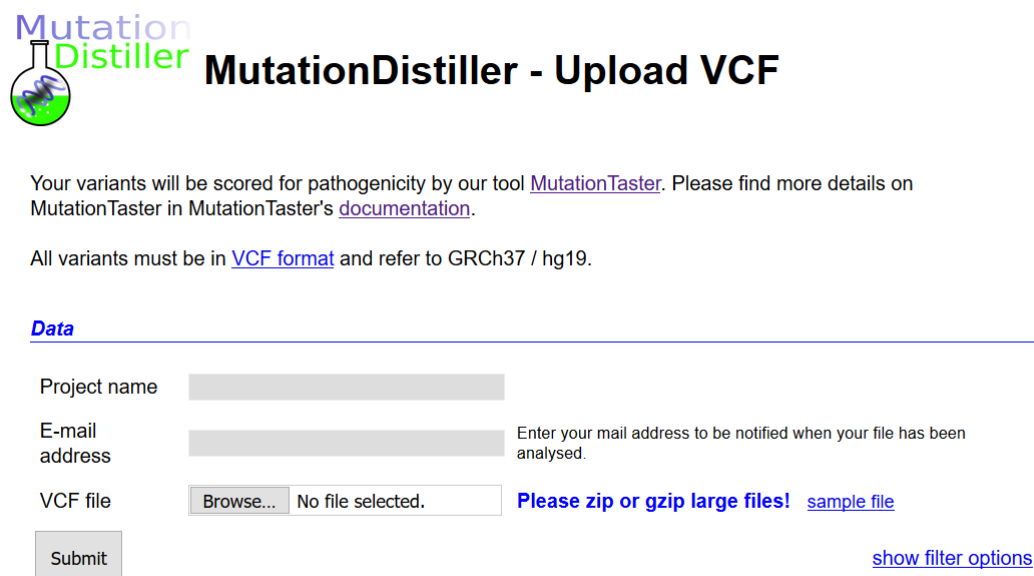
VCF file, enter additional information and pre-filter their variants to speed up the upload process:

**Project name:** A project name can be allocated to make the management of several projects easier.

**Email address:** As upload and analysis might take a short amount of time, we recommend entering an email address. We will send a notification with a project ID for the convenient retrieval of the submitted project at a later time. However, this information is not mandatory as we wish to provide the option for users to remain anonymous.

**Filtering:** The user can decide which types of variants they wish to analyse. Heterozygous variants can be excluded, which might be useful in recessive disorders, especially in consanguineous families. In addition, low-coverage variants may be discarded and polymorphisms stored in the 1000G or ExAC database may be excluded depending on their genotype: By default, variants appearing 4 times in 1000G or 10 times in ExAC in a homozygous state are removed.

**Analysis settings:** The analysis can include the entire VCF file or be restricted to a certain chromosome, region, or to only exons and flanking regions. This feature is aimed at users who have already determined a candidate region via homozygosity mapping, for instance. In addition, users can exclude given areas from analysis to speed up the process. However, as file upload and initial analysis only need a few minutes, we encourage users to upload their entire file and apply region or candidate gene filters at later stages. Figure 5.2 shows a screenshot of MutationDistiller's upload page.



**MutationDistiller - Upload VCF**

Your variants will be scored for pathogenicity by our tool [MutationTaster](#). Please find more details on MutationTaster in MutationTaster's [documentation](#).

All variants must be in [VCF format](#) and refer to GRCh37 / hg19.

---

**Data**

Project name

E-mail address  Enter your mail address to be notified when your file has been analysed.

VCF file  No file selected. [Please zip or gzip large files!](#) [sample file](#)

[show filter options](#)

FIGURE 5.2: Screenshot of MutationDistiller's Query Engine upload page.

### 5.1.3 User modes

The user modes cater for clinicians, human geneticists and researchers coming from different backgrounds, and determine which parts of MutationDistiller’s query interface will be displayed. With each user mode selected from the landing page (figure 5.1), different parts of the the query interface will be shown or hidden:

The query interface page is built from HTML elements called *divs*. By clicking on a link within the page, the visibility setting for a div is changed from *hidden* to *visible*, and the content of the div is displayed to the user. Similarly, when selecting a specific user mode by clicking on one of the clickable bricks from the landing page, the visibility settings for the different interface sections are set to display the relevant areas of the website while hiding others. In addition, MutationDistiller creates hyperlinks to a user’s project and their settings, allowing them to re-load their analysis with all their entries and settings.

For clinicians with a clear idea of their patient’s phenotype, the **Phenotype mode** displays the project section together with the phenotype section. **Gene Panels** displays the project section and the candidate genes, regions, or panels section and is aimed at human geneticists or clinicians with an idea of promising candidate genes. The **Functional mode** shows the gene function section containing GO data as well as pathways, whereas **Expression** opens the gene expression panel.

The user modes are meant to support first-time users who might be overwhelmed by the many options offered by MutationDistiller. However, users do not restrict themselves by selecting one of the modes - all hidden options can easily be displayed, added and selected with one click as described above.

### 5.1.4 Query interface

The query interface is where the user can add their project-specific information and criteria in order to best rank the submitted variants according to the patient’s disease phenotype. Depending on the selected user mode, different parts of the website will be displayed, hidden or pre-selected as described above. Figure 5.3 displays a screenshot of MutationDistiller’s main page in the Phenotype mode (see section 5.1.3). This page is generated by a Perl CGI script, which employs the module *HTML::Template* to dynamically fill a HTML template with values from our database (such as expression data sources), add user entries, and check or uncheck HTML checkboxes depending on user settings.

FIGURE 5.3: **Screenshot of MutationDistiller’s query interface.** MutationDistiller’s query interface in *phenotype* mode, displaying the project and phenotype sections.

The only mandatory information for the program is a MutationDistiller ID plus access code, which was allocated by the Query Engine. This is sufficient to display the most harmful variants in the file. However, a wide range of options are available to sort the data. We organised these options in a number of sections:

### Project

In this section, the project ID is entered. As described before, we decided to allow project access only via the ID plus security code (rather than the name) to ensure unique access – while several users might allocate the same name to their projects, the ID is unique for each case. In addition, the access code ensures privacy and data protection.

The project section also provides the option to select a mode of inheritance and a maximum number of genes to be displayed in the output. Depending on their case, users can filter genes out which do not match the indicated mode of inheritance (*strict* setting) or decide to simply rank matching genes higher.

### Variant Selection

For further refinement, users can select which variants they wish to have considered in their analysis. By default, we include variants labelled ‘pathogenic’ in the ClinVar



database, together with frame-shift mutations or nonsense variants leading to a premature stop-codon, and variants inducing amino acid changes that were designated as disease mutations by MutationTaster. In this selection, we also include variants located within a splice site. In addition, users can decide to include alterations considered disease causing by MutationTaster that are located near a splice site ( $\pm 10$  base pairs), or display all variants predicted to be disease causing. Users can also view all variants in the VCF file, but for performance reasons we only permit this option if the analysis is restricted to a gene region or candidate genes.

By checking a HTML checkbox, simple and complex amino acid variants considered to be harmless by MutationTaster can be included (while excluding known polymorphisms from databases). This setting might be especially useful when recessive inheritance is suspected but only one strong heterozygous candidate mutation is found. In the case of compound heterozygosity, all variant settings can be made separately for the second variant, thus allowing a strict filter for one variant and a more lenient one for the second. Figure 5.4 shows the query interface’s variant selection. The choices a user makes here determine whether a new table for compound heterozygosity (table *comphet\_ID\_VARSSEL*) needs to be generated in the prioritisation protocol as described in section 4.1.1

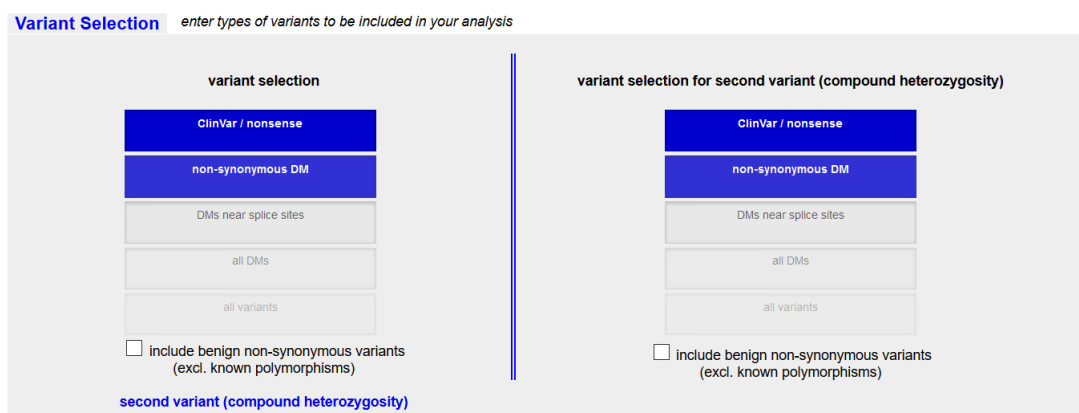


FIGURE 5.4: **Screenshot of MutationDistiller’s variant selection.** This figure shows MutationDistiller’s variant selection section with the selection option for compound heterozygous variants displayed.

### Candidate Genes, Regions, or Panels

This section allows users to restrict the analysis to promising candidate genes. These can be entered manually as a gene list or by region. Moreover, we offer a number of gene panels to be selected here. Several panels can be selected at once to increase the search radius across multiple panels. More information on the included gene panels can be found in the introduction in section 1.4.1.4. The panel resources are loaded dynamically from our database before the page is displayed.

## Phenotype and Gene Function

In these two sections, users can enter identifiers that link to various data sources concerning phenotype information (HPO, OMIM and OrphaNet) or gene function (GO, WikiPathways and Reactome) in auto-completion fields via *AJAX (Asynchronous JavaScript and XML)*. After typing the first four letters of a symptom or disorder, the auto-completion list is loaded and allows users to select the relevant term. The search becomes more precise by entering more letters. Several terms might be clicked at once from the list. In addition, specific non-relevant HPO terms can also be entered via auto-completion in a separate field. Genes linked with these terms are then excluded from analysis. Moreover, highlighting options allow users to decide which MGD disease groups should be stressed in the output.

## Expression

Expression data sets obtained from ExpressionAtlas can be picked in this section. These sets can be selected by developmental status (adult or fetal) or experiment (RNAseq, protein expression or RNA-CAGE). Like the virtual gene panels, the expression resources are loaded dynamically before the page is displayed.

Users can decide if they wish to use the data for filtering, and whether they wish to see expression levels displayed. Data on expression levels are, however, only shown on request to keep the output as lean as possible. If a user decides to display the expression levels, it is denoted whether the expression of a given gene in a tissue of interest is high (within the 75th percentile) or very high (within the 90th percentile). Please refer to section 2.2.5.2 for further details on how we integrated expression data into MutationDistiller.

If the filtering option is selected, genes that are expressed below median in a tissue of interest are not displayed. In this case, candidate genes will only be included in the result list if there is clear data in one of the selected data sources indicating that it is indeed expressed in the given tissue: For example, consider a case where the clinician expects the candidate gene to be expressed in the kidney and decides to only have a look at genes which are clearly expressed in this tissue. MutationDistiller will now remove all genes that are expressed below the median, but also those genes for which no expression data for the kidney is available.

We decided on this restrictive approach to enable convenient filtering, as more lenient filters might possibly still flood the user with too many candidate genes. In our documentation, we explain this feature and recommend to remove the filter (and instead use expression data for display) if a user finds it too restrictive.

### 5.1.5 Output page

MutationDistiller uses all submitted data entered in the sections described above to distil the most fitting variants from the sea of candidates. The steps it undertakes to do this are described in chapter 4. After these steps of filtering, scoring, and sorting have been concluded, MutationDistiller generates a comprehensive output page. This page is divided into three main sections: At the top of the page, a short list reminds the user of their submitted entries. Directly below, a summary table lists crucial data for the resulting variant and gene data. Further down the page, detailed information is listed for each candidate gene. Figure 5.5 shows the user entry list and the summary table, while figure 5.6 displays the gene details.

If only a VCF file was submitted without any further user-defined information, the variants appear in random order, with known disease mutations on top. In all other cases, MutationDistiller displays its data following user instructions: Variants not located in a candidate gene or not scored will be excluded, while the order of all other variants and their genes is determined by MutationDistiller's score.

In the summary table, basic gene information (gene symbol and title) is listed together with reported diseases and mutations. In addition, the overall MutationDistiller score and its percentage of the maximum score reached in the analysis are shown as well as some basic information on variants found in the gene. For each variant, its genomic location and coverage in the submitted VCF file, its genotype, its predicted effect(s) on the amino acid sequence and whether it is a known disease mutation are indicated. Moreover, frequencies in 1000G, dbSNP, and ExAC are listed. Details about the variants' effects can be studied with a hyperlink to MutationTaster's results page for each variant. Basic information from the result table can be exported for external storage and further downstream applications. Figure 5.5 provides a screenshot of an example result table.

Below the result table, MutationDistiller lists more in-depth information on each candidate gene. This information can be accessed by scrolling down or by clicking on the gene symbol in the result table. Here, the user is not only presented with the MutationDistiller score and its composition, but also with detailed up-to-date data on the gene of interest. Moreover, hyperlinks provide access to external sources, allowing users to assess the relevance of a gene with ease.

The data sources used for providing detailed gene information are explained in chapter 2. In the output page, we group these sources in logical sections: First, general data such as ClinVar, Modes of Inheritance, and relevant links are listed. Here, we indicate the overall MutationDistiller score together with this sub-scores. In the next section,

rank	genesymbol	title	score	%	reported diseases & mutations	variants
1	<a href="#">NSD1</a>	nuclear receptor binding SET domain protein 1	2432.9	100%	<ul style="list-style-type: none"> <li>BECKWITH-WIEDEMANN SYNDROME (BWS)</li> <li>NUCLEAR RECEPTOR-BINDING SET DOMAIN PROTEIN (NSD1)</li> <li>SOTOS SYNDROME (SOTOS1)</li> <li>5q35 microduplication syndrome</li> <li>Beckwith-Wiedemann syndrome due to NSD1 mutation</li> <li>Deletion 5q35</li> <li>Sotos syndrome</li> <li>Weaver syndrome</li> <li>germline, loss of function, role in phenotype, candidate gene tested, autosomal dominant, autosomal recessive</li> </ul>	<a href="#">5:176638438C-G</a> <b>het</b> <a href="#">IGV</a> <a href="#">170x</a> <a href="#">NMD1PTC_S910_S1013_S744*</a> <i>neither in ExAC nor 1000G</i>
2	<a href="#">JUP</a>	junction plakoglobin	1293.7	53%	<ul style="list-style-type: none"> <li>ARRHYTHMOGENIC RIGHT VENTRICULAR DYSPLASIA, FAMILIAL (ARVD12)</li> <li>JUNCTION PLAKOGLOBIN (JUP)</li> <li>NAXOS DISEASE (NAXD)</li> <li>Familial isolated arrhythmogenic ventricular dysplasia, biventricular form</li> <li>Familial isolated arrhythmogenic ventricular dysplasia, left dominant form</li> <li>Familial isolated arrhythmogenic ventricular dysplasia, right dominant form</li> <li>Ethral acantholytic epidermolysis bullosa</li> <li>Naxos disease</li> <li>germline, autosomal recessive</li> </ul>	<a href="#">17:39778762A-C</a> <b>het</b> <a href="#">IGV</a> <a href="#">19x</a> <a href="#">F332V</a> , splicing impaired <a href="#">rs198897890</a> <b>hom</b> carriers <a href="#">1000G</a> - - <a href="#">ExAC</a> 0 769
3	<a href="#">ANOS</a>	anoctamin 5	797.5	32%	<ul style="list-style-type: none"> <li>ANOCTAMIN (ANOS)</li> <li>GNATHODIAPHYSEAL DYSPLASIA (GDD)</li> <li>MUSCULAR DYSTROPHY, LIMB-GIRDLE, TYPE (LGM2)</li> <li>MUSCULAR DYSTROPHY, LIMB-GIRDLE, TYPE (LGM2L)</li> <li>Autosomal recessive limb-girdle muscular dystrophy</li> <li>Distal anoctaminopathy</li> <li>Gnathodiaphyseal dysplasia</li> <li>germline, autosomal dominant, autosomal recessive</li> </ul>	<a href="#">11:22225388T&gt;C</a> <b>het</b> <a href="#">IGV</a> <a href="#">13x</a> splice site <i>neither in ExAC nor 1000G</i>
4	<a href="#">EVC2</a>	EVC ciliary complex subunit 2	797.5	32%	<ul style="list-style-type: none"> <li>ELLIS-VAN CREVELD SYNDROME (EVC)</li> <li>EVC2 GENE (EVC2)</li> </ul>	<a href="#">4:5624281C&gt;T</a> <b>het</b> <a href="#">IGV</a> <a href="#">12x</a> <a href="#">W828*</a> , <a href="#">NMD1PTC_W748*</a>

FIGURE 5.5: Screenshot of MutationDistiller’s result table. This figure shows the result table for a HPO-based example project.

pathway data and protein information are listed. If a pathway receives a match, this term will be scored and highlighted in bold. The following section provides information on symptoms and diseases: For HPO, all terms linked with the gene are listed. Direct, ancestor or descendant matches are highlighted and their score is indicated. For OMIM, GO and Orphanet, we list all entries for the gene highlight matches. Finally, we list Ensembl transcripts with hyperlinks to the relevant Ensembl webpage. If a user selected to display additional information (such as expression data), we list these data below. Figure 5.6 provides a screenshot of the detailed gene view.

genesymbol type	description	chr.	startpos	endpos	synonyms
<a href="#">NSD1</a> #1	protein-coding <b>nuclear receptor binding SET domain protein 1</b>	5	176560026	176727214	SOTOS, FLJ44628, FLJ22263, ARA267, DKFZp666C163, STO, KMT3B,
reported mutations	germline, loss of function, role in phenotype, candidate gene tested, autosomal dominant, autosomal recessive				
overall score		2432.9	100%		
HPO		2432.9			
links	<a href="#">NCBI</a> <a href="#">ENSEMBL</a> <a href="#">SwissProt</a> <a href="#">GeneCards</a> <a href="#">STRING</a> <a href="#">PubMed</a> <a href="#">PubMed-phenotype</a>				
KEGG pathways	<a href="#">Lysine degradation</a>				
WikiPathways	<a href="#">Histone Modifications</a> , <a href="#">Pathways Affected in Adenoid Cystic Carcinoma</a>				
PFAM	<a href="#">PHD-finger</a> , <a href="#">PWWP domain</a> , <a href="#">SET domain</a>				
InterPro domains	<a href="#">PWWP domain</a> , <a href="#">SET domain</a> , <a href="#">Zinc finger, PHD-type</a> , <a href="#">Post-SET domain</a> , <a href="#">AWS</a> , <a href="#">Zinc finger, FYVE/PHD-type</a> , <a href="#">Zinc finger, PHD-finger</a>				
paralog	<a href="#">SUV39H1</a> (3%), <a href="#">NSD2</a> (23%), <a href="#">SETDB1</a> (5%), <a href="#">EHMT2</a> (6%), <a href="#">SETBP1</a> (8%), <a href="#">SETD2</a> (10%), <a href="#">NSD3</a> (22%), <a href="#">ASH1L</a> (14%), <a href="#">SUV39H2</a> (21%), <a href="#">EHMT1</a> (6%), <a href="#">SETDB2</a> (3%)				
HPO	<ul style="list-style-type: none"> <li><b>Large hands</b> direct match score: 348.9 <a href="#">collapse</a></li> <li><b>Muscular hypotonia</b> direct match score: 12.0 <a href="#">collapse</a></li> <li><b>Global developmental delay</b> direct match score: 12.1 <a href="#">collapse</a></li> <li><b>Overgrowth</b> direct match score: 620.3 <a href="#">collapse</a></li> <li><b>Cardiomyopathy</b> direct match score: 134.5 <a href="#">collapse</a></li> <li><b>Neonatal hypoglycemia</b> direct match score: 507.5 <a href="#">collapse</a></li> <li><b>Advanced eruption of teeth</b> direct match score: 797.5 <a href="#">collapse</a></li> <li>[all] stature parent 1 score: 0.0 <a href="#">collapse</a></li> <li>Macrocephaly parent 1 score: 0.0 <a href="#">collapse</a></li> <li>Hypoglycemia parent 1 score: 0.0 <a href="#">collapse</a></li> <li>Neonatal hypotonia child 1 score: 0.0 <a href="#">collapse</a></li> </ul>				
OMIM	<a href="#">SOTOS SYNDROME 1 (SOTOS1)</a> phenotype (molecular basis known) <a href="#">117550</a> synopsis:				
	<b>INHERITANCE:</b> Isolated cases <b>GROWTH:</b> [Height] Mean full term birth length 55.2cm Length at or greater than 97th percentile through early adolescence Adult height often normal Mean male adult height 184.9cm				
	<a href="#">BECKWITH-WIEDEMANN SYNDROME (BWS)</a> phenotype (molecular basis known) <a href="#">130650</a> synopsis:				

FIGURE 5.6: Screenshot of MutationDistiller’s detailed view. The detailed view provides an insight into MutationDistiller’s scoring system together with in-depth data for every candidate gene.

To enable flexible analyses, the program also allows for interactive refinement of the

search: A hyperlink takes the user back to the entry page but keeps the previously entered terms and selections. This hyperlink can be bookmarked to resume the analysis later and to exchange prioritisation settings with colleagues. Moreover, HPO terms can be added or excluded flexibly without having to re-load the entry page. Both features are achieved using CGI scripts that hand the selected values and properties to the relevant scripts for the generation of the interface.

## **5.2 Manual and tutorial pages**

In order to make it as easy as possible for users to start working with MutationDistiller, the tool comes with extensive documentation and tutorial pages. On these pages, we explain how to get started with MutationDistiller and provide information on updates and changes. The tutorial is a step-by-step analysis of an example case, which is intended to get users acquainted with MutationDistiller's many options. The tutorial can be found at <https://mutationdistiller.org/info/tutorial.html>, the manual is located at <https://mutationdistiller.org/info/documentation.html>. Both can be accessed easily at any stage through hyperlinks.

## 6 Implementation and Tools

### 6.1 Software development

We developed MutationDistiller in an iterative fashion: Instead of following a detailed plan, we took user input and newly emerged data sources into account during each step of software development. In this way, we ensured that the resulting program would be up-to-date and easily accessible for the intended users.

#### 6.1.1 MutationDistiller

The program and functions of MutationDistiller were written in the programming language Perl. Central modules – collections of functions – contain all the relevant sub-routines, grouped by their purpose for the program. A number of freely available Perl modules were incorporated into the program, which we obtained via the operating system’s package management system or the central Perl repository CPAN (Comprehensive Perl Archive Network)<sup>1</sup>:

- Apache2::Reload
- CGI
- CGI::Carp
- DBD::Pg
- DBI
- Email::Valid
- Encode
- HTML::Entities
- HTML::Template
- JSON
- Mail::Sendmail
- Net::SMTP::SSL
- PBS::Client
- Sort::Naturally
- Statistics::Basic::Correlation
- Time::HiRes

---

<sup>1</sup>[www.cpan.org](http://www.cpan.org)

### 6.1.2 Query Engine

The MutationDistiller QueryEngine (QE) was written in Perl as well. Job scheduling is handled by the TORQUE Resource Manager<sup>2</sup>. The single Perl scripts that make up MutationDistiller's QE are called via shell scripts. The user submissions entered in the start page are being read out using the Perl module CGI, the communication with TORQUE is handled by the Perl module PBS::Client.

## 6.2 Manuscript

The Entity Relationship Diagrams (figure 2.3 and Appendix C) displayed in this thesis were generated with the great, freely available database tool DBeaver (<https://dbeaver.io/>, accessed 28.12.2018). I conducted statistical analyses, designed plots and printed appendix tables using the programming language R (version 3.4.3) [113] and its packages plyr [114], dplyr [115], xtable [116], ggplot2 [117] and reshape [118]. The flowcharts describing the workflow of MutationDistiller's prioritisation algorithm and its Query Engine (figures 4.1 and 3.1) were generated with the freely available online diagram software draw.io (<https://www.draw.io/>, accessed 12.06.2019).

## 6.3 Hardware

All MutationDistiller applications run on a 48-CPU system with 512 GB RAM under Linux (CentOS 6). All program scripts are written in Perl (5.10) and run in an Apache 2.2 web server with modperl2. User interfaces are written in HTML with JavaScript functions. The database is run on PostgreSQL 9.5.

---

<sup>2</sup><http://www.adaptivecomputing.com/products/open-%20source/torque/>, accessed 20.12.2018

## 7 MutationDistiller: Optimisation and validation

### 7.1 Determination of HPO weights

#### 7.1.1 Training Data

We built MutationDistiller to find the most likely disease causing candidate(s) from a sea of potentially harmful alterations. In this endeavour, the tool will be faced with a large variety of genetic variants, which may be linked with many different diseases or phenotypes. To get MutationDistiller up and ready for the task, we trained and optimised its HPO score and weights using information that represents and resembles the data it will encounter in real-life cases.

As real patient data and disease-gene connections are hard to come by or not available due to data protection issues, current variant prioritisation tools have usually been trained and optimised using somewhat artificial data sets. For example, PhenIX [75] was developed and tested by randomly selecting HPO terms from the list of HPO terms annotated for a gene of interest. eXtasy [71], on the other hand, used gene-phenotype associations generated by the tool Phenomizer [119], a procedure that guarantees ideal associations which do not usually occur in clinical day-to-day life.

We have optimised the matching procedure of HPO terms by choosing an approach that attempts to be as realistic as possible while still accounting for data and patient protection: The variant database ClinVar [44] contains a range of disease mutations with HPO identifiers as associated phenotype information. These identifiers have been submitted by users – mainly clinicians and researchers – and thus can be expected to resemble a real-life situation more closely than artificial data. Moreover, ClinVar data covers a relatively wide range of different diseases and gene groups, thus enabling to represent various medical fields. We obtained all ClinVar entries with at least two HPO terms that were labelled as *pathogenic*. In total, we were able to compile a set of 188 cases linked with 142 different genes. We then integrated the ClinVar alterations into a freely available 1000G exome VCF file (HG00377) and sent the resulting VCF files to MutationDistiller in order to optimise the HPO scoring system.



### 7.1.2 HPO weight parameter selection

As described in section 4.1.2.2, the ontology structure of the HPO allows us to base the HPO scoring on information content. In addition, we can not only score direct matches, but also ancestor and descendant terms. This helps to account for phenotyping errors and gaps: If a user enters a certain term for which a gene is not directly annotated, it would not be scored at all, even if the first descendant is annotated. Therefore, we devised a system that scores direct matches, but also ancestors and descendants. In order to evaluate in which way they should be weighted against each other to reach optimal results, we used the ClinVar-set described above to iterate through a range of weight combinations (245 combinations in total) between the three categories (direct matches, ancestors, descendants). We chose this approach rather than dynamically searching for the optimal weight distribution to avoid overfitting on this relatively small data set.

Table 7.1 shows the different weights we combined for the three categories. We then recorded the ranks given to the genes containing the indicative alteration and only regarded the first 100 ranks, labelling any cases beyond that as 'not found'. Genes with the exact same score were given the same rank – this means that for each case, usually more than 100 genes were included in the analysis.

HPO match type	weight
direct	0.2, 0.5, 1, 2, 5
ancestor/descendant	0, 0.05, 0.1, 0.2, 0.5, 1, 2

TABLE 7.1: **HPO optimisation weights.** Displays the different weights tested for direct, ancestor and descendant hits.

For all cases in which the disease mutation was found, we then observed the rank distribution for the disease genes across all weight combinations. We only regarded the first 100 ranks, denoting any cases beyond that as 'not found'. Genes with the exact same score were given the same rank. For each combination, we evaluated how many indicated disease genes were ranked on rank 1, ranks 1 to 5, greater than 10, or not found at all. We also calculated the mean rank for all disease genes across each combination.

We found that, for each of the various weight-combinations, a relatively high number of cases (at least 22.3%) could not be solved, indicating that the phenotypes entered into ClinVar are not always identical to the phenotypes associated with the disease genes. This is consistent with a real-life situation in the clinic with phenotyping errors and inconsistencies. In addition, it was evident that a high weight for direct hits was consistently better at ranking the alteration of interest amongst the top positions.

We then assessed the resulting weight-combinations under specific considerations to find the best suited solution:

- a) Balanced for direct hits, ancestors and descendants:** To represent all three categories in the weighting process, we excluded all candidates with zero-scores in one of the categories, but included small values (0.05 at the lowest).
- b) Able to detect the causative variant:** We only included combinations for closer consideration that were able to find the disease-relevant gene in the majority of cases. As mentioned above, a high fraction of cases (minimum 22.3%) were not found within the top 100 in any of the weight combinations.
- c) Low mean rank:** To ensure that the gene of interest frequently shows up in the top ranks, we excluded combinations with a high mean rank across all disease-relevant genes.
- d) Causative variants on rank 1:** In addition, we ensured that a large proportion of the genes of interest get ranked on top.

Together, we expect these criteria to ensure that MutationDistiller is capable of ranking the most likely candidate genes within the top ranks for a majority of the cases the program encounters. After careful consideration, we decided on a weight of 5 for direct hits, 0.05 for descendants and 2 for ancestors as this combination showed a comparatively low loss-rate (22.8%) while ranking the indicative genes on the top rank in 37.2% of the cases. In addition, it reached a low mean rank for the genes of interest (5.82). We then incorporated this combination into MutationDistiller and used this version of the program for further testing and comparison with other tools (see below). We also found that while the loss rate was relatively high to begin with, in the groups with the lowest mean rank for the indicative genes it did not change much – it was always around 23%. This indicates that the loss rate is not linked with the weighting but with the phenotyping or the HPO annotation process itself.

Table 7.2 shows the parameters of the weight-combinations with the lowest loss rate. A summary of all tested combinations can be found in table A.1 in the appendix.

weights			indicated disease mutation			
direct hit	ancestor	descendant	top rank	top 3	not found	mean rank
1	2	0.5	0.04	0.25	0.223	9.5
0.2	0.5	0.1	0.03	0.21	0.223	10.1
0.2	1	0.1	0.02	0.21	0.223	11.5
0.2	2	0.1	0.02	0.13	0.223	11.9
5	0.05	2	0.37	0.5	0.228	5.8

TABLE 7.2: **HPO weight iterations.** Displays the top iterations for direct hits, ancestor and descendant weights by minimal drop-out rate (not found: disease mutation not listed within the first 100 ranks) for the ClinVar-HPO set.

## 7.2 Testing and validation

### 7.2.1 Test set

To test how MutationDistiller would fare in a real-life scenario, we compiled a test set of 101 existing patient cases from Charité Berlin. These cases of rare, early onset genetic disorders were provided by clinicians and researchers working in the Neuropaediatrics and the Medical Genetics departments. The patient had given consent for research use. The clinicians provided patient VCF files together with the causative variant(s) and the relevant genes, the HPO terms that were used in the quest to find the disease-relevant alteration, and information on the expected mode of inheritance (if available). We ensured that there was no overlap between the ClinVar cases used for program optimisation and the validation data set.

We had originally planned to compare MutationDistiller with online versions of other tools. To account for patient data protection, we hence spiked the known causative variant for each case into the same 1000G VCF file used for optimisation of MutationDistiller (HG00377). Due to performance reasons, however, we had to rely on downloaded versions of the program.

### 7.2.2 Validation

We sent the resulting VCF files containing the disease mutation(s), the HPO identifiers and mode of inheritance information submitted by the clinicians to MutationDistiller to validate its performance on this real-life data set. For the test, we used the weight settings determined in the optimisation step. For the mode of inheritance, we chose the same weight as for a direct HPO match (5) to avoid it being underrepresented. The goal for this test was to determine MutationDistiller's capabilities for detecting disease-relevant alterations in a HPO-centric search and to compare them with other state-of-the-art tools. In MutationDistiller, known pathogenic variants from the ClinVar database are given a ClinVar match score as described in 4.1.2. However, the tools included into this comparison do not provide this function. We therefore decided to not allocate MutationDistiller's ClinVar score at this stage. However, this means that in real life, the MutationDistiller results can be expected to be slightly better for known disease mutations.

We then observed which rank the gene containing the known disease-relevant alteration was given by the program. As in the optimisation step, we only regarded the first 100

ranks, labelling cases where the gene of interest could not be located within these ranks as 'not found'. For genes that obtained the exact same score, we allocated the same rank. We found that MutationDistiller was capable of finding all but one of the disease-relevant genes within the first 100 ranks. In total, MutationDistiller reached a mean rank of 6.52 for the indicative gene across the test set. In the vast majority of cases, the disease gene indicated by the clinicians was ranked within the first 10 ranks (82.2%). Table 7.3 shows the number of disease genes ranked within ranks 1 to 10 for the set of 101 cases.

rank	number of genes	cumulative
1	39	39
2	9	48
3	9	57
4	5	62
5	7	69
6	4	73
7	6	79
8	2	81
9	2	83
10	0	83

TABLE 7.3: **Validation set ranks.** Shows the first 10 ranks for the validation set as absolute and cumulative numbers. If several genes reach the same rank, they are all allocated the best rank. Total number of cases: 101.

As can be seen in this table, over half of the indicated disease genes were ranked within the first 3 ranks by MutationDistiller, and over two-thirds were ranked within the first five ranks. The distribution across all ranks from 1 to 100 is depicted in figure 7.1.

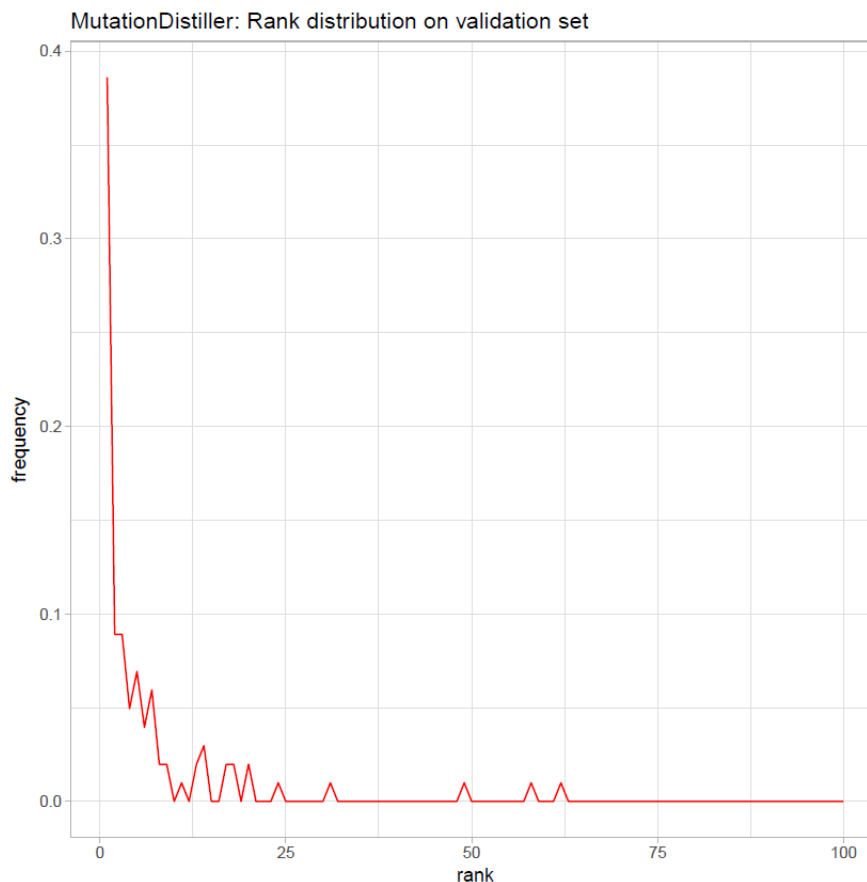


FIGURE 7.1: **MutationDistiller rank distribution for the validation set.** Displays the rank distribution (in percentage) for the indicated disease genes in the validation data set. The cumulative distribution can be found in the following figure, 7.2.

### 7.2.3 Comparison with state-of-the-art prioritisation tools

To assess MutationDistiller’s performance in comparison to other, frequently used variant prioritisation tools, we decided to send the validation data set to a number of similar programs. We included algorithms into our test that are freely available online and do not require any software installation or user login (to avoid data security issues). In addition, we only included tools that can work with VCF files and offer HPO-centric prioritisation. We also excluded candidates such as Phen-Gen [72], which requires trio VCFs (usually data from unaffected parents and an affected child) as this would have substantially reduced the potential test cases. Moreover, we had to remove a number of tools that unfortunately were not functioning at the time of testing.

We were thus able to compare MutationDistiller to three different algorithms, the PhenIX [75] and HiPhive [120] algorithms incorporated into Exomiser [121] as well as eXtasy [71]. We used Exomiser version exomiser-cli-10.0.1 and the eXtasy version 2013-07-04 (the

latest version available from their GitHub page<sup>1</sup>. We stuck to default settings, which is what an untrained user is expected to do. For each of these algorithms, we had to rely on locally installed versions as the online tools were not working reliably or fast enough for our purposes.

The Exomiser generates a number of different scores, the case-relevance of which is not easily obvious to users from the clinic. Hence, we decided to limit our assessment to the so-called 'Exomiser gene pheno score', which we deemed to be most fitting to the task at hand, namely the matching of genes to phenotype data. As the eXtasy algorithm is not capable of working with all HPO terms, for this tool we removed the terms not found in eXtasy's database from our set. This limited our eXtasy analysis to 88 cases. Moreover, eXtasy's entry options are limited to 10 HPO terms per case. We thus randomly removed all terms exceeding 10 from the 7 cases where this was necessary.

To assess and compare the capabilities of the different algorithms, we sent the validation set (VCF files, HPO annotations and mode of inheritance information, if available) to them and recorded, for each case, the rank of the indicated disease gene. For eXtasy, we had to distinguish between cases in which only one HPO term was used for analysis and cases with more than one term. In single HPO cases, we ranked the files by ordering them by the result score; in combined cases we ranked them by the provided statistical score as the program outputs a result score for each HPO term separately.

We then examined which proportion of cases were ranked at which position and compared the outcomes between the different programs. To ensure that the results from the various algorithms can be compared, we also capped the search at rank 100, as for the MutationDistiller test. Cases in which the gene of interest was not located within the first 100 ranks we hence considered to not have been solved.

When comparing the cumulative ranks allocated to the disease genes, we found that eXtasy failed in a large majority of the provided cases. To start with, due to the lack of HPO terms in its database, the analysis was limited to 88 cases of the 101 test cases. Of these cases, eXtasy found less than 30% of the causative alterations within the first 100 ranks, which might be due to the fact that the underlying gene-phenotype associations were updated more than 5 years ago.

For HiPhive and PhenIX, we found that those algorithms were capable of detecting the causative gene within the top 100 positions in the majority of cases. However, MutationDistiller was capable of solving considerably more cases than the other tools (99% for MutationDistiller, 81.2 % for PhenIX and HiPhive). This was shown to be the same for genes of interest that were ranked within the first 10 (82.2% for MutationDistiller, 68.3%

---

<sup>1</sup><https://github.com/asifrim/eXtasy/blob/master/README>, accessed Aug 2018)

for PhenIX, 63.3% for HiPhive) or 20 positions (94.1% for MutationDistiller, 73.3% for HiPhive, 70.3% for PhenIX). Figure 7.2 displays the accuracy of the tested tools as the cumulative percentage of indicated disease genes ranked within each rank group from top 1 to top 100. To obtain this figure, we calculated the cumulative percentage of correct disease genes ranked within each group (on the first rank, within the first two, three, four, and so on) and plotted the distribution up to rank 100 for each of the four tested tools.

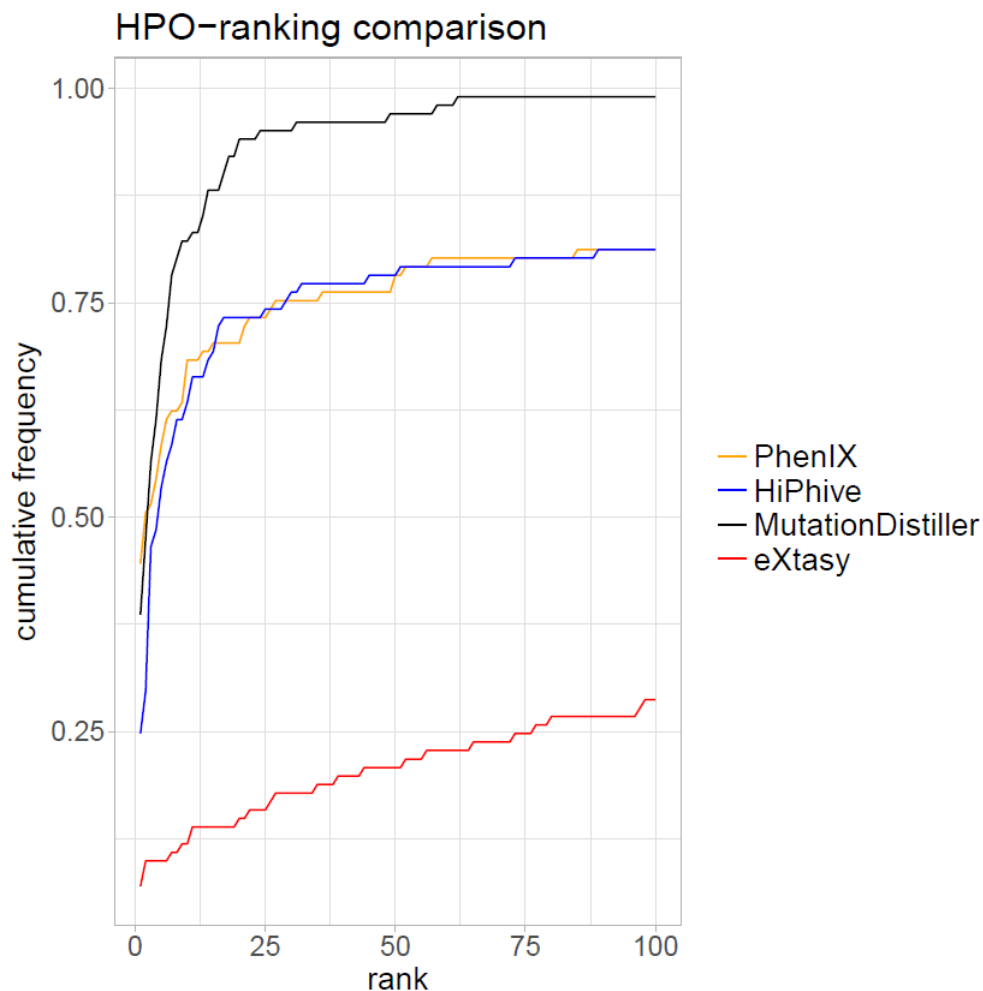


FIGURE 7.2: **Tool Comparison: Cumulative rank frequencies.** Cumulative rank frequencies for the HPO-based detection of disease mutations in a set of 101 patient files for MutationDistiller (black), PhenIX (orange), HiPhive (blue) and eXtasy (red). For each tool, the accuracy is depicted as the cumulative percentage of indicated disease genes sorted within each rank group (top 1 to top 100). Published in Hombach D *et al.* MutationDistiller – user-driven identification of pathogenic DNA variants. NAR Web Server Issue. 2019. doi:10.1093/nar/gkz330

## 8 Discussion

### 8.1 Data selection process

MutationDistiller is a tool to prioritise genetic variants based on genetic, clinical and biological data. As such, its performance and success stand and fall with the data it bases its decisions on. These data come in two kinds: On the one hand, the program depends on the quality of the integrated data sources; on the other hand its scoring and sorting success relies on the training and optimisation cases used during program development. In addition, MutationDistiller's performance also heavily depends on the quality of the phenotyping, but the responsibility lies with the user and cannot be addressed by us.

In order to span a wide variety of cases and to cater to needs from different user groups, we decided to include a wide range of data and information sources (see section 2). We integrated up-to-date data covering a plethora of genetic fields. As some of these sources are still based on genome version GRCh37 – which has also been used by all groups which were involved in the development of MutationDistiller – we also decided to base MutationDistiller on this genome build. Even though the later build GRCh38 has been available for several years now, it has not yet completely entered the field: some secondary data sources employed by MutationDistiller and MutationTaster, such as the ExAC data, are based on the previous build. However, in the future, with more and more potential users and secondary data sources migrating to the new build, we plan to also update MutationDistiller and its databases to accommodate scientific advances made in deciphering the human genome.

#### 8.1.1 Integrated data types

In the following, I will discuss the main data sources and the reasoning behind choosing them for inclusion into the program.

##### **Phenotype data**

We decided to develop MutationDistiller as a phenotype-based prioritisation tool based on ontologies and disease repositories such as OMIM [59], OrphaNet [60] and especially the HPO [58], as they are widely used and well-accepted in the field of human genetics. For the HPO, this wide acceptance is mirrored by the number of HPO-based approaches and tools that help users in the phenotyping effort, e.g. Phenomizer [119], PhenoTips



[122], or Phenotero [123], which was developed in our group. In addition, a number of phenotype-driven platforms such as DECIPHER [95], DDD (Deciphering Developmental Disorders) [124] or Phenopolis [125] base their analysis on the HPO as well.

However, if a complete diagnosis is available for a patient, translating this into sets of phenotypes (i.e. HPO terms) adds an unnecessary layer of uncertainty. In those cases, it makes more sense and is much easier for the clinician in charge to simply choose the relevant diagnosis for data analysis. Because we did not wish to limit our users to HPO data, we also integrated OMIM and OrphaNet data. We expect the existence of a secured clinical diagnosis to be of greater relevance for a case than a collection of HPO symptoms. Thus, we allocate a much higher weight to Orphanet or OMIM entries. However, it has to be noted that the weighting system is not set in stone and will be updated and improved as user feedback comes in.

In addition, we also include a phenotype-genotype resource not linked with human data; the MGD [126]. This repository stores data on phenotypes observed in mice and their genetic background. Mice have been used as model organisms in genetic disease for a long time. As a consequence, a large set of phenotype-genotype connections are known; more than for humans: While the HPO contains around 12,000 genotype-phenotype relations, The MGD stores over 300,000 mouse phenotype annotations<sup>1</sup>. This knowledge can be particularly helpful in the discovery of unknown disease genes. The limitation to known disease genes is one of the main drawbacks when relying on repositories such as the HPO, OMIM or Orphanet, as it limits the number of cases that can be solved by these means. We therefore incorporated 20 disease classes from the *MGD:human disease portal* into MutationDistiller, hoping to enable the detection of new disease genes.

Together, we expect the phenotype data available on MutationDistiller to enable flexible analyses while limiting options to the most relevant and reliable data sources.

### Gene panels

While NGS methods allow the analysis of entire exomes or even genomes, they come at a high cost: They generate large data sets of variants which all have to be considered for further testing. One way to get around this is to apply *virtual panels*, which allow users to restrict their search to certain candidate genes. This panel-based usage of NGS data has previously been suggested as a time- and cost-effective method [101]. We thus decided to include this option in MutationDistiller.

While incorporating commonly used panels, we decided to not allow MutationDistiller users to deposit their user-specific panels on our servers as this would cause issues of data privacy and usability: Either, these panels would have to be available and visible

---

<sup>1</sup><http://www.informatics.jax.org/>, accessed 05.06.2019

to all users, or each user would have to create an account to keep their panels safe and secret. Therefore, instead we decided to allow users to simply upload or copy their gene lists whenever they run an analysis.

### Gene function data

While most tools with a similar aim to MutationDistiller focus their options on disease symptoms (i.e. via the HPO), we decided to offer to search via gene function data such as the Gene Ontology (GO), gene expression or metabolic pathways.

The GO offers comprehensive data on the function and properties of genes. By including this resource as an option, we allow users to identify genes of disease relevance which have not yet been identified as such. While symptom- or diagnosis-based searches (HPO, OMIM, Orphanet) are of great importance in routine clinical cases, they do not allow the detection of hitherto unknown disease genes (see also section 8.3).

In addition, in a number of rare diseases, only specific tissues or organs are affected. An example is *Cutis Laxa*, a group of connective tissue disorders that manifest in the skin. In those and similar cases, the search for the causative gene can be rendered easier by limiting the search to genes that are expressed in a tissue of interest (e.g. the skin for *Cutis Laxa*). While this is not yet exploited by most tools, we are convinced that this feature can be of great help in a number of cases. We thus decided to offer users the option to include expression information for their analysis as an alternative method, or in addition to other data. We incorporated a number of data sets obtained from ExpressionAtlas as this source offers curated sets while including a wide range of experiments (see chapter 2.2.5.2 for more information).

Currently, MutationDistiller allows users to search and filter their data for genes that are expressed in a tissue of interest. It might be conceivable, however, that users are interested in genes that are explicitly not expressed in a given tissue, for example in the case of promoter mutations. In the future, we are planning to include a feature into MutationDistiller that enables users to find genes that are not expressed (or expressed below median) in a tissue or tissue group.

Transcriptome data – data on all (m)RNA molecules present in one cell or a population of cells – depicts the amount of gene expression that is present at a given time. It allows us to see gene expression changes in tissues as affected by disease, enabling detailed assessments of affected organs or systems. As disrupted gene expression is known to play a role in rare disease [127, 128], including transcriptome data into analysis has been found to increase diagnostic yield in rare disease [129, 130] as well. Currently, the availability of public transcriptome data – especially linked with monogenic diseases – is limited. Recently, however, we have observed an increase in the research on transcriptomes of

rare disease patients [131–133] and we expect this field to grow further in the future. Thus, with the rise relevant data sources, we are planning to include these options as well.

Similarly to gene expression, certain pathways are known to be involved in the development of specific rare disorders. For instance, the pentose phosphate pathway (PPP) has recently been found to be affected in rare kidney diseases [134] while the *mammalian target of rapamycin (mTOR)* pathway is a frequent target in neurodevelopmental disorders [135–137]. Thus, including pathway data might improve diagnostic yield in certain cases. We currently include three main data sources: Reactome and WikiPathway can be used for active analyses while KEGG pathway data is displayed in the results. As KEGG pathway information is no longer freely available, the data included into MutationDistiller is somewhat dated. We therefore decided to only display the data rather than allowing users to actively search for KEGG pathways. In this way, clinicians have the advantage of being able to include the latest pathways into the ranking of their candidate variants while at the same time having the long-standing information of KEGG at their disposal.

The ability to include expression or pathway data directly within MutationDistiller instead of having to go through the HPO can thus free users from an additional load of work. While the HPO is currently one of the main used resources, it can be difficult to use and cause problems due to the complex procedure of phenotyping. We are therefore convinced that the option to not have to rely on the HPO can help many users – especially non-geneticists who suspect a genetic cause in a patient’s disease – in their daily work.

### 8.1.2 Testing and training data

The second type of data shaping MutationDistiller, the cases used for training, optimising, and testing, were selected in an attempt to mirror real-life patient cases as closely as possible. As described in chapter 7, the program’s HPO score was developed using variants with known phenotype associations obtained from ClinVar and tested using actual patient data from the Medical Genetics and Neuropaediatrics departments at Charité.

Our rationale for this patient-centred approach was the drive to develop a tool that is able to model real-life cases as truthfully as possible. Due to the aforementioned lack of combined genotype and phenotype data, most other similar tools have been developed using somewhat artificial data. PhenIX [75], for instance, was tested using modified sets of HPO terms generated from the OMIM gene entries – which is not the same as clinical sets with their errors and lack of exactness. Exomiser [73], on the other hand,

was tested with curated, 'optimal' HPO sets as obtained from HGMD. Both approaches deviate from the real situation found in the clinic: Here, a clinician examines a patient and then describes their symptoms in medical terms that can then be translated into HPO identifiers.

In daily clinical routine, clinicians rarely encounter the model patient who displays all (and only those) symptoms listed for a specific disease. In reality, the symptoms of two patients diagnosed with the same disorder can be quite different. In addition, even the lists of symptoms assigned to the same patient by different physicians might differ strongly, which is what we found in several cases in our data. These problems found in phenotyping cannot be reliably mirrored by artificial means but severely influence the outcome of a WES analysis. We thus decided for a different approach, even though this restricts our training and testing data to a relatively small number of cases, since, largely due to data regulation and confidentiality issues, real patient data of disease-relevant alterations and phenotype links is difficult or impossible to use. For MutationDistiller, we were able to collect 188 training cases from ClinVar and 101 in-house clinical data sets from patients who had given consent to scientific use of their data. While under the given circumstances we considered these data sets to be large enough for our purposes, the program's performance could potentially be improved by adding more training or test cases. We thus hope to be able to increase the data set sizes in the future. However, we are convinced that the benefits of having realistic data outweigh the comparably low number of test cases.

## 8.2 Scoring

### 8.2.1 HPO score optimisation

MutationDistiller scores and weighs variants and their genes according to user-defined criteria and sorts them accordingly. The tool covers a wide range of information sources and weighs the different types of data against each other. We decided on a pre-set scoring system that cannot be altered by the user as we learned from a previous program developed in our group, GeneDistiller [81], that allowing users to manually alter weights can be overwhelming and alarming for new users, even if they never apply any changes. Thus, MutationDistiller does not support this option. Instead, on the result page, the program lists in detail which data source contributed to the score in which way. In this way, the tool enables users to draw their own conclusions on the reliability of their scores. For most data types, such as pathways, expression data, or MGD entries, the underlying scores were determined by biological and clinical considerations. For instance, if an

OMIM entry – a clinical diagnosis – has been entered and a match is found, this is rewarded with a high score as this type of data is of great clinical relevance and hence of importance for the user. Matching pathway data, on the other hand, is scored with a lower score as the clinical significance of pathways cannot usually directly be deduced from available data.

Thanks to the availability of genotype data in connection with relevant HPO terms (our ClinVar and in-house data sets), we were able to optimise and test this scoring for HPO entries (see section 7). For all other data types, this was not yet feasible due to a lack of suitable data or users. All our cooperation partners thus far have based their analyses on the HPO as well. However, over time, we hope to receive feedback from users of diverse backgrounds, which would enable us to update MutationDistiller and to optimise the weights for other data sources.

## 8.2.2 Mutation severity

In our scoring system, and in opposition to many other tools such as PhenIX and Exomiser, we do not include the 'gravity' of a variant's predicted effect. MutationDistiller receives its pathogenicity prediction from MutationTaster, which employs a Naïve Bayes classifier to sort the variants into either 'harmless' or 'harmful'. While the classifier also delivers a probability value, this does not mirror how severe an alteration is but only how certain the classifier is with its decision. Instead, the severity of a mutation can be seen as its capacity to cause harm to the gene product – thus, a nonsense variant can be expected to be more harmful than most other variants (which is indicated by MutationTaster's *disease-causing (automatic)* classification). We therefore do not include MutationTaster's probability value into the scoring system, but allow users to investigate a variant in detail by offering a direct hyperlink to MutationTaster's prediction for it. In addition, we allow users to filter variants by severity via our variant classes. Moreover, we plan to improve and increase this functionality as described in section 8.5.6.

## 8.3 Phenotype data variety

### 8.3.1 Detection of new disease genes

The notion that diagnosis of rare diseases can be improved or accelerated through the inclusion of phenotype data is mirrored by the number of phenotype-based analysis tools that have been developed in recent years (see section 1.4.2). Most of these tools are based

on the HPO, as this resource is currently the most widely used in human genetics. Thus, we expect most of our users to be used to and to already rely on HPO-based phenotyping data, and hence focused the development of MutationDistiller on this resource. Indeed, most of our users to date have based their search on the HPO. However, it has to be noted that analyses based on the HPO or resources such as OMIM and Orphanet are only able to detect genes which are already known to be involved in the development of genetic disorders.

However, new ways of analysing rare disease cases are needed, as the diagnostic yield in NGS projects is usually reported to lie between 25-30% [76, 138, 139]. Reasons for this are manifold; the disease-causing mutation might not be covered (sufficiently), or be located outside of the coding sequence. In addition, it might not be recognised as a disease mutation, or be located in a gene that has not yet been discovered to be disease-relevant.

However, as described above, the inclusion of functional data, such as GO, expression or pathway resources, allows us to detect 'new' disease genes. This approach, which is feasible using MutationDistiller, can thus help in elucidating currently unsolved cases.

Therefore, we expect our tool to be of assistance in a range of cases where current means have not been able to identify the causative variant. The benefit of re-analysis of WES or WGS data which has previously not led to success has recently been demonstrated by the Deciphering Developmental Disorders (DDD) team in the UK [124]. MutationDistiller's flexibility lends itself to attempt re-analyses using new resources or with novel insights about the disease, and we hence expect the tool to be of use in many currently unsolved cases.

### 8.3.2 Symptom annotations

A lack of annotated symptoms can be a problem when relying solely on HPO-based analyses. The HPO obtains symptom annotations via OMIM and Orphanet, and depending how quickly and reliably these links are established and updated, there might be quite some lag-time. By offering a wide range of options for entering patient-related data, MutationDistiller can find mutations in genes that are not yet sufficiently annotated:

As an example, one of our collaborating clinicians provided whole exome data from a patient diagnosed with congenital myasthenia suffering from *areflexia* (HP:0001284) and *muscular hypotonia* (HP:0001252). The pathogenic variant in this case had previously been determined to be located in the *SLC5A7* gene. However, when trying to assess this case using the HPO terms described above in MutationDistiller and other software

tools, the causative variant could not be found within the best ranks. Despite being listed in OMIM as a molecular cause for *congenital preysnaptic myasthenic syndrome 20* (*CMS20*, *OMIM:617143*) and *distal hereditary motor neuropathy type VIIA* (*HMN7A*, *OMIM: 158580*), the *SLC5A7* gene was not linked with these symptoms in the HPO, and hence the programs could establish a connection. In MutationDistiller, users can chose one of multiple alternative approaches to overcome this obstacle: By restricting the search to a relevant virtual panel – in this case the congenital myasthenia panel from PanelApp – the clinician was able to identify the causative alteration. Alternative ways to come to the same result would have been to enter the clinical diagnosis via OMIM, or to upload an in-house virtual panel.

## 8.4 Comparison with state-of-the-art tools

To evaluate MutationDistiller’s ability to prioritise relevant variants, we compared its performance to similar state-of-the art tools. As described in section 7.2.3, we decided to limit our comparison to tools that are freely available online without any need to install or log in. However, it has to be noted that we had to fall back to downloaded versions of the tools as using them online would have been too slow.

We compared MutationDistiller’s prioritisation capabilities to three other algorithms, eXtasy, PhenIX and Exomiser. In a recent comparison analysis which did not include MutationDistiller, PhenIX was found to deliver the best results on 21 exomes [140]. This study was conducted without the involvement of any authors of PhenIX.

Using our set of 101 variants obtained from the Charité, we found that MutationDistiller was capable of placing the causative variant within the top 10 in over 80% of the cases, thus out-competing the other tools included in our comparison.

However, in addition to HPO terms and in contrast to many other software options, MutationDistiller offers a wide range of input data. Unfortunately, due to a lack of both testing data and candidate tools, we were not able to compare MutationDistiller quantitatively in this respect and had to limit our comparison to HPO data.

When designing MutationDistiller, we aimed to generate a comprehensive and user-friendly software tool for clinicians and researchers. This becomes obvious when comparing the output and surrounding information of the four tools: MutationDistiller provides a wide range of information in the output page rather than a battery of scores. In addition, the tool displays the final score and its contributing sub-scores to allow users to make an educated decision about their case. Moreover, our program provides comprehensive tutorial and manual pages, aiming at making its usage as easy as possible.

We are convinced that these features facilitate MutationDistiller’s use in the clinic, as suggested by Shyr *et al* [70].

It has to be noted that comparisons of multi-faceted tools such as MutationDistiller are difficult to achieve as many different factors have to be taken into account. A realistic comparison would have to be conducted by researchers who do not have any stakes in any of the tested tools. In this way, one could study how much time a trained physician spends to identify a mutation which they truly believe to be causal. Unfortunately, so far no one volunteered for this time intensive task.

## 8.5 Outlook

MutationDistiller in its current form supports a wide range of input data to determine the most likely disease-relevant alterations for a given case. Nevertheless, as in any project, there is always room for development and improvement. For the future of MutationDistiller, a number of development opportunities remain, which I will discuss in the following sections.

### 8.5.1 Family analyses

In contrast to other means such as linkage analysis, NGS analyses are able to determine the most likely disease-relevant alterations even if there is only the patient’s genotype available. However, the hunt can be made much simpler when using *family data* by adding data sets obtained from (healthy) relatives. The most common approach is the analysis of trios consisting of the patient and her or his parents. In previous comparison studies, this approach has been found to increase diagnostic yield [138, 139, 141]. Moreover, this approach has great advantages when filtering against variant databases such as ExAC, as only the variants occurring in the family have to be taken into account and the issue of variant frequency can be neglected.

Adding the parents’ or siblings’ sequencing information to the analysis allows the exclusion of a large number of potential alterations: In recessive disorders with complete penetrance, all alterations that can be found in a homozygous state in a healthy individual can be safely removed from further investigation. In fully penetrant dominant disorders, even inherited heterozygous alterations can be discarded. Thus, in dominant modes of inheritance, *de novo* mutations can be specifically searched for. MutationDistiller already allows the analysis of trio data in an indirect way: A user can create separate projects and then compare the results to exclude non-relevant data. This approach, however, is rather cumbersome. Thus, we plan to update MutationDistiller to



allow analyses of families in addition to the analysis of singletons.

We have thus far not been able to introduce this function due to a lack of training data, as trio analyses have not yet been introduced in routine care due to the higher cost compared to singleton sequencing. As sequencing costs drop and the awareness of the benefits of trio analyses rise, however, a higher rate of trio analyses is to be expected even in routine settings. This would allow for the development of suitable tools while simultaneously increasing the demand for such software. We hence plan to add this feature to MutationDistiller in the near future.

### 8.5.2 Genome version

Currently, MutationDistiller is based on genome build GRCh37, even though a more recent version, GRCh38, has been available for a number of years. We made this decision due to the fact that secondary data used by MutationDistiller and MutationTaster is only available for GRCh37. While mapping between the two versions (a process called *liftover*) is possible, this is a tedious process which we have not yet seen the need for as our users are currently still relying on GRCh37. Therefore, we decided to stick with GRCh37 for MutationDistiller's first version. However, we are well aware that in the future, the demand for tools compatible with GRCh38 might increase. We are thus planning to add this genome version to MutationDistiller as demand arises.

### 8.5.3 WGS data

In theory, MutationDistiller would be capable of addressing WGS projects already. However, MutationTaster can only handle a small part of WGS projects as it is limited to protein-coding genes (but analyses non-coding variants contained in these). When starting the project, WGS analysis was still prohibitively expensive and therefore not used in routine clinical research, leading to a lack of data sources. We have thus decided to limit MutationDistiller to WES projects and to use MutationTaster as the variant effect predictor. In the course of the project, however, both of these points have changed – thanks to the efforts of the ENCODE consortium and other projects, data on non-coding regions and their regulatory relevance are readily available, while lower sequencing costs have led to an increase in WGS usage [32].

In light of these developments, we are aware that the need for programs like MutationDistiller to cope with WGS data is rising steadily. In our research group, we have since developed RegulationSpotter [142], a tool to analyse WGS projects in order to find alterations located in areas of regulatory relevance. Moreover, RegulationSpotter is able

to take HPO terms as input to identify variants located in regulatory regions of genes that are connected with a given phenotype. However, the prediction quality for variants located outside of transcript regions is currently not sufficient for direct incorporation into clinical tools: with a mean of 3 million variants per WGS experiment, the tools simply drown in false positives. With an increased use of WGS, this might change as more training data becomes available, i.e. experimentally confirmed disease mutations outside of protein-coding genes. We therefore decided to keep the two tools separate for now. We are, however, considering to merge MutationDistiller and RegulationSpotter in the future to develop software that is capable of both analysing WGS data and coping with the variety of input options currently offered by MutationDistiller.

### 8.5.4 Mitochondrial DNA

In contrast to many other tools, MutationDistiller can detect mutations located in mitochondrial DNA (mtDNA). However, as the software is based on diploidy, the program cannot take heteroplasmy into account. As described in section 1.3.3, in many mitochondrial disorders, the degree of heteroplasmy plays a role and only individuals carrying a high amount of mutated mtDNA will be affected by a disease.

To achieve this, we would have to read the degree of heteroplasmy from the VCF and incorporate mitochondrial databases into MutationDistiller. Moreover, we would have to alter our database structure accordingly and change the filters for external databases and trios. While these changes are feasible, they require substantial changes to our database, the integrated data, and how we call the data within MutationDistiller. We are therefore planning to achieve this in a second version of the program.

### 8.5.5 gnomAD

We have currently implemented the variant database ExAC into MutationDistiller. As described in section 2.2.3.2, this source contains human exome sequencing data from over 60,000 individuals. During the course of the development of MutationDistiller, gnomAD [43], a genome-wide version, has been established. GnomAD contains over 125,000 exome sequences and over 15,000 whole-genome sequences. These data were obtained from a range of studies, both on diseases and on healthy populations. Previously, inclusion criteria into gnomAD were unclear and did not allow to distinguish easily between sequences from healthy individuals or patients suffering from genetic disorders. This has prevented us from incorporating gnomAD data into MutationDistiller. However, this issue was solve recently as gnomAD now separates the data into control and patient populations. Therefore, we are currently working on adding gnomAD to MutationDistiller.

## 8.5.6 Classification bins

MutationDistiller currently outputs an ordered list of genes and variants, sorted by how well they match the user-defined criteria. However, depending on these criteria, a number of genes and variants may receive the same score and are hence allocated to the same rank. This is the case especially when users do not provide a wide range of criteria or apply only few restrictions. For reasons of simplicity and to keep run-times fast, we are currently displaying those variants in no particular order. In the future, we wish to completely overhaul the ranking system and instead sort variants into bins indicating how relevant they are for the given case. In this system, we would be able to not only take the phenotypic relevance of a variant into account but to also sort alterations by their predicted effect on the protein (e.g. missense alteration vs. NMD). Rather than ranking the candidate variants, we would provide several bins of alterations that are deleterious, while at the same time matching the phenotype of interest to varying degrees, expressed by variant flags. Users could then toggle several switches depending on their focus to show the predicted phenotype, effect, location, or gene function. For instance, one flag would be whether the gene matches the phenotype description (in three stages, e.g. green/yellow/red), another flag would denote the predicted effect (ClinVar/NMD/splice site etc.), and another flag would be reserved for the mode of inheritance. Moreover, additional flags could be added in later stages of the program, thus allowing for great flexibility.

We have thus far decided not to implement this system yet for two main reasons. First, all other variant prioritisation tools use ranked list in one way or another – hence users are well acquainted with this approach. Rather than pushing users to learn how to use a new program and a new sorting system at the same time, we decided to take one step at a time. Moreover, we were still suffering from a lack of training data in order to generate reliable thresholds for the binning system. However, with the increased usage of NGS data in routine clinical settings, we are convinced that this problem is just a matter of time. We are thus optimistic to be able to update MutationDistiller to a binning system in the near future.

## 8.5.7 Data management

### 8.5.7.1 User data sustainability

In the development process of MutationDistiller, we have opted against an automated process for removing data. Thus, we are currently manually running a script to delete

user data at regular intervals. Now that the testing and development phase is over, we are planning to automate this process. In addition to our deletion of user data in regular intervals, and to provide further data security, users can currently simply delete their projects by project ID and security code.

### 8.5.7.2 Data sources

We have not yet implemented an automated way of updating the data sources integrated into MutationDistiller. Instead, we update data sets manually when we become aware of relevant changes, which can be cumbersome: With GeneDistiller, MutationDistiller's parent tool that has been running for over 10 years now, we have experienced updates to be complex and time consuming. However, an automated update protocol would make our lives simpler while at the same time ensuring more up-to-date data. This would be especially beneficial for data sources that we always want to keep as up-to-date as possible (such as OMIM or HPO).

However, automated updates are not practical, necessary or feasible for all the different data types MutationDistiller is using, since data structures can change and services be discontinued. We are therefore planning to automate updates for selected data sources that provide easy access to their data and do not alter data structure from one update to the next. Current candidates for automatic updates are PanelApp, the HPO and WikiPathways, but this list can be changed and broadened in the future.

## 8.6 Clinical use

Thanks to technical advances, the field of genomics has been catapulted into the digital age. NGS methods allow for easy, fast and cheap sequencing, thus enabling work on projects and cases that could not be handled before. However, the vast amounts of data generated in NGS projects pose major obstacles and thus prevent clinicians, researchers and genetic counsellors from attempting such endeavours [70, 143, 144]. With MutationDistiller, we have attempted to respond to the need for dedicated expert software that is easy to use, provides a convenient user interface, and allows the analysis of large data sets without having to obtain a bioinformatics degree first. MutationDistiller has been designed as a tool to support rare disease research as well as clinical assessments. However, it is not and cannot be a diagnostic tool or a medical device, since to achieve this status strict regulations have to be followed, which are beyond the scope of this research group.

Nevertheless, the tool has already entered the clinic: To date, over 14,000 individual cases

have been uploaded to our database by clinicians and researchers. In recent months, MutationDistiller has seen over 1,000 cases every 30 days. The tool has been used in projects from all around the world, and we expect this to increase still as clinicians more routinely sequence their patients. We therefore hope that the work presented in this thesis can bring some contributions to the field of genomics, and be of benefit for the numerous patients suffering from so-called rare disorders.

## Abbreviations

<b>A</b>	adenine
<b>AJAX</b>	Asynchronous JavaScript and XML
<b>bp</b>	base pair
<b>C</b>	cytosine
<b>CGI</b>	common gateway interface
<b>ChIP</b>	chromatin immunoprecipitation
<b>CNV</b>	copy number variant
<b>ddNTP</b>	dideoxynucleotide
<b>DNA</b>	deoxyribonucleic acid
<b>DBMS</b>	database management system
<b>DW</b>	data warehouse
<b>CADD</b>	Combined Annotation Dependent Depletion
<b>cDNA</b>	complementary DNA
<b>ENCODE</b>	Encyclopedia of DNA elements
<b>ExAC</b>	Exome Aggregation Consortium
<b>G</b>	guanine
<b>GMT</b>	Gene Matrix Transposed
<b>GOF</b>	gain of function
<b>GO</b>	Gene Ontology
<b>GRC</b>	Genome Reference Consortium
<b>GWAS</b>	genome wide association study
<b>HGMD</b>	Human Gene Mutation Database
<b>hPDI</b>	Human Protein-DNA Interactome
<b>HPO</b>	Human Phenotype Ontology
<b>HT</b>	high-throughput
<b>HTML</b>	HyperText Markup Language
<b>InDel</b>	insertion/deletion
<b>KEGG</b>	Kyoto Encyclopedia of Genes and Genomes
<b>LOF</b>	loss of function
<b>LD</b>	linkage disequilibrium
<b>MoI</b>	mode of inheritance
<b>mtDNA</b>	mitochondrial DNA
<b>mut</b>	mutation
<b>RNA</b>	ribonucleic acid
<b>mRNA</b>	messenger RNA
<b>NGS</b>	next generation sequencing

---

<b>OMIM</b>	Online Mendelian Inheritance in Man
<b>ORF</b>	open reading frame
<b>PKU</b>	phenylketonuria
<b>QE</b>	query engine
<b>rSNP</b>	regulatory single nucleotide polymorphism
<b>SQL</b>	Structured Query Language
<b>SNP</b>	single nucleotide polymorphism
<b>SNV</b>	single nucleotide variant
<b>T</b>	thymine
<b>TAD</b>	transactivating domain
<b>TF</b>	transcription factor
<b>TSS</b>	transcription start site
<b>TSV</b>	tab separated values
<b>U</b>	uracil
<b>UCSC</b>	University of California, Santa Cruz
<b>UI</b>	user interface
<b>var</b>	variant
<b>VCF</b>	Variant Call Format
<b>WES</b>	whole exome sequencing
<b>WGS</b>	whole genome sequencing
<b>wt</b>	wildtype
<b>XML</b>	Extensible Markup Language

## A Appendix – HPO optimisation weights

The following table displays summaries for all weight combinations ordered by lowest number of unsolved cases. The abbreviations indicate as follows: `anc_weight`: weight assigned to ancestor matches. `desc_weight`: weight assigned to descendant matches. `n`: number of cases in total. `first`: number of indicated genes ranked on rank 1. `one_five`: number of indicated genes ranked on ranks 1-5, respectively. `gr_ten`: number of indicated genes ranked higher than rank 10. `mean_rank`: mean rank allocated to the indicated genes for the given combination. `not_found`: number of cases ranked higher than 100. Combination 210, marked with an asterisk, is currently implemented into MutationDistiller.

combination ID	direct_weight	anc_weight	desc_weight	first	one_five	gr_ten	mean_rank	not_found
145	1.00	2.00	0.50	7	74	37	9.55	42
31	0.20	0.50	0.10	6	65	38	10.10	42
38	0.20	1.00	0.10	3	49	48	11.46	42
45	0.20	2.00	0.10	3	46	52	11.92	42
*210	5.00	0.05	2.00	70	108	26	5.82	43
224	5.00	0.20	2.00	70	109	26	5.86	43
217	5.00	0.10	2.00	70	108	26	5.87	43
160	2.00	0.05	1.00	67	108	26	5.96	43
60	0.50	0.05	0.20	70	107	24	5.99	43
231	5.00	0.50	2.00	70	107	24	5.99	43
110	1.00	0.05	0.50	67	108	26	6.02	43
167	2.00	0.10	1.00	67	108	26	6.02	43
117	1.00	0.10	0.50	67	107	26	6.12	43
174	2.00	0.20	1.00	67	107	26	6.12	43
67	0.50	0.10	0.20	69	106	24	6.17	43
238	5.00	1.00	2.00	69	106	24	6.17	43
124	1.00	0.20	0.50	66	106	25	6.25	43
111	1.00	0.05	1.00	56	104	27	6.28	43
168	2.00	0.10	2.00	56	104	27	6.28	43
10	0.20	0.05	0.10	66	105	25	6.28	43
181	2.00	0.50	1.00	66	105	25	6.28	43
161	2.00	0.05	2.00	56	103	28	6.30	43
61	0.50	0.05	0.50	56	105	27	6.37	43
118	1.00	0.10	1.00	56	105	27	6.37	43
175	2.00	0.20	2.00	56	105	27	6.37	43
11	0.20	0.05	0.20	56	104	27	6.52	43
68	0.50	0.10	0.50	56	103	28	6.52	43
125	1.00	0.20	1.00	56	103	28	6.52	43
182	2.00	0.50	2.00	56	104	27	6.52	43
17	0.20	0.10	0.10	61	99	25	6.64	43
131	1.00	0.50	0.50	61	99	25	6.64	43
188	2.00	1.00	1.00	61	99	25	6.64	43



combination ID	direct_weight	anc_weight	desc_weight	first	one_five	gr_ten	mean_rank	not_found
75	0.50	0.20	0.50	55	101	26	6.66	43
112	1.00	0.05	2.00	42	102	27	6.83	43
18	0.20	0.10	0.20	55	98	27	6.84	43
132	1.00	0.50	1.00	55	98	27	6.84	43
189	2.00	1.00	2.00	55	98	27	6.84	43
62	0.50	0.05	1.00	42	100	28	6.87	43
119	1.00	0.10	2.00	42	100	28	6.87	43
24	0.20	0.20	0.10	39	90	27	7.43	43
138	1.00	1.00	0.50	39	90	27	7.43	43
195	2.00	2.00	1.00	39	90	27	7.43	43
25	0.20	0.20	0.20	37	90	28	7.60	43
82	0.50	0.50	0.50	37	90	28	7.60	43
139	1.00	1.00	1.00	37	90	28	7.60	43
196	2.00	2.00	2.00	37	90	28	7.60	43
63	0.50	0.05	2.00	37	88	28	7.63	43
89	0.50	1.00	0.50	8	69	32	8.99	43
146	1.00	2.00	1.00	8	69	32	8.99	43
88	0.50	1.00	0.20	8	75	36	9.12	43
90	0.50	1.00	1.00	11	67	35	9.15	43
147	1.00	2.00	2.00	11	67	35	9.15	43
87	0.50	1.00	0.10	8	75	38	9.22	43
144	1.00	2.00	0.20	8	75	38	9.22	43
142	1.00	2.00	0.05	8	74	37	9.28	43
86	0.50	1.00	0.05	8	74	38	9.28	43
143	1.00	2.00	0.10	8	74	38	9.28	43
32	0.20	0.50	0.20	6	60	34	9.59	43
91	0.50	1.00	2.00	14	62	39	9.60	43
33	0.20	0.50	0.50	9	57	38	9.68	43
30	0.20	0.50	0.05	6	66	39	9.77	43
34	0.20	0.50	1.00	13	53	43	10.10	43
35	0.20	0.50	2.00	15	54	42	10.23	43
93	0.50	2.00	0.05	3	51	45	10.72	43
94	0.50	2.00	0.10	3	51	46	10.73	43
97	0.50	2.00	1.00	3	46	47	10.74	43
95	0.50	2.00	0.20	3	51	44	10.75	43
96	0.50	2.00	0.50	3	47	44	10.77	43
98	0.50	2.00	2.00	8	45	52	10.85	43
37	0.20	1.00	0.05	3	50	47	10.99	43
39	0.20	1.00	0.20	3	45	47	11.05	43
40	0.20	1.00	0.50	3	44	54	11.18	43
41	0.20	1.00	1.00	7	42	55	11.20	43
42	0.20	1.00	2.00	12	42	55	11.21	43
44	0.20	2.00	0.05	3	47	50	11.46	43

combination ID	direct_weight	anc_weight	desc_weight	first	one_five	gr_ten	mean_rank	not_found
46	0.20	2.00	0.20	3	42	52	11.50	43
49	0.20	2.00	2.00	6	38	60	11.70	43
47	0.20	2.00	0.50	3	40	57	11.74	43
48	0.20	2.00	1.00	3	41	61	12.06	43
159	2.00	0.05	0.50	69	110	23	5.45	44
209	5.00	0.05	1.00	69	108	23	5.45	44
216	5.00	0.10	1.00	69	108	23	5.46	44
208	5.00	0.05	0.50	65	106	24	5.50	44
223	5.00	0.20	1.00	69	107	23	5.51	44
109	1.00	0.05	0.20	69	107	23	5.51	44
166	2.00	0.10	0.50	69	108	23	5.52	44
207	5.00	0.05	0.20	63	106	24	5.54	44
158	2.00	0.05	0.20	65	106	24	5.55	44
215	5.00	0.10	0.50	65	106	24	5.55	44
222	5.00	0.20	0.50	64	106	25	5.58	44
173	2.00	0.20	0.50	68	108	24	5.58	44
214	5.00	0.10	0.20	62	106	25	5.60	44
157	2.00	0.05	0.10	63	106	25	5.61	44
108	1.00	0.05	0.10	64	106	25	5.61	44
165	2.00	0.10	0.20	64	106	25	5.61	44
107	1.00	0.05	0.05	62	104	24	5.62	44
164	2.00	0.10	0.10	62	104	24	5.62	44
206	5.00	0.05	0.10	62	106	25	5.62	44
221	5.00	0.20	0.20	61	104	24	5.65	44
156	2.00	0.05	0.05	61	104	24	5.65	44
220	5.00	0.20	0.10	61	102	24	5.65	44
59	0.50	0.05	0.10	66	107	25	5.65	44
116	1.00	0.10	0.20	66	107	25	5.65	44
230	5.00	0.50	1.00	66	107	25	5.65	44
213	5.00	0.10	0.10	61	104	24	5.67	44
163	2.00	0.10	0.05	61	102	24	5.68	44
205	5.00	0.05	0.05	61	104	24	5.69	44
58	0.50	0.05	0.05	63	103	25	5.69	44
115	1.00	0.10	0.10	63	103	25	5.69	44
172	2.00	0.20	0.20	63	103	25	5.69	44
229	5.00	0.50	0.50	63	103	25	5.69	44
212	5.00	0.10	0.05	61	103	24	5.70	44
219	5.00	0.20	0.05	61	102	23	5.71	44
114	1.00	0.10	0.05	62	102	25	5.73	44
171	2.00	0.20	0.10	62	102	25	5.73	44
228	5.00	0.50	0.20	62	102	24	5.74	44
227	5.00	0.50	0.10	62	102	24	5.76	44
170	2.00	0.20	0.05	62	102	24	5.77	44

combination ID	direct_weight	anc_weight	desc_weight	first	one_five	gr_ten	mean_rank	not_found
65	0.50	0.10	0.05	64	102	24	5.78	44
122	1.00	0.20	0.10	64	102	24	5.78	44
236	5.00	1.00	0.50	64	102	24	5.78	44
235	5.00	1.00	0.20	64	103	24	5.81	44
66	0.50	0.10	0.10	65	103	23	5.81	44
123	1.00	0.20	0.20	65	103	23	5.81	44
178	2.00	0.50	0.10	65	103	22	5.81	44
237	5.00	1.00	1.00	65	103	23	5.81	44
226	5.00	0.50	0.05	61	102	24	5.81	44
121	1.00	0.20	0.05	64	102	24	5.82	44
179	2.00	0.50	0.20	65	103	22	5.82	44
9	0.20	0.05	0.05	66	104	23	5.84	44
180	2.00	0.50	0.50	66	104	23	5.84	44
177	2.00	0.50	0.05	65	103	22	5.85	44
234	5.00	1.00	0.10	64	103	24	5.86	44
233	5.00	1.00	0.05	64	103	25	5.88	44
72	0.50	0.20	0.05	63	99	22	6.03	44
243	5.00	2.00	0.50	63	99	22	6.03	44
74	0.50	0.20	0.20	64	100	23	6.04	44
245	5.00	2.00	2.00	64	100	23	6.04	44
73	0.50	0.20	0.10	63	99	23	6.05	44
244	5.00	2.00	1.00	63	99	23	6.05	44
242	5.00	2.00	0.20	63	99	24	6.10	44
16	0.20	0.10	0.05	61	99	23	6.12	44
187	2.00	1.00	0.50	61	99	23	6.12	44
241	5.00	2.00	0.10	63	99	24	6.15	44
240	5.00	2.00	0.05	63	99	24	6.15	44
130	1.00	0.50	0.20	61	99	23	6.18	44
129	1.00	0.50	0.10	61	100	25	6.23	44
186	2.00	1.00	0.20	61	100	25	6.23	44
128	1.00	0.50	0.05	61	100	25	6.24	44
185	2.00	1.00	0.10	61	100	25	6.24	44
184	2.00	1.00	0.05	61	100	26	6.29	44
69	0.50	0.10	1.00	42	99	27	6.40	44
126	1.00	0.20	2.00	42	99	27	6.40	44
76	0.50	0.20	1.00	41	98	26	6.53	44
133	1.00	0.50	2.00	41	96	25	6.58	44
12	0.20	0.05	0.50	41	96	27	6.63	44
19	0.20	0.10	0.50	41	93	26	6.89	44
81	0.50	0.50	0.20	39	93	27	7.03	44
23	0.20	0.20	0.05	39	94	28	7.06	44
194	2.00	2.00	0.50	39	94	28	7.06	44
80	0.50	0.50	0.10	39	92	28	7.07	44

combination ID	direct_weight	anc_weight	desc_weight	first	one_five	gr_ten	mean_rank	not_found
137	1.00	1.00	0.20	39	92	28	7.07	44
79	0.50	0.50	0.05	38	92	30	7.08	44
136	1.00	1.00	0.10	38	92	30	7.08	44
193	2.00	2.00	0.20	38	92	30	7.08	44
135	1.00	1.00	0.05	36	92	30	7.15	44
192	2.00	2.00	0.10	36	92	30	7.15	44
70	0.50	0.10	2.00	37	87	27	7.17	44
77	0.50	0.20	2.00	38	87	26	7.22	44
191	2.00	2.00	0.05	35	92	30	7.23	44
83	0.50	0.50	1.00	31	82	28	7.30	44
140	1.00	1.00	2.00	31	82	28	7.30	44
13	0.20	0.05	1.00	35	85	29	7.35	44
20	0.20	0.10	1.00	36	83	26	7.48	44
26	0.20	0.20	0.50	30	76	28	7.53	44
14	0.20	0.05	2.00	34	80	31	7.60	44
21	0.20	0.10	2.00	34	78	29	7.76	44
84	0.50	0.50	2.00	29	72	30	7.94	44
27	0.20	0.20	1.00	28	70	30	8.12	44
28	0.20	0.20	2.00	26	69	32	8.33	44
85	0.50	1.00	0.00	8	74	29	8.15	48
141	1.00	2.00	0.00	8	74	29	8.15	48
29	0.20	0.50	0.00	6	66	31	8.64	48
92	0.50	2.00	0.00	3	52	36	9.47	48
36	0.20	1.00	0.00	3	51	37	9.70	48
43	0.20	2.00	0.00	3	48	40	10.00	48
211	5.00	0.10	0.00	61	104	14	4.49	49
204	5.00	0.05	0.00	61	104	15	4.50	49
218	5.00	0.20	0.00	61	103	14	4.50	49
106	1.00	0.05	0.00	61	103	14	4.50	49
155	2.00	0.05	0.00	61	103	14	4.50	49
162	2.00	0.10	0.00	61	103	14	4.50	49
57	0.50	0.05	0.00	61	103	15	4.58	49
113	1.00	0.10	0.00	61	103	15	4.58	49
169	2.00	0.20	0.00	61	103	15	4.58	49
225	5.00	0.50	0.00	61	103	15	4.58	49
64	0.50	0.10	0.00	64	104	17	4.66	49
120	1.00	0.20	0.00	64	104	17	4.66	49
232	5.00	1.00	0.00	64	104	17	4.66	49
8	0.20	0.05	0.00	65	104	16	4.66	49
176	2.00	0.50	0.00	65	104	16	4.66	49
71	0.50	0.20	0.00	63	99	18	4.93	49
239	5.00	2.00	0.00	63	99	18	4.93	49
15	0.20	0.10	0.00	61	100	20	5.09	49

---

combination ID	direct_weight	anc_weight	desc_weight	first	one_five	gr_ten	mean_rank	not_found
127	1.00	0.50	0.00	61	100	20	5.09	49
183	2.00	1.00	0.00	61	100	20	5.09	49
22	0.20	0.20	0.00	35	92	23	6.04	49
78	0.50	0.50	0.00	35	92	23	6.04	49
134	1.00	1.00	0.00	35	92	23	6.04	49
190	2.00	2.00	0.00	35	92	23	6.04	49
53	0.50	0.00	0.20	74	105	19	4.78	53
203	5.00	0.00	2.00	74	105	19	4.78	53
3	0.20	0.00	0.10	71	106	20	4.93	53
103	1.00	0.00	0.50	71	106	20	4.93	53
153	2.00	0.00	1.00	71	106	20	4.93	53
4	0.20	0.00	0.20	58	103	23	5.36	53
54	0.50	0.00	0.50	58	103	23	5.36	53
104	1.00	0.00	1.00	58	103	23	5.36	53
154	2.00	0.00	2.00	58	103	23	5.36	53
55	0.50	0.00	1.00	43	100	21	5.91	53
105	1.00	0.00	2.00	43	100	21	5.91	53
5	0.20	0.00	0.50	42	96	23	6.17	53
56	0.50	0.00	2.00	38	92	23	6.61	53
6	0.20	0.00	1.00	36	90	24	6.81	53
7	0.20	0.00	2.00	35	86	25	7.03	53
52	0.50	0.00	0.10	73	105	16	4.29	54
102	1.00	0.00	0.20	73	105	16	4.29	54
202	5.00	0.00	1.00	73	105	16	4.29	54
100	1.00	0.00	0.05	68	104	16	4.31	54
150	2.00	0.00	0.10	68	104	16	4.31	54
51	0.50	0.00	0.05	70	104	16	4.31	54
101	1.00	0.00	0.10	70	104	16	4.31	54
151	2.00	0.00	0.20	70	104	16	4.31	54
201	5.00	0.00	0.50	70	104	16	4.31	54
200	5.00	0.00	0.20	67	104	16	4.31	54
2	0.20	0.00	0.05	73	106	16	4.31	54
152	2.00	0.00	0.50	73	106	16	4.31	54
199	5.00	0.00	0.10	67	104	16	4.33	54
149	2.00	0.00	0.05	67	104	16	4.34	54
198	5.00	0.00	0.05	67	103	16	4.35	54
1	0.20	0.00	0.00	65	104	6	3.02	63
50	0.50	0.00	0.00	65	104	6	3.02	63
99	1.00	0.00	0.00	65	104	6	3.02	63
148	2.00	0.00	0.00	65	104	6	3.02	63
197	5.00	0.00	0.00	65	104	6	3.02	63

---

## B Appendix – Expression tissue groups

This table provides a summary of the the groups and subgroups developed to structure expression data downloaded from ExpressionAtlas.

datasource	class	group	sub-tissues
E-MTAB-4344	organs	brain	brain
E-MTAB-4344	organs	kidney	kidney
E-MTAB-4344	organs	liver	liver
E-MTAB-4344	organs	lung	lung
E-MTAB-4344	organs	heart	heart
E-MTAB-4344	organs	gastrointestinal tract	sigmoid colon
E-MTAB-4344	organs	gastrointestinal tract	small intestine
E-MTAB-4344	organs	reproductive organs	ovary
E-MTAB-4344	organs	reproductive organs	testis
E-MTAB-4344	tissues	adipose tissue	adipose tissue
E-MTAB-4344	systems	nervous system	brain
E-MTAB-4344	systems	circulatory/respiratory system	heart
E-MTAB-4344	systems	circulatory/respiratory system	lung
E-MTAB-4344	systems	immune system	spleen
E-MTAB-4344	systems	reproductive system	ovary
E-MTAB-4344	systems	reproductive system	testis
E-MTAB-4344	systems	food intake/digestion	sigmoid colon
E-MTAB-4344	systems	food intake/digestion	small intestine
E-MTAB-4344	systems	urinary system	kidney
E-MTAB-4344	systems	endocrine system	adrenal gland
E-MTAB-4344	systems	endocrine system	pancreas
E-MTAB-3358	organs	brain	amygdala
E-MTAB-3358	organs	brain	brain
E-MTAB-3358	organs	brain	caudate nucleus
E-MTAB-3358	organs	brain	cerebellum
E-MTAB-3358	organs	brain	cerebral meninges
E-MTAB-3358	organs	brain	diencephalon
E-MTAB-3358	organs	brain	dura mater
E-MTAB-3358	organs	brain	globus pallidus
E-MTAB-3358	organs	brain	hippocampus
E-MTAB-3358	organs	brain	locus coeruleus
E-MTAB-3358	organs	brain	medulla oblongata
E-MTAB-3358	organs	brain	middle frontal gyrus
E-MTAB-3358	organs	brain	middle temporal gyrus
E-MTAB-3358	organs	brain	occipital cortex
E-MTAB-3358	organs	brain	occipital lobe
E-MTAB-3358	organs	brain	parietal lobe
E-MTAB-3358	organs	brain	putamen

datasource	class	group	sub-tissues
E-MTAB-3358	organs	brain	substantia nigra
E-MTAB-3358	organs	brain	thalamus
E-MTAB-3358	organs	gastrointestinal tract	appendix
E-MTAB-3358	organs	gastrointestinal tract	colon
E-MTAB-3358	organs	heart	artery
E-MTAB-3358	organs	heart	heart
E-MTAB-3358	organs	heart	left atrium
E-MTAB-3358	organs	heart	left ventricle
E-MTAB-3358	organs	heart	mitral valve
E-MTAB-3358	organs	heart	pulmonary valve
E-MTAB-3358	organs	heart	tricuspid valve
E-MTAB-3358	organs	kidney	kidney
E-MTAB-3358	organs	lung	lung
E-MTAB-3358	organs	reproductive organs	cervix
E-MTAB-3358	organs	reproductive organs	epididymis
E-MTAB-3358	organs	reproductive organs	ovary
E-MTAB-3358	organs	reproductive organs	penis
E-MTAB-3358	organs	reproductive organs	placenta
E-MTAB-3358	organs	reproductive organs	prostate
E-MTAB-3358	organs	reproductive organs	seminal vesicle
E-MTAB-3358	organs	reproductive organs	testis
E-MTAB-3358	organs	reproductive organs	uterus
E-MTAB-3358	organs	reproductive organs	vagina
E-MTAB-3358	organs	reproductive organs	vas deferens
E-MTAB-3358	organs	skin	skin
E-MTAB-3358	organs	gallbladder	gallbladder
E-MTAB-3358	organs	olfactory apparatus	olfactory apparatus
E-MTAB-3358	systems	circulatory/respiratory system	artery
E-MTAB-3358	systems	circulatory/respiratory system	heart
E-MTAB-3358	systems	circulatory/respiratory system	left atrium
E-MTAB-3358	systems	circulatory/respiratory system	left ventricle
E-MTAB-3358	systems	circulatory/respiratory system	lung
E-MTAB-3358	systems	circulatory/respiratory system	mitral valve
E-MTAB-3358	systems	circulatory/respiratory system	pulmonary valve
E-MTAB-3358	systems	circulatory/respiratory system	tricuspid valve
E-MTAB-3358	systems	endocrine system	pancreas
E-MTAB-3358	systems	endocrine system	pineal gland
E-MTAB-3358	systems	endocrine system	pituitary gland
E-MTAB-3358	systems	food intake/digestion	appendix
E-MTAB-3358	systems	food intake/digestion	colon
E-MTAB-3358	systems	food intake/digestion	parotid gland
E-MTAB-3358	systems	food intake/digestion	submandibular gland
E-MTAB-3358	systems	food intake/digestion	tongue

datasource	class	group	sub-tissues
E-MTAB-3358	systems	immune system	lymph node
E-MTAB-3358	systems	immune system	spleen
E-MTAB-3358	systems	nervous system	amygdala
E-MTAB-3358	systems	nervous system	brain
E-MTAB-3358	systems	nervous system	caudate nucleus
E-MTAB-3358	systems	nervous system	cerebellum
E-MTAB-3358	systems	nervous system	cerebral meninges
E-MTAB-3358	systems	nervous system	diencephalon
E-MTAB-3358	systems	nervous system	dura mater
E-MTAB-3358	systems	nervous system	globus pallidus
E-MTAB-3358	systems	nervous system	hippocampus
E-MTAB-3358	systems	nervous system	locus coeruleus
E-MTAB-3358	systems	nervous system	medulla oblongata
E-MTAB-3358	systems	nervous system	middle frontal gyrus
E-MTAB-3358	systems	nervous system	middle temporal gyrus
E-MTAB-3358	systems	nervous system	occipital cortex
E-MTAB-3358	systems	nervous system	occipital lobe
E-MTAB-3358	systems	nervous system	parietal lobe
E-MTAB-3358	systems	nervous system	putamen
E-MTAB-3358	systems	nervous system	spinal cord
E-MTAB-3358	systems	nervous system	substantia nigra
E-MTAB-3358	systems	nervous system	thalamus
E-MTAB-3358	systems	neuromuscular	smooth muscle
E-MTAB-3358	systems	reproductive system	cervix
E-MTAB-3358	systems	reproductive system	epididymis
E-MTAB-3358	systems	reproductive system	ovary
E-MTAB-3358	systems	reproductive system	penis
E-MTAB-3358	systems	reproductive system	placenta
E-MTAB-3358	systems	reproductive system	prostate
E-MTAB-3358	systems	reproductive system	seminal vesicle
E-MTAB-3358	systems	reproductive system	testis
E-MTAB-3358	systems	reproductive system	uterus
E-MTAB-3358	systems	reproductive system	vagina
E-MTAB-3358	systems	reproductive system	vas deferens
E-MTAB-3358	systems	skin	skin
E-MTAB-3358	systems	urinary system	kidney
E-MTAB-3358	tissues	bone marrow	bone marrow
E-MTAB-3358	tissues	mammary tissue	breast
E-MTAB-3358	tissues	skin	skin
E-MTAB-3358	tissues	smooth muscle	smooth muscle
E-MTAB-5214	organs	brain	amygdala
E-MTAB-5214	organs	brain	anterior cingulate cortex (BA24)
E-MTAB-5214	organs	brain	caudate (basal ganglia)



datasource	class	group	sub-tissues
E-MTAB-5214	organs	brain	cerebellar hemisphere
E-MTAB-5214	organs	brain	cerebellum
E-MTAB-5214	organs	brain	cerebral cortex
E-MTAB-5214	organs	brain	frontal cortex
E-MTAB-5214	organs	brain	hippocampus
E-MTAB-5214	organs	brain	hypothalamus
E-MTAB-5214	organs	brain	nucleus accumbens (basal ganglia)
E-MTAB-5214	organs	brain	putamen (basal ganglia)
E-MTAB-5214	organs	brain	substantia nigra
E-MTAB-5214	organs	gastrointestinal tract	esophagus muscularis mucosa
E-MTAB-5214	organs	gastrointestinal tract	gastroesophageal junction
E-MTAB-5214	organs	gastrointestinal tract	mucosa of esophagus
E-MTAB-5214	organs	gastrointestinal tract	sigmoid colon
E-MTAB-5214	organs	gastrointestinal tract	stomach
E-MTAB-5214	organs	gastrointestinal tract	terminal ileum of small intestine
E-MTAB-5214	organs	gastrointestinal tract	transverse colon
E-MTAB-5214	organs	heart	aorta
E-MTAB-5214	organs	heart	coronary artery
E-MTAB-5214	organs	heart	left ventricle
E-MTAB-5214	organs	kidney	cortex of kidney
E-MTAB-5214	organs	liver	liver
E-MTAB-5214	organs	lung	lung
E-MTAB-5214	organs	reproductive organs	cervix
E-MTAB-5214	organs	reproductive organs	fallopian tube
E-MTAB-5214	organs	reproductive organs	ovary
E-MTAB-5214	organs	reproductive organs	prostate
E-MTAB-5214	organs	reproductive organs	testis
E-MTAB-5214	organs	reproductive organs	uterus
E-MTAB-5214	organs	reproductive organs	vagina
E-MTAB-5214	organs	skin	skin of lower leg
E-MTAB-5214	organs	skin	skin of suprapubic region
E-MTAB-5214	systems	circulatory/respiratory system	aorta
E-MTAB-5214	systems	circulatory/respiratory system	atrial appendage of heart
E-MTAB-5214	systems	circulatory/respiratory system	coronary artery
E-MTAB-5214	systems	circulatory/respiratory system	left ventricle
E-MTAB-5214	systems	circulatory/respiratory system	lung
E-MTAB-5214	systems	circulatory/respiratory system	tibial artery
E-MTAB-5214	systems	circulatory/respiratory system	whole blood
E-MTAB-5214	systems	endocrine system	adrenal gland
E-MTAB-5214	systems	endocrine system	pancreas
E-MTAB-5214	systems	endocrine system	pituitary gland
E-MTAB-5214	systems	endocrine system	thyroid
E-MTAB-5214	systems	food intake/digestion	esophagus muscularis mucosa

datasource	class	group	sub-tissues
E-MTAB-5214	systems	food intake/digestion	gastroesophageal junction
E-MTAB-5214	systems	food intake/digestion	minor salivary gland
E-MTAB-5214	systems	food intake/digestion	mucosa of esophagus
E-MTAB-5214	systems	food intake/digestion	sigmoid colon
E-MTAB-5214	systems	food intake/digestion	stomach
E-MTAB-5214	systems	food intake/digestion	terminal ileum of small intestine
E-MTAB-5214	systems	food intake/digestion	transverse colon
E-MTAB-5214	systems	immune system	spleen
E-MTAB-5214	systems	nervous system	amygdala
E-MTAB-5214	systems	nervous system	anterior cingulate cortex (BA24)
E-MTAB-5214	systems	nervous system	caudate (basal ganglia)
E-MTAB-5214	systems	nervous system	cerebellar hemisphere
E-MTAB-5214	systems	nervous system	cerebellum
E-MTAB-5214	systems	nervous system	cerebral cortex
E-MTAB-5214	systems	nervous system	frontal cortex
E-MTAB-5214	systems	nervous system	hippocampus
E-MTAB-5214	systems	nervous system	hypothalamus
E-MTAB-5214	systems	nervous system	nucleus accumbens (basal ganglia)
E-MTAB-5214	systems	nervous system	putamen (basal ganglia)
E-MTAB-5214	systems	nervous system	spinal cord (cervical c-1)
E-MTAB-5214	systems	nervous system	substantia nigra
E-MTAB-5214	systems	nervous system	tibial nerve
E-MTAB-5214	systems	neuromuscular	skeletal muscle
E-MTAB-5214	systems	reproductive system	cervix
E-MTAB-5214	systems	reproductive system	fallopian tube
E-MTAB-5214	systems	reproductive system	ovary
E-MTAB-5214	systems	reproductive system	prostate
E-MTAB-5214	systems	reproductive system	testis
E-MTAB-5214	systems	reproductive system	uterus
E-MTAB-5214	systems	reproductive system	vagina
E-MTAB-5214	systems	skin	skin of lower leg
E-MTAB-5214	systems	skin	skin of suprapubic region
E-MTAB-5214	systems	urinary system	bladder
E-MTAB-5214	systems	urinary system	cortex of kidney
E-MTAB-5214	tissues	adipose tissue	subcutaneous adipose tissue
E-MTAB-5214	tissues	adipose tissue	visceral adipose tissue
E-MTAB-5214	tissues	cellular	EBV-transformed lymphocyte
E-MTAB-5214	tissues	cellular	leukemia cell line
E-MTAB-5214	tissues	cellular	transformed fibroblast
E-MTAB-5214	tissues	cellular	whole blood
E-MTAB-5214	tissues	mammary tissue	breast
E-MTAB-5214	tissues	muscle	skeletal muscle
E-MTAB-5214	tissues	skin	skin of lower leg

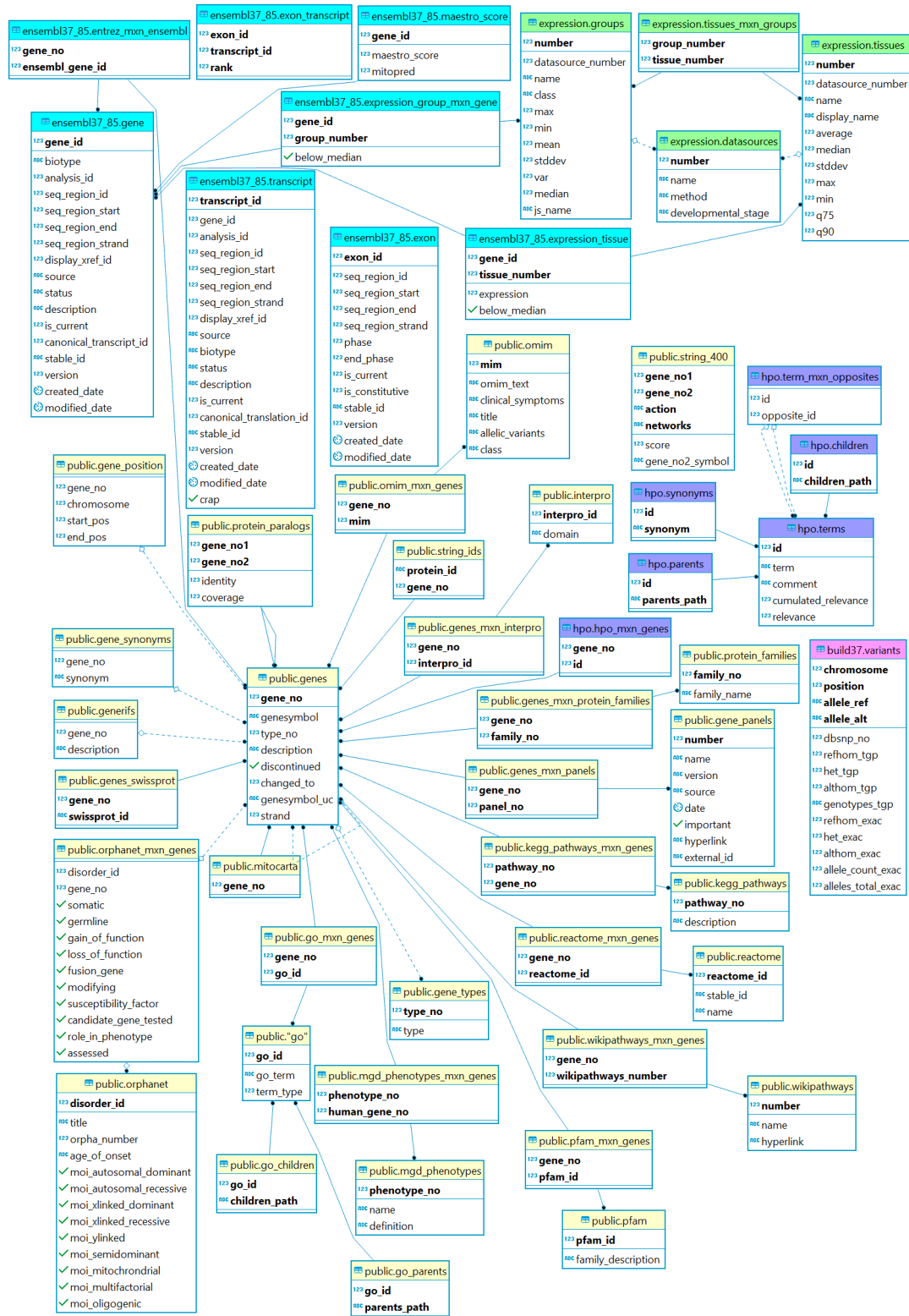
datasource	class	group	sub-tissues
E-MTAB-5214	tissues	skin	skin of suprapubic region
E-PROT-3,E-MTAB-2836	organs	brain	cerebral cortex
E-PROT-3,E-MTAB-2836	organs	gastrointestinal tract	appendix
E-PROT-3,E-MTAB-2836	organs	gastrointestinal tract	colon
E-PROT-3,E-MTAB-2836	organs	gastrointestinal tract	duodenum
E-PROT-3,E-MTAB-2836	organs	gastrointestinal tract	esophagus
E-PROT-3,E-MTAB-2836	organs	gastrointestinal tract	rectum
E-PROT-3,E-MTAB-2836	organs	gastrointestinal tract	small intestine
E-PROT-3,E-MTAB-2836	organs	gastrointestinal tract	stomach
E-PROT-3,E-MTAB-2836	organs	heart	heart
E-PROT-3,E-MTAB-2836	organs	kidney	kidney
E-PROT-3,E-MTAB-2836	organs	liver	liver
E-PROT-3,E-MTAB-2836	organs	lung	lung
E-PROT-3,E-MTAB-2836	organs	reproductive organs	endometrium
E-PROT-3,E-MTAB-2836	organs	reproductive organs	fallopian tube
E-PROT-3,E-MTAB-2836	organs	reproductive organs	ovary
E-PROT-3,E-MTAB-2836	organs	reproductive organs	placenta
E-PROT-3,E-MTAB-2836	organs	reproductive organs	prostate
E-PROT-3,E-MTAB-2836	organs	reproductive organs	testis
E-PROT-3,E-MTAB-2836	organs	skin	skin
E-PROT-3,E-MTAB-2836	organs	gallbladder	gallbladder
E-PROT-3,E-MTAB-2836	systems	circulatory/respiratory system	heart
E-PROT-3,E-MTAB-2836	systems	circulatory/respiratory system	lung
E-PROT-3,E-MTAB-2836	systems	endocrine system	adrenal gland
E-PROT-3,E-MTAB-2836	systems	endocrine system	pancreas
E-PROT-3,E-MTAB-2836	systems	endocrine system	thyroid
E-PROT-3,E-MTAB-2836	systems	food intake/digestion	appendix
E-PROT-3,E-MTAB-2836	systems	food intake/digestion	colon
E-PROT-3,E-MTAB-2836	systems	food intake/digestion	duodenum
E-PROT-3,E-MTAB-2836	systems	food intake/digestion	esophagus
E-PROT-3,E-MTAB-2836	systems	food intake/digestion	rectum
E-PROT-3,E-MTAB-2836	systems	food intake/digestion	salivary gland
E-PROT-3,E-MTAB-2836	systems	food intake/digestion	small intestine
E-PROT-3,E-MTAB-2836	systems	food intake/digestion	stomach
E-PROT-3,E-MTAB-2836	systems	immune system	lymph node
E-PROT-3,E-MTAB-2836	systems	immune system	spleen
E-PROT-3,E-MTAB-2836	systems	immune system	tonsil
E-PROT-3,E-MTAB-2836	systems	nervous system	cerebral cortex
E-PROT-3,E-MTAB-2836	systems	neuromuscular	skeletal muscle
E-PROT-3,E-MTAB-2836	systems	neuromuscular	smooth muscle
E-PROT-3,E-MTAB-2836	systems	reproductive system	endometrium
E-PROT-3,E-MTAB-2836	systems	reproductive system	fallopian tube
E-PROT-3,E-MTAB-2836	systems	reproductive system	ovary

datasource	class	group	sub-tissues
E-PROT-3,E-MTAB-2836	systems	reproductive system	placenta
E-PROT-3,E-MTAB-2836	systems	reproductive system	prostate
E-PROT-3,E-MTAB-2836	systems	reproductive system	testis
E-PROT-3,E-MTAB-2836	systems	skin	skin
E-PROT-3,E-MTAB-2836	systems	urinary system	bladder
E-PROT-3,E-MTAB-2836	systems	urinary system	kidney
E-PROT-3,E-MTAB-2836	tissues	adipose tissue	adipose tissue
E-PROT-3,E-MTAB-2836	tissues	bone marrow	bone marrow
E-PROT-3,E-MTAB-2836	tissues	muscle	skeletal muscle
E-PROT-3,E-MTAB-2836	tissues	muscle	smooth muscle
E-PROT-3,E-MTAB-2836	tissues	skin	skin
E-MTAB-513	organs	brain	brain
E-MTAB-513	organs	gastrointestinal tract	colon
E-MTAB-513	organs	heart	heart
E-MTAB-513	organs	kidney	kidney
E-MTAB-513	organs	liver	liver
E-MTAB-513	organs	lung	lung
E-MTAB-513	organs	reproductive organs	ovary
E-MTAB-513	organs	reproductive organs	prostate
E-MTAB-513	organs	reproductive organs	testis
E-MTAB-513	systems	circulatory/respiratory system	heart
E-MTAB-513	systems	circulatory/respiratory system	lung
E-MTAB-513	systems	endocrine system	adrenal gland
E-MTAB-513	systems	endocrine system	thyroid
E-MTAB-513	systems	food/digestion	colon
E-MTAB-513	systems	immune system	leukocyte
E-MTAB-513	systems	immune system	lymph node
E-MTAB-513	systems	nervous system	brain
E-MTAB-513	systems	neuromuscular	skeletal muscle
E-MTAB-513	systems	reproductive system	ovary
E-MTAB-513	systems	reproductive system	prostate
E-MTAB-513	systems	reproductive system	testis
E-MTAB-513	systems	urinary system	kidney
E-MTAB-513	tissues	adipose tissue	adipose tissue
E-MTAB-513	tissues	cellular	leukocyte
E-MTAB-513	tissues	mammary tissue	breast
E-MTAB-513	tissues	muscle	skeletal muscle
E-PROT-1	organs	brain	frontal cortex
E-PROT-1	organs	gastrointestinal tract	colon
E-PROT-1	organs	gastrointestinal tract	esophagus
E-PROT-1	organs	gastrointestinal tract	rectum
E-PROT-1	organs	heart	heart
E-PROT-1	organs	kidney	kidney

datasource	class	group	sub-tissues
E-PROT-1	organs	liver	liver
E-PROT-1	organs	lung	lung
E-PROT-1	organs	reproductive organs	ovary
E-PROT-1	organs	reproductive organs	prostate
E-PROT-1	organs	reproductive organs	testis
E-PROT-1	organs	gallbladder	gallbladder
E-PROT-1	systems	circulatory/respiratory system	heart
E-PROT-1	systems	circulatory/respiratory system	lung
E-PROT-1	systems	circulatory/respiratory system	platelet
E-PROT-1	systems	endocrine system	adrenal gland
E-PROT-1	systems	endocrine system	pancreas
E-PROT-1	systems	food intake/digestion	colon
E-PROT-1	systems	food intake/digestion	esophagus
E-PROT-1	systems	food intake/digestion	rectum
E-PROT-1	systems	immune system	B cell
E-PROT-1	systems	immune system	CD4-positive T cell
E-PROT-1	systems	immune system	CD8-positive T cell
E-PROT-1	systems	immune system	monocyte
E-PROT-1	systems	immune system	natural killer cell
E-PROT-1	systems	nervous system	frontal cortex
E-PROT-1	systems	nervous system	spinal cord
E-PROT-1	systems	reproductive system	ovary
E-PROT-1	systems	reproductive system	prostate
E-PROT-1	systems	reproductive system	testis
E-PROT-1	systems	urinary system	kidney
E-PROT-1	systems	urinary system	bladder
E-PROT-1	tissues	cellular	B cell
E-PROT-1	tissues	cellular	CD4-positive T cell
E-PROT-1	tissues	cellular	CD8-positive T cell
E-PROT-1	tissues	cellular	monocyte
E-PROT-1	tissues	cellular	natural killer cell
E-PROT-1	tissues	cellular	platelet
E-PROT-1	organs	eye	retina

## **C Appendix – MutationDistiller database ERD**

In this figure, I provide a comprehensive ERD of the tables and schemas used by MutationDistiller. Please note that this might change with updates and new versions. Symbols indicate data types: 123 - numeric; ABC - text; tick - boolean; clock - date. A more legible summarised ERD can be found in 2.3.



## Bibliography

1. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (Apr. 1953).
2. Franklin, R. E. & Gosling, R. G. Molecular configuration in sodium thymonucleate. 1953. *Nature* **421**, 400–401, discussion 396 (Jan. 23, 2003).
3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (Feb. 15, 2001).
4. Venter, J. C. *et al.* The sequence of the human genome. *Science (New York, N.Y.)* **291**, 1304–1351 (Feb. 16, 2001).
5. Hamacher, R. *et al.* Interleukin 1 beta gene promoter SNPs are associated with risk of pancreatic cancer. *Cytokine* **46**, 182–186 (May 2009).
6. Mathelier, A., Shi, W. & Wasserman, W. W. Identification of altered cis-regulatory elements in human disease. *Trends in Genetics* **31**, 67–76 (Feb. 2015).
7. Zheru, D. *et al.* Association of PPAR $\gamma$  gene polymorphisms with osteoarthritis in a southeast Chinese population. *Journal of Genetics* **93**, 719–723 (Dec. 2014).
8. Fearon, E. *et al.* Identification of a chromosome 18q gene that is altered in colorectal cancers. *Science* **247**, 49–56 (Jan. 5, 1990).
9. Drilon, A. MET Exon 14 Alterations in Lung Cancer: Exon Skipping Extends Half-Life. *Clinical Cancer Research* **22**, 2832–2834 (June 15, 2016).
10. Eswaran, J. *et al.* RNA sequencing of cancer reveals novel splicing alterations. *Scientific Reports* **3** (Dec. 2013).
11. Boon, K.-L. *et al.* prp8 mutations that cause human retinitis pigmentosa lead to a U5 snRNP maturation defect in yeast. *Nature Structural & Molecular Biology* **14**, 1077–1083 (Nov. 2007).
12. Wilkie, S. E. *et al.* Disease mechanism for retinitis pigmentosa (RP11) caused by missense mutations in the splicing factor gene PRPF31. *Molecular Vision* **14**, 683–690 (Apr. 18, 2008).
13. Drmanac, R. *et al.* Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* **327**, 78–81 (Jan. 1, 2010).
14. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463–5467 (Dec. 1977).



15. Troiano, G. The “Angelina Jolie effect” Gianmarco Troiano. *European Journal of Public Health* **27** (suppl\_3 Nov. 1, 2017).
16. Lippi, G. The risk of unjustified BRCA testing after the “Angelina Jolie effect”: how can we save (laboratory) medicine from the Internet? *Clinical Chemistry and Laboratory Medicine (CCLM)* **56**, e33–e35 (2018).
17. Nohdurft, E., Long, E. & Spinler, S. Was Angelina Jolie Right? Optimizing Cancer Prevention Strategies Among BRCA Mutation Carriers. *Decision Analysis* **14**, 139–169 (July 12, 2017).
18. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (Mar. 24, 2017).
19. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics* **97**, 199–215 (Aug. 2015).
20. Følling, I. The discovery of phenylketonuria. *Acta Paediatrica (Oslo, Norway: 1992). Supplement* **407**, 4–10 (Dec. 1994).
21. Van der Kloot, W. A. *et al.* The psychological burden of an initially unexplained illness: patients with sternocostoclavicular hyperostosis before and after delayed diagnosis. *Health and Quality of Life Outcomes* **8**, 97 (Sept. 2010).
22. Krabbenborg, L. *et al.* Understanding the Psychosocial Effects of WES Test Results on Parents of Children with Rare Diseases. *Journal of Genetic Counseling* **25**, 1207–1214 (Dec. 2016).
23. Burgstaller, J. P., Johnston, I. G. & Poulton, J. Mitochondrial DNA disease and developmental implications for reproductive strategies. *Molecular Human Reproduction* **21**, 11–22 (Jan. 2015).
24. Burgstaller, J. P. *et al.* MtDNA segregation in heteroplasmic tissues is common in vivo and modulated by haplotype differences and developmental stage. *Cell Reports* **7**, 2031–2041 (June 26, 2014).
25. Johnston, I. G. *et al.* Stochastic modelling, Bayesian inference, and new in vivo measurements elucidate the debated mtDNA bottleneck mechanism. *eLife* **4**, e07464 (June 2, 2015).
26. Turnbull, C. Introducing Whole Genome Sequencing into routine cancer care: The Genomics England 100,000 Genomes project. *Annals of Oncology* (Feb. 15, 2018).
27. Sun, Y. *et al.* Next-generation diagnostics: gene panel, exome, or whole genome? *Human Mutation* **36**, 648–655 (June 2015).

28. Lionel, A. C. *et al.* Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genetics In Medicine* **20**, 435 (Aug. 3, 2017).
29. Ng, S. B. *et al.* Targeted Capture and Massively Parallel Sequencing of Twelve Human Exomes. *Nature* **461**, 272–276 (Sept. 10, 2009).
30. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics* **42**, 30–35 (Jan. 2010).
31. Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences* **112**, 5473–5478 (Apr. 28, 2015).
32. Meienberg, J., Bruggmann, R., Oexle, K. & Matyas, G. Clinical sequencing: is WGS the better WES? *Human Genetics* **135**, 359–362 (2016).
33. Schwarze, K., Buchanan, J., Taylor, J. C. & Wordsworth, S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics in Medicine* **20**, 1122–1130 (Oct. 1, 2018).
34. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Research* **46**, D754–D761 (D1 Jan. 4, 2018).
35. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Research* **12**, 996–1006 (June 1, 2002).
36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–1760 (July 15, 2009).
37. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).
38. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (Nov. 23, 2006).
39. Bateman, A. *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45**, D158–D169 (D1 Jan. 4, 2017).
40. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**, 308–311 (Jan. 1, 2001).
41. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68 (Sept. 30, 2015).
42. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285 (Aug. 17, 2016).

43. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes: Supplementary Information. *bioRxiv* (Jan. 30, 2019).
44. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research* **44**, D862–D868 (Database issue Jan. 4, 2016).
45. Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics* **136**, 665–677 (June 2017).
46. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nature Methods* **11**, 361–362 (Apr. 2014).
47. Schwarz, J. M., Rödelberger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods* **7**, 575–576 (Aug. 2010).
48. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248–249 (Apr. 2010).
49. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nature Protocols* **11**, 1–9 (Dec. 3, 2015).
50. Hu, H. *et al.* VAAST 2.0: Improved Variant Classification and Disease-Gene Identification Using a Conservation-Controlled Amino Acid Substitution Matrix: VAAST 2.0. *Genetic Epidemiology* **37**, 622–634 (Sept. 2013).
51. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* (Oct. 29, 2018).
52. Yen, J. L. *et al.* A variant by any name: quantifying annotation discordance across tools and clinical databases. *Genome Medicine* **9**, 7 (Jan. 26, 2017).
53. McCarthy, D. J. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine* **6**, 26 (2014).
54. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**, 310–315 (Mar. 2014).
55. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (June 6, 2016).

56. Bieber, T. *et al.* Clinical phenotypes and endophenotypes of atopic dermatitis: Where are we, and where should we go? *Journal of Allergy and Clinical Immunology* **139**, S58–S64 (Apr. 2017).
57. Snoek, R., van Eerde, A. M. & Knoers, N. V. Importance of reliable variant calling and clear phenotyping when reporting on gene panel testing in renal disease. *Kidney International* **92**, 1325–1327 (Dec. 2017).
58. Köhler, S. *et al.* The Human Phenotype Ontology in 2017. *Nucleic Acids Research* **45**, D865–D876 (D1 Jan. 4, 2017).
59. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* **43**, D789–D798 (Database issue Jan. 28, 2015).
60. INSERM. *Orphanet: an online database of rare diseases and orphan drugs*. 1997.
61. Kingsmore, S. Comprehensive Carrier Screening and Molecular Diagnostic Testing for Recessive Childhood Diseases. *PLoS Currents* **4** (May 2, 2012).
62. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**, 25–29 (May 2000).
63. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45**, D353–D361 (D1 Jan. 4, 2017).
64. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Research* **46**, D649–D655 (D1 Jan. 4, 2018).
65. Slenter, D. N. *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research* **46**, D661–D667 (D1 Jan. 4, 2018).
66. Petryszak, R. *et al.* Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Research* **42**, D926–D932 (D1 Jan. 2014).
67. Lambrix, P., Tan, H., Jakoniene, V. & Strömbäck, L. in *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences* (eds Baker, C. J. O. & Cheung, K.-H.) 85–99 (Springer US, Boston, MA, 2007).
68. Gruber, T. R. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies* **43**, 907–928 (Nov. 1995).

69. Resnik, P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* **11**, 95–130 (July 1, 1999).
70. Shyr, C., Kushniruk, A., van Karnebeek, C. D. & Wasserman, W. W. Dynamic software design for clinical exome and genome analyses: insights from bioinformaticians, clinical geneticists, and genetic counselors. *Journal of the American Medical Informatics Association : JAMIA* **23**, 257–268 (Mar. 2016).
71. Sifrim, A. *et al.* eXtasy: variant prioritization by genomic data fusion. *Nature Methods* **10**, 1083–1084 (Nov. 2013).
72. Javed, A., Agrawal, S. & Ng, P. C. Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nature Methods* **11**, 935–937 (Sept. 2014).
73. Robinson, P. N. *et al.* Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research* **24**, 340–348 (Feb. 2014).
74. Singleton, M. V. *et al.* Phevor Combines Multiple Biomedical Ontologies for Accurate Identification of Disease-Causing Alleles in Single Individuals and Small Nuclear Families. *American Journal of Human Genetics* **94**, 599–610 (Apr. 3, 2014).
75. Zemojtel, T. *et al.* Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science Translational Medicine* **6**, 252ra123 (Sept. 3, 2014).
76. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature Protocols* **10**, 1556–1566 (Oct. 2015).
77. Antanaviciute, A. *et al.* OVA: Integrating molecular and physical phenotype data from multiple biomedical domain ontologies with variant filtering for enhanced variant prioritization. *Bioinformatics*, btv473 (Aug. 12, 2015).
78. Alemán, A., Garcia-Garcia, F., Salavert, F., Medina, I. & Dopazo, J. A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. *Nucleic Acids Research* **42**, W88–93 (Web Server issue July 2014).
79. Bertoldi, L. *et al.* QueryOR: a comprehensive web platform for genetic variant analysis and prioritization. *BMC Bioinformatics* **18**, 225 (Apr. 28, 2017).
80. Glanzmann, B. *et al.* A new tool for prioritization of sequence variants from whole exome sequencing data. *Source Code for Biology and Medicine* **11**, 10 (July 1, 2016).

81. Seelow, D., Schwarz, J. M. & Schuelke, M. GeneDistiller—Distilling Candidate Genes from Linkage Intervals. *PLoS ONE* **3** (ed Awadalla, P.) e3874 (Dec. 5, 2008).
82. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (Sept. 5, 2012).
83. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* **39**, D52–D57 (Database issue Jan. 2011).
84. Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science (New York, N.Y.)* **185**, 862–4 (Oct. 1, 1974).
85. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**, 1034–1050 (Aug. 2005).
86. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20**, 110–121 (Jan. 1, 2010).
87. Tatusova, T. & Madden, T. BLAST 2 SEQUENCES, a new tool for comparing protein and nucleotide sequences (vol 174, pg 247, 1999). *FEMS microbiology letters* **174**, 247–50 (June 1, 1999).
88. Yeo, G. & Burge, C. B. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *Journal of Computational Biology* **11**, 377–394 (Mar. 2004).
89. Tabaska, J. E. & Zhang, M. Q. Detection of polyadenylation signals in human DNA sequences. *Gene* **231**, 77–86 (Apr. 29, 1999).
90. Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D. Improved Splice Site Detection in Genie. *Journal of Computational Biology* **4**, 311–323 (Jan. 1997).
91. Wilming, L. G. *et al.* The vertebrate genome annotation (Vega) database. *Nucleic Acids Research* **36**, D753–D760 (Database Dec. 23, 2007).
92. Pruitt, K. D. *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research* **19**, 1316–1323 (July 1, 2009).
93. Kinsella, R. J. *et al.* Ensembl BioMart: a hub for data retrieval across taxonomic space. *Database* **2011**, bar030–bar030 (July 23, 2011).
94. Blake, J. A. *et al.* Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Research* **45**, D723–D729 (Database issue Jan. 4, 2017).

95. Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics* **84**, 524–533 (Apr. 2009).
96. Noguchi, S. *et al.* FANTOM5 CAGE profiles of human and mouse samples. *Scientific Data* **4**, 170112 (Aug. 29, 2017).
97. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (Oct. 11, 2017).
98. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419–1260419 (Jan. 23, 2015).
99. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research* **44**, 11033–11033 (Dec. 15, 2016).
100. Daoud, H. *et al.* Next-generation sequencing for diagnosis of rare diseases in the neonatal intensive care unit. *Canadian Medical Association Journal* **188**, E254–E260 (Aug. 9, 2016).
101. Dohrn, M. F. *et al.* Frequent genes in rare diseases: panel-based next generation sequencing to disclose causal mutations in hereditary neuropathies. *Journal of Neurochemistry* **143**, 507–522 (Dec. 1, 2017).
102. Kalia, S. S. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genetics in Medicine: Official Journal of the American College of Medical Genetics* **19**, 249–255 (2017).
103. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44**, D279–D285 (D1 Jan. 4, 2016).
104. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Research* **42**, D222–D230 (D1 Jan. 2014).
105. Finn, R. D. *et al.* InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research* **45**, D190–D199 (D1 Jan. 4, 2017).
106. Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research* **45**, D183–D189 (D1 Jan. 4, 2017).
107. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research* **46**, D493–D496 (D1 Jan. 4, 2018).
108. Von Mering, C. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research* **33**, D433–D437 (Database issue Dec. 17, 2004).

109. Herst, P. M., Rowe, M. R., Carson, G. M. & Berridge, M. V. Functional Mitochondria in Health and Disease. *Frontiers in Endocrinology* **8** (Nov. 3, 2017).
110. Calvo, S. *et al.* Systematic identification of human mitochondrial disease genes through integrative genomics. *Nature Genetics* **38**, 576–582 (May 2006).
111. Calvo, S. E., Clauser, K. R. & Mootha, V. K. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Research* **44**, D1251–D1257 (D1 Jan. 4, 2016).
112. Guda, C., Guda, P., Fahy, E. & Subramaniam, S. MITOPRED: a web server for the prediction of mitochondrial proteins. *Nucleic Acids Research* **32**, W372–W374 (Web Server July 1, 2004).
113. Team, R. C. *R: A Language and Environment for Statistical Computing* Vienna, Austria, 2017.
114. Wickham, H. The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*. <http://www.jstatsoft.org/v40/i01/> **40**, 1–29 (2011).
115. Wickham, H., Francois, R., Henry, L. & Müller, K. *dplyr: A Grammar of Data Manipulation* (2017).
116. Dahl, D., Scott, D., Roosen, C., Magnusson, A. & Swinton, J. *xtable: Export Tables to LaTeX or HTML* 2018.
117. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (2009).
118. Wickham, H. Reshaping data with the reshape package. *Journal of Statistical Software* **21** (2007).
119. Köhler, S. *et al.* Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *The American Journal of Human Genetics* **85**, 457–464 (Oct. 2009).
120. Bone, W. P. *et al.* Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genetics in Medicine* **18**, 608–617 (June 2016).
121. Smedley, D. *et al.* Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols* **10**, 2004–2015 (Dec. 2015).
122. Girdea, M. *et al.* PhenoTips: Patient Phenotyping Software for Clinical and Research Use. *Human Mutation* **34**, 1057–1065 (Aug. 2013).
123. Hombach, D. *et al.* Phenotero: Annotate as you write. *Clinical Genetics* **95**, 287–292 (Feb. 2019).
124. Wright, C. F. *et al.* Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *GENETICS in MEDICINE* (Jan. 11, 2018).



125. Pontikos, N. *et al.* Phenopolis: an open platform for harmonization and analysis of genetic and phenotypic data. *Bioinformatics* **33**, 2421–2423 (Mar. 15, 2017).
126. Smith, C. L. *et al.* Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Research* **46**, D836–D842 (D1 Jan. 4, 2018).
127. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239 (Oct. 11, 2017).
128. Zhao, J. *et al.* A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. *The American Journal of Human Genetics* **98**, 299–309 (Feb. 2016).
129. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine* **9**, eaal5209 (Apr. 19, 2017).
130. Li, X. *et al.* Transcriptome Sequencing of a Large Human Family Identifies the Impact of Rare Noncoding Variants. *The American Journal of Human Genetics* **95**, 245–256 (Sept. 2014).
131. Shovlin, S. & Tropea, D. Transcriptome level analysis in Rett syndrome using human samples from different tissues. *Orphanet Journal of Rare Diseases* **13** (Dec. 2018).
132. Mears, A. J. *et al.* Mining the transcriptome for rare disease therapies: a comparison of the efficiencies of two data mining approaches and a targeted cell-based drug screen. *npj Genomic Medicine* **2** (Dec. 2017).
133. Feiglin, A., Allen, B. K., Kohane, I. S. & Kong, S. W. Comprehensive Analysis of Tissue-wide Gene Expression and Phenotype Data Reveals Tissues Affected in Rare Genetic Disorders. *Cell Systems* **5**, 140–148.e2 (Aug. 2017).
134. Grayson, P. C. *et al.* Metabolic pathways and immunometabolism in rare kidney diseases. *Annals of the Rheumatic Diseases*, annrheumdis–2017–212935 (May 3, 2018).
135. Vafai, S. B. & Mootha, V. K. A Common Pathway for a Rare Disease? *Science* **342**, 1453–1454 (Dec. 20, 2013).
136. Wong, M. Mammalian Target of Rapamycin (mTOR) Pathways in Neurological Diseases. *Biomedical Journal* **36**, 40 (2013).
137. Reijnders, M. R. F. *et al.* Variation in a range of mTOR-related genes associates with intracranial volume and intellectual disability. *Nature Communications* **8** (Dec. 2017).
138. Lee, H. *et al.* Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. *JAMA* **312**, 1880 (Nov. 12, 2014).

139. Taylor, J. C. *et al.* Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics* **47**, 717–726 (July 2015).
140. Pengelly, R. J. *et al.* Evaluating phenotype-driven approaches for genetic diagnoses from exomes in a clinical setting. *Scientific Reports* **7** (Dec. 2017).
141. Retterer, K. *et al.* Clinical application of whole-exome sequencing across clinical indications. *Genetics In Medicine* **18**, 696 (Dec. 3, 2015).
142. Schwarz, J. M. *et al.* RegulationSpotter: annotation and interpretation of extratranscriptomic DNA variants. *Nucleic Acids Research* (May 20, 2019).
143. Shyr, C., Kushniruk, A. & Wasserman, W. W. Usability study of clinical exome analysis software: Top lessons learned and recommendations. *Journal of Biomedical Informatics* **51**, 129–136 (Oct. 2014).
144. Machini, K., Douglas, J., Braxton, A., Tsipis, J. & Kramer, K. Genetic Counselors' Views and Experiences with the Clinical Integration of Genome Sequencing. *Journal of Genetic Counseling* **23**, 496–505 (Aug. 2014).

## **Statement of independent work**

Herewith I confirm that I wrote this Dissertation in its entirety and that no additional assistance was provided, other than from the sources listed.

Signed: \_\_\_\_\_