Enabling Versatile And Comprehensive Analysis Of Genomic Variant Data

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Sebastian Ginzel
aus Dormagen

Düsseldorf, April 2019

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf
Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1.(Referent) Prof. Dr. Egon Wanke

2.(Korreferent) Prof. Dr. Arndt Borkhardt

3.(Korreferent) Prof. Dr. Ralf Thiele

Tag der mündlichen Prüfung: 28.06.2019

EIDESSTATTLICHE ERKLÄRUNG (AFFIDAVIT)

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Ort, Datum

Sebastian Ginzel

ABSTRACT

Next generation sequencing (NGS) is able to identify hundreds of thousands of mutations per individual, revealing new insights for research and medical treatment. This has lead to an improved characterization of tumors, discoveries of new disease-causing mechanisms in genetic diseases, as well as the identification of new treatment options. Medical sequencing experiments are multidisciplinary efforts that require members with varying professions and degrees of expertise to generate and process the data. Identify variants of significance in a haystack of mutations is done through interpretation by genomic experts (e.g. molecular biologists and medical doctors).

The initially large number of variants is reduced by applying custom variant annotation and filtering procedures. This requires complex software toolchains to be set up and data sources to be integrated. Furthermore, increasing study sizes subsequently require higher efforts to manage datasets in a multi-user and multi-institution environment. It is common practice to expect numerous iterations of continuative respecification and refinement of filter strategies, when the cause for a disease or phenotype is unknown. Data analysis support during this phase is fundamental, because handling the large volume of data is not possible or inadequate for users with limited computer literacy. Constant feedback and communication is necessary when filter parameters are adjusted or the study grows with additional samples. Consequently, variant filtering and interpretation becomes time-consuming and hinders a dynamic and explorative data analysis by experts.

In this work I present SNuPy, an interactive tool that empowers genomic experts to analyze their own variant datasets. A user-friendly interface allows to manage datasets and filter small variants (SNV/Indel), as well as copy number variants from thousands of samples in parallel. Utilizing SNuPy, genomic experts can perform quality control to verify the correctness of datasets, execute parameterized multi-criterial queries to find mutations of interest, and are enabled to refine queries without additional bioinformatic support. I present a variant discovery platform that addresses the short-comings of current solutions for this task.

SNuPy was deployed in the sequencing facility in one of Germanys largest pediatric oncologies to handle hundreds of millions of genotyped variants in a user-friendly platform, managing more than 5000 variant datasets. It has successfully contributed to a broad range of research projects as part of oncological (7 times), immunological (6 times), drug-resistance and clinical diagnostic studies in human and mice (2 times).

ZUSAMMENFASSUNG

Die Next Generation Sequenzierung (NGS) identifiziert hunderttausende Mutationen pro Individuum und eröffnet der Forschung und medizinischen Behandlung neue Erkenntnisse. Dies führte zur verbesserten Charakterisierung von Tumoren, der Entdeckung neuer krankheitsverursachender Mechanismen bei genetischen Erkrankungen sowie der Identifizierung neuer Behandlungsmöglichkeiten. Sequenzierungexperimente sind multidisziplinäre Projekte, die die Fachkenntnisse unterschiedlicher Experten benötigen, um Daten zu generieren und zu verarbeiten. Die Ergebnisse werden von Genomikern (z.B. Molekularbiologen und Mediziner) interpretiert, um signifikante Varianten in einem Heuhaufen von Mutationen zu finden.

Die anfänglich große Anzahl der Varianten wird durch die Ausführung angepasster Annotations- und Filterprozesse reduziert. Dies erfordert den Aufbau komplexer Software-Werkzeugketten und die Integration unterschiedlicher Datenquellen. Darüber hinaus bedeuten steigende Studiengrößen einen zunehmend höheren Verwaltungsaufwand der Datensätzen in einer Multi-User und -Institutionsumgebung. Die Unterstützung der Datenanalyse in dieser Phase ist von grundlegender Bedeutung, da das große Datenvolumen für Benutzer mit eingeschränkter Computerkenntnis nicht oder unzureichend handhabar ist. Wenn die Ursache für eine Krankheit oder einen Phänotyp unbekannt ist, ist es üblich, dass über mehrere Iterationen hinweg Filterstrategien und Problemspezifizierungen verfeinert werden. Ständige Rückmeldung und Kommunikation ist notwendig, wenn Parameter angepasst werden müssen oder sich die Datengrundlage ändert (z.B. wenn weitere Proben hinzugefügt werden). Die Variantenfilterung und -interpretation werden so zeitaufwendig und erschweren eine dynamische und explorative Datenanalyse durch Experten.

In dieser Arbeit stelle ich SNuPy vor, ein interaktives Werkzeug, das Forschern mit eingeschränkten Computerkenntnisse die Möglichkeit gibt, ihre eigenen Variantendatensätze zu analysieren. SNuPy erlaubt Genomik-Experten mit Qualitätskontrollen die Korrektheit von Datensätzen zu überprüfen und diese zu verwalten. Kleine Varianten (SNV/Indel) sowie Kopienzahl-Varianten aus hunderten Proben können über eine benutzerfreundliche Oberfläche parallel gefiltert werden. Dazu ist es möglich parametrisierte und aus multiplen Kriterien bestehende Abfragen durchzuführen, so relevante Mutationen zu finden, oder verfeinerte Abfragen ohne bioinformatische Unterstützung neu zu stellen. Ich präsentiere eine Variantenentdeckungsplattform, die die Mängel aktueller Werkzeuge in diesen Aspekten löst.

SNuPy wurde in der Sequenziereinheit in einer der größten pädiatrischen Onkologien Deutschlands eingesetzt, um hunderte Millionen genotypisierter Varianten

über eine benutzerfreundliche Plattform zu verarbeiten und mehr als 5000 Variant-Datensätze zu verwalten. Es wurde erfolgreich in einem breiten Spektrum von Forschungsprojekten im Rahmen von onkologischen- (7 mal), immunologischen- (6 mal), medikamentenresistenz- und klinischen Diagnose- Studien an Menschen und Mäusen (2 mal) eingesetzt.

CONTENTS

Part I

INTRODUCTION

1

MOTIVATION

The advent and success of Next-Generation Sequencing (NGS) technology in research and clinical applications has undoubtedly shaped the way genetic diseases are studied and treated. With sequencing cost of a human genome dropping from hundreds of million dollars to hundreds of dollars and the time required dropping from years to days or even hours[1] this technology is now widely available for laboratories and hospitals in the developed world.

Recent years have seen an increase in genetic and genomic information as part of population sequencing studies as well as disease centered studies[2,3,4,5]. The Leiden Open Variation Database (LOVD)[6] project currently lists 81 projects, reporting on different disease specific or population wide DNA variations summing up to 3,953,657 unique variations (as of 24th Sep. 2018). ClinVar[7], a database of clinically described variations holds more than 400,000 variations of different significance levels (as of 24th Sep. 2018)

However, there remains a discrepancy between the availability and potential usage of genomic data in clinical research and applications, as identified by Yang et al.:

> *"Many patients with genetic diseases are not given a specific diagnosis. The standard of practice involves the recognition of specific phenotypes [. . .] or the selection of candidate-gene tests, including single-gene analysis and gene-panel tests. The majority of patients remain without a diagnosis."*[8]

Efforts in the field of precision medicine (PM) are working towards overcoming this discrepancy[9,10].

> *"PM seeks to improve stratification and timing of health care by utilizing biological information and biomarkers on the level of molecular disease pathways, genetics, proteomics as well as metabolomics."*[11]

These efforts are not new, *"blood typing, for instance, has been used to guide blood transfusions for more than a century"*[9] and even the use of genetic markers have been used successfully for decades (e.g. HER-2 in breast cancer[12]). However, the era of NGS technology now provides unprecedented resolution and availability as a toolkit for diagnosis and research, and consequently has entered medical practice[13,14,15].

Oncology has been, and still is the vanguard in application and implementation of NGS based diagnosis and treatment decisions, in part because it is able to unravel

the heterogeneous genetic landscape of cancer[16]. Examples for this are sequenced-based identification of minimal residual disease (MRD)[17,18], the identification of leukemia subtypes[19,20] as well pharmacogenomics approaches, a subtype of PM aiming to assess drug efficacy[21,22,23]. Recently, new disease causing mechanisms for leukemia have been revealed using NGS technology hinting at the role of common pathogens in early leukemia development[24,25].

Although PM is met with skepticism by some[26], and has limits[27], NGS and its potential to transform PM is part of the solution for future developments[9,28,29]. Especially when new mechanisms that influence disease progression, such as microbiota[29] are getting more attention.

Moreover, since 2007 the drop in sequencing costs have outperformed the highly referenced Moore's law of semiconductor cost decrease[30] constantly. It has been estimated that sequencing instruments in 2013 produced 15 petabytes (10^6 gigabytes) of data, this number is forecast to reach 1 zettabyte (10^6 petabytes) by 2025[31].

As a consequence researchers and clinicians are faced with the *"DNA data deluge"*[32]. Between 20,000 and 80,000 variants fall in the coding region of human individuals[33,34,35], which make up around 2% of the whole genome. Besides all the clinical, medical and biological challenges PM and molecular biology research faces, the shear amount of data that researchers and clinicians need to sift through is immense and close to being useless without appropriate tools[32].

In order to identify significant variants, genomic scientists have to be able to query the datasets, interpret the result and possibly refine the previous query. The required filter conditions are complex, possibly involving hundreds of samples, a multitude of annotation features from difference sources to filter by, usage of shared controls and all is performed in a collaborative environment that requires multiple disciplines and institutions to work together (see Brownstein and Others, Amendola et al.).

The integration of bioinformaticians into research groups to support data analysis has been standard practice in recent years. This requires constant communication and feeback when filter criteria need to be adjusted in response to a previous query result in an iterative fashion. Consequently, such setups hinder dynamic and explorative approaches that can empower genomic experts to query complex variant datasets and interpret the results independently.

1.0.1 *Outline*

This thesis aims to systematically analyze the requirements that are necessary for genomic scientists to perform filtering, analysis and interpretation of genomic variants from NGS experiments. I will further show how to design a platform that empowers genomic scientists to perform queries on genomic variant datasets without additional bioinformatic support. Additionally I will show how variant inter-

pretation, data management and quality control can be supported to enable the identification of significant variants from hundreds of samples.

Because the work is set in a interdisciplinary field at the intersection between bioinformatic, molecular and clinical research I will give a brief introduction into the bio-medical background in the rest of this introductory part. Next we will look at existing solutions and the *State of the Art* for variant discovery tools. Here we will examine the current software landscape, that is available to researchers to filter and process variant datasets. The third part will analyze the *Requirements* for a versatile and comprehensive variant discovery platform. That part consists of the variant analysis requirements on one hand and the technical platform requirements on the other. The two subsequent parts will present a *Concept* and its *Implementation*. These parts will focus on a design that meets the requirements and the development of modules for a web application platform that is made available to genomic experts. The two final chapters will present statistics about the feasibility of our approach and demonstrate how the software aided in the clinical and molecular research over the last years of its development, as well as give an outlook for the road ahead.

2

GENOMICS & NEXT GENERATION SEQUENCING

2.1 FROM DNA TO DISEASE

2.1.1 *From DNA to Protein*

Genetic information is encoded in nucleic acids (adenine(A), guanine(G), cytosine(C) & thymine(T)) that form the double-helix structure of deoxyribonucleic acids (DNA). A only binds to T, and G only to C when they are on opposite sides of the double helix, hence their name: basepairs (bp). Long chains of these structures are packaged into so called chromosomes, which humans have 23 pairs of, 22 autosomes (no. 1-22) and one pair of gonosomes (X & Y, in combinations of XX or XY), which determine the sex[a]. These chromosomes form the genome of an organism and describe all inherited information that is passed down to the next generation.

Because each chromosome is redundant there are always two copies (alleles), which are independent from one another. Sex gonosomes are an exception because human male carry XY and female carry XX, thus only females have redundancy in the X-chromosome.

Genes are species-specific regions of the genome that contain the necessary blueprints for proteins - the structures that provide biological functions by interacting with each other or the environment. The process how genes, which are tightly packaged into large chromosomes, are extracted and used as blueprints is called protein biosynthesis. It is split into two subprocesses called transcription and translation, one responsible for extracting and copying genetic information (transcription) and the other using this information to create proteins from amino acid (translation).

Transcription starts in the region before the gene ("upstream", so called 5'UTR) and ends after the gene ("downstream", so called 3'UTR). The process is started by the binding and accumulation of transcript factors, guiding the so called RNA polymerase to the start of the relevant genetic information. An RNA polymerase catalyzes the duplication of a single strand from the double-stranded DNA structure, so called ribonucleic acid in single stranded form (hence RNA)[b]. The bases comprising an RNA as also referred to as basepairs, although they are single stranded and are usually not bound to another nucelic/ribonucleic acid.

A single string of RNA consists of two components, exons and introns. The latter are subunits of the RNA that are removed from the RNA in a process called

a Other mammals and organisms may have different autosome/gonosome names and distributions

b During this process thymine is replaced by uracil(denoted U)

splicing, leaving only a continuous string of exons. This transcribed and spliced structure is called mRNA and is then passed from the cell nucleus to the cytoplasm, where it is used as a template for the translation process. The regions mentioned above, namely 5'UTR, 3'UTR and introns are however not useless in the process, but may act as (auto-)regulatory elements for the transcriptional or translational process.

The synthesis of proteins is done during the translational process, using mRNA strings as templates, which are virtually subdivided into basepair triplets, so called codons. A protein group called Ribosomes start the translational process that subsequently recruits amino acids one-by-one, as a function of a codon and using structures called transport RNA (tRNA). These RNA structures are comprised of a RNA binding site that is three basepairs long and matches the compliment of the RNA codon sequence. On the other side it carries a single amino acid. The process stops once a special codon sequence is detected and the fully functional protein is released from the ribosome. In total a triplet of basepairs allows for 64 possible combinations, but only 21 amino acids are available in humans, leading to a redundancy in codon codes. Addiontally a set of special triplets called stop codons are not associated to amino acids, but rather quit the translational process.

Post translational processes and interactions with other proteins then make up the biological function required in cells to perform basic tasks (house keeping) or more specialized tasks, depending on the tissue the cell developed into.

The body of genetic DNA information is called genome and genes only make up 2% of the around 3 billion base pairs of genetic information the human genome contains. The 98% of genetic information was seen as evolutionary junk for decades, but has been revealed to carry so called epigenetic functionality that explains inheritance without change of DNA code. Other larger parts of DNA act as regulatory regions that are able to regulate mRNA expression and protein abundance. Smaller parts, so called microRNA act as a regulatory mechanism for mRNA abundance and regulate degradation of them. Most importantly, when it comes to whole-exome sequencing (see section2.2.2): The body of information that are templates for protein products is called *exome* because only exons contain parts of DNA relevant to protein synthesis.

2.1.2 *From Cell to Cell and parent to child*

Cells, the smallest living parts of an organism use cell division to replicate themselves, while larger living organism reproduce. The main difference, is that replication lead to a copy of the same cell, while reproduction results in a new cell with two(or more) predecessors. The composition of genetic information in the first case stays the same, but changes in the latter.

The life of a cell is called cell cycle, which is comprised of 5 phases:

GAP 0 (G_0) The cell does not divide any more.

GAP 1 (G_1) Cell size increases and a check (G_1 checkpoint) is performed that ensures everything is ready for S-phase.

SYNTEHSIS (S) Is the synthesis of DNA, which replicates the genome. The genome is now doubled and 46 chromosomes are present.

GAP 2 (G_2) The cell continuous to increase and an G_2 checkpoint makes sure that M-phase is possible.

MITOSIS (M) This is the cell division phase that results in the 46 chromosomes being correctly divided up between the two new cells.

The result are two genetically identical cells.

Reproduction on the other hand requires two cells with distinct genetic information to fuse into a single one. The first step of this is a process called meiosis, whereby a cell looses half of its genetic information and retains no redundant copy of the chromosomes. Furthermore, the four sister cells contain a shuffled version of each chromosome (sister chromatids) because of recombination events that occur during the process. These make up the so called germline of an individual, the genetic information that is passed on to its offspring. The result are sex cells (gametes), carrying only a single (possible recombined) copy of the source cell. Gametes are used in sexual reproduction to form a new single cell with two chromosome pairs from two individuals, from which an offspring is born.

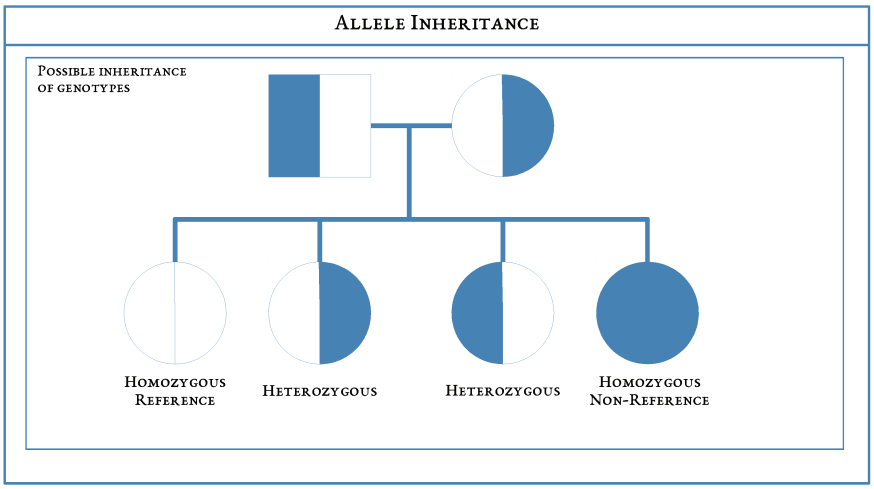


Figure 2.1: Possible inheritance of alleles for a single loci. In general each copy is inherited independently from the other.

Although most of the genes in humans are conserved, generally two humans do not carry the same genetic information (with the exception of identical twins), every individual carries different variants[c] of the same gene. These variations are

c The term mutation has seen a decline in its use in favor of the more neutral terms variation or variant[38]

passed on to offspring, resulting in four possible combinations (see Figure 2.1). In total there are four classes of variants that are relevant for genetics:

SINGLE NUCLEOTIDE VARIANT (SNV) indicating only a single basepair change.

INSERTION/DELETION (INDEL) Insertion/Deletion, local insertion or deletion of genetic information. The range is considerably large, reaching from 50 bp[39] to 1kb[40].

COPY NUMBER VARIANT (CNV) a larger deletions or duplication of chromosomal DNA (>1000 bases) possibly spanning multiple gene regions[40].

STRUCTURAL VARIANT (SV) Large scale fusions of chromosomal regions.

The term SNP (for Single Nucleotide Polymorphism) is often used and historically has the same meaning as SNV. However, in recent years SNP and in lights of large sequencing projects the term SNP specifically refers to SNV which are rare in the common human population(<10%).

A variant is always a change in comparison to the human reference genome, which is based on DNA of thirteen anonymous individuals from the Buffalo, New York area[41,42].

2.1.3 *From defect to disease*

Protein biosynthesis is a complex molecular machinery that has remarkable error detection and repair mechanisms during the different stages. If errors are detected there are fail-safe mechanisms to prevent erroneous RNA from being transcribed or translated.

The same is true for cells and their life cycles. If checkpoints fail the cell will start a suicide process called apoptosis, triggering internal metabolic processes that eventually lead to a fragmentation of the cell, which is cleared up by other cells of the body. Apoptosis can be triggered extrinsically as well, when certain receptors on the cell surface are activated by other cells via proteins or compounds (e.g. drugs).

However, when these failsafe mechanisms are not successful the fine balanced biologic machinery gets disrupted, resulting in dysfunctional proteins being built or genetic copies of the cell carrying new (de novo) variants. If proteins of critical biological functions or function cascades (pathways) are affected by modifications, this can lead to possibly life threatening genetic disease (as compared to bacterial, viral or fungal diseases).

A common theme in oncology are cells that are not able to initiate or complete an apoptotic process, or evade the immune response, resulting in uncontrolled and unchecked cell division. The sister cells of the first founding cell with a genetic defect then carry the same genetic defect, proliferate the same way without being able to go into apoptosis, replicating again and eventually driving out healthy cells

from their environment, leading to a collapse of the biological function supporting the organism.

The reasons for possible DNA damage are plenty, reaching from intrinsic failures in the DNA copy mechanism, missing gene redundancy due to consanguine background to extrinsic factors from the environment, radiation, drugs or pathogens. Technology allows to observe the consequences of genetic aberrations on the full spectrum of the protein biosynthesis.

In general, when a biological measure can be used for diagnosis or treatment decisions, these measures are called biomarker. There are biomarkers that are easily obtained and that have been long used in medicine, such as heart rate or blood type. Molecular biomarkers are more complex to obtain, but play an important role in oncology today (e.g HER-2 receptors for breast cancer[12]). With the advent of NGS the prospect is that precise genetic markers will aid in PM approaches to find the best treatment options utilizing the whole genome.

2.2 INTRODUCTION TO NEXT GENERATION SEQUENCING

Sanger-sequencing was the prevailing method to sequence DNA strands for decades. The general process behind Sanger-sequencing is to break a single piece of DNA into pieces that share the same sequence prefix, sorting them by length and use a readout method to determine the last basepair of each piece. This results in a signal for each possible nucleic acid at each position of the originally shattered DNA sequence. It was the fundamental technology for the assembly of the human genome and is still used today to validate results or confirm variants in small areas of a gene.

So called next generation sequencing summarize high-throughput sequencing technologies, not based on Sanger-sequencing. These technologies use different readouts using light sensors or semi-conductors to record millions of single molecular reactions in parallel and compile them into sequences.

Available technologies are called PacBio (by Pacific Biosciences), Ion torrent sequencing (by Ion Torrent), Pyrosequencing (454 Life Sciences), synthesis based sequencing (by Illumina), ligation sequencing (by SOLiD) and Nanopore sequencing (by Oxford Nanopore). The technologies differ in the required biological protocols and technical readout techniques, but most importantly in their error rate, maximal achievable read length and their cost per basepair (see Levy and Myers [43] for a review on available technologies).

Illumina is currently the most widely used technology, able to sequence DNA fragments up to reads of 300 basepairs, while other technologies such as PacBio provide read lengths of up to 20.000 basepairs[44]. Even longer reads in the range of 400.000-500.000 basepairs are also being developed[43]. Long-read technologies are mostly used to assemble previously unknown genomes (de novo assembly), phasing or haplotyping, allowing investigators to track the paternal origin of specific alleles.

Sequencing itself must be seen as a sampling process, in which the DNA of a random set of cells is fragmented and each fragment is randomly chosen to be successfully sequenced. One way to overcome this is to be able to sequence all possible fragments that went into the sequencer, but this is usually limited by the amount of basepairs a machine can detect per run.

2.2.1 *Fields of application*

Next generation sequencing technologies has been used in a wide range of application, among them:

WHOLE GENOME SEQUENCING (WGS) Sequencing of the entire genomes.

WHOLE EXOME SEQUENCING (WES) Sequencing of the entire exome.

RNA-SEQ sequencing of the mRNA content of cells.

CHROMATIN IMMUNOPRECIPITATION (CHIP) used to identify genomic sites where transcription starts (transcription factor binding sites).

BISULFITE SEQUENCING used to identify which chromosomal regions are available for transcription.

MICRO BIOME & METAGENOMICS allows the sequencing of environmental samples, such as water, soil, food and the gut.

The result of sequencing is always a list of reads that were produced by the readout of individual fragments. For large genomes, where the technology does not allow the complete coverage of the genome with a single read, this step is followed-up with an alignment of the reads against the reference genome (alignment). The complexity of this process depends on the length of the reads and error rate at which bases are classified incorrectly during sequencing.

RNA-Seq, ChIP and bisulfite sequencing do not look at genetic information, but rather expression, regulatory or epigenetic factors. Micro biome (also called metagenomics) mostly target the bacterial or viral composition of a sample, by targeted sequencing of species specific sites. This work will focus on WGS and WES to identify SNV , InDel and CNV for the rest of the work because these applications allow the detection of clinically relevant genetic aberrations. Another reason I will focus on these variation types is the large number of variations and the genes they affect, which pose a major analysis challenge for users with limited computer literacy.

The final step is to find statistically significant differences from the reference, often referred to as variant or copy number variation calling. Longer reads allow to detect larger events, such as copy number variations with higher resolution and accuracy, while shorter reads today result in a higher number of fragments supporting a single genetic locations variant status.

2.2.2 *Whole-genome & Whole-exome sequencing*

Whole genome sequencing, as the name suggests, allows to sequence the complete genome of an individual. The size of the human reference genome is around 3 billion nucleotides, and only for a small fraction of it the biological function is known. While the complete genome of an individual can be sequenced, due the current limitations, only a hand full or dozens of fragments can be sequenced per site. This is very low compared to the thousand or millions of cells that are present in a sample. This can be overcome by performing and merging multiple whole genome runs, but has to be traded-off against the associated cost.

For this reason researchers often chose whole-exome sequencing, which amplifies the exome of a sample, hence shifting the probabilistic distribution in favor of genetically active genetic sites. This results in higher coverage (many reads supporting a site) in exonic sites and virtually no coverage of intergenic regions. A special case of whole exome sequencing is targeted and panel sequencing where only very few targets of interest are sequenced with the benefit of extensive coverage (often 3 magnitudes larger than whole exome sequencing).

This targeted sequencing approach is used to detect variants at very low frequencies, or perform genetic testing on known disease-associated genes. WES provides a trade-off between unbiased detection in all genetically relevant (or interpretable) regions and the cost for a comprehensive WGS.

Using the NGS results of multiple samples allows to detect variants in each sample individually and to compare changes in the allele distributions between samples. Using this enables to detect tumor-specific (somatic) variants, inherited variants (germline), acquired variants of the germline (de-novo) and an estimate of copy number variants in exonic regions.

2.3 FROM SAMPLE TO VARIANT

The complete workflow from sample to a list of variants is depicted in Figure 2.2.

First DNA is extracted from a sample, which can come from solid or liquid tissue (e.g. blood). Second the DNA is sheared and denaturated into smaller, single strand pieces, allowing the material to be amplified in the third phase (if necessary). Depending on the technology used adapters may be added to DNA fragments to later support the sequencing or allow multiple samples to be multiplexed during a single sequencing run. The result of the sequencing process is a list of unsorted reads that are subsequently aligned to a reference genome. This alignment then allows to analyze the samples and identify statistically relevant variants within a sample, possibly taking into consideration the allele distribution in other samples (e.g. de-novo, germline, somatic calling). The de-facto standard file format for lists of variants is the Variant Call Format (developed by the 1000 genomes project)

(VCF) . It contains dynamic attributes for variant and sample specific information and allows to describe small and large genetic changes making it very versatile.

Genomic scientists will then work with these VCF files, detect commonalities, differences, systematic similarities or simply check for mutations, which have been previously linked to a disease. The results of this *variant discovery* then aids in decision making, be it in clinical or research setting.

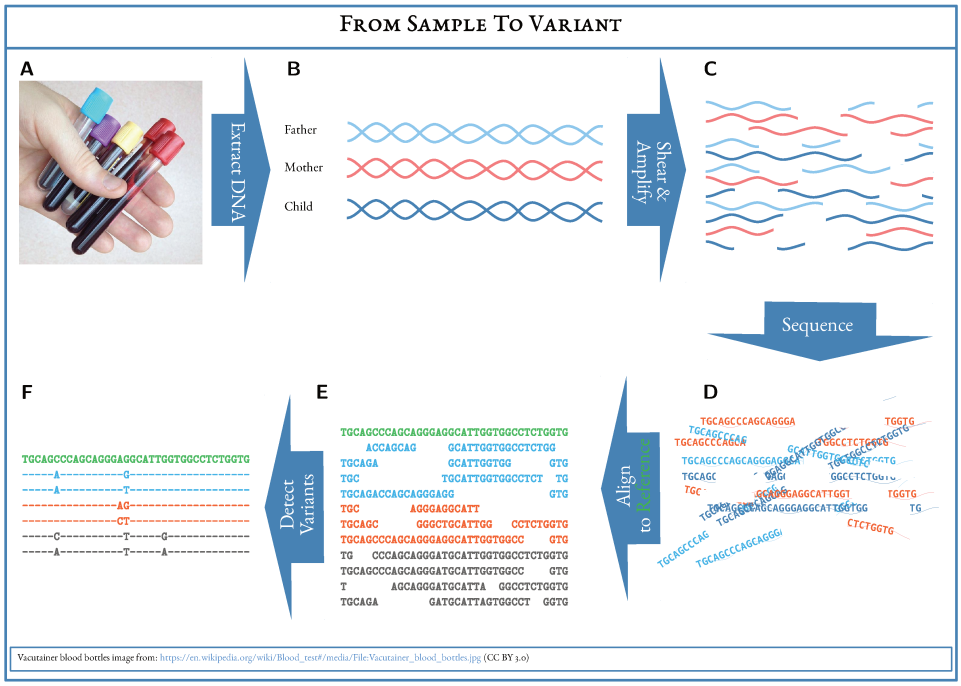


Figure 2.2: Simplified workflow of how biological samples are sequenced and variants are identified. (A) A biological specimen is taken. (B) The DNA of several samples (in this care a family trio) is extracted. (C) The DNA is sheared (split) and DNA fragments are put into the sequencing process. (D) After sequencing an unsorted list of DNA reads is available for further in-silico processing. (E) These reads from DNA fragments need to be aligned to a reference genome. (F) Based on the alignment variants can be called in comparison to the reference and other reference samples (child vs mother & father).

3

VARIANT ANNOTATION

Variant annotation describes the augmentation and enrichment of variants with relevant information from several resources. It is fundamental for the degree to which the impact of variants on a biological function or organism can be described and interpreted.

3.1 ANNOTATION RESOURCES

A wide range of biological databases are publicly available to researchers. These cover primary genetic information, such as genetic location of genes, and secondary information that is derived from primary data through other methods (e.g. experimental, such as gene expression profiles or computational, such as conservation scores).

3.1.1 *Transcript sets & gene annotation*

Large institutional databases, such as Ensembl[a] and National Center for Biotechnology Information (NCBI) [b] provide rich state-of-the-art information about genome assembly and annotation. This includes the GENCODE [45] and RefSeq [46] transcript sets, which are widely accepted and used as standard transcript sets. They provide information about the genomic location of genes, their transcripts[c], their exons and proteins. Using this information the consequence of a nucleotide exchange on the protein product can be calculated when an in-silico translation is performed. Because different codons encode the same amino acid in the final gene product, it is possible for a single nucleotide variant to have no direct affect on the amino acid composition. These are called synonymous variants, compared to missense variants, which result in a different amino acid.

Futhermore these databases allow to translate between the absolute genomic location of a variant and the possibly relative Human Genome Variation Society (HGSV) nomenclature [47].

In general, not all amino acids of a protein have the same relevance for the protein function. So called proteins domains, are parts of a protein that support a specific protein function. Proteins can contain different, often independent protein do-

a www.ensembl.org

b www.ncbi.nlm.nih.gov

c transcripts are the product of the splice process, and one gene can possibly encode many different alternatively spliced transcripts. The transcript most likely to be the protein coding transcript is called canonical transcript

mains. Databases that contain these and similar information about the functional information are for example PFAM[48], SMART[49] and Prosite[50].

3.1.2 *Variant databases*

The dbSNP database[51] is a public resource to report genetic variations from multiple organisms. It is a repository that is used for a wide range of applications and strives to provide a catalog of previously observed variants. Large population sequencing projects such as 1000 genomes[3] have provided millions of variations to this catalog.

Furthermore, population sequencing projects, such as 1000 genomes, ExAc[4] and UK10k project[2] provide information about the frequency at which a variant is observed in different human populations. These are valuable resources for human genetics because it enables researchers to exclude variants from their research which are frequent in humans.

3.1.3 *Disease associations*

Disease association resource combine variation databases with information about diseases and phenotypes, e.g. HGMD[52], COSMIC[53], TCGA[5], ClinVar[7] and OMIM[54].

COSMIC and TCGA focus on mutations in cancer and provide multiple measures per sample (e.g. gene expression, variation and other) from different platforms as well. HGMD, ClinVar and OMIM on the other hand focus on inherited diseases and phenotypes and differ mostly in the level of curation for reported variants and genes.

These databases also link to relevant clinical information such as drug resistance (COSMIC), detailed literature reviews (OMIM) or clinical evidence of pathogenicity (Clinvar). Such information is used to find co-located mutations and identify variants of significance.

The *Leiden Open Variantion Database*[6] is a project that publishes a catalog of disease specific specialized databases, where researchers can publish disease associations for genes and variants. This allows the creation of a comprehensive list.

3.1.4 *Conservation*

During evolution some parts of a genome are more frequently subject to changes than others. These mutations can possibly manifest through generations. When a gene or protein domain provides critical survival or fitness benefits, it is less likely to be altered between generations and related species. Conservation scores such as PhyloP[55], GERP[56] or phastCons[57] capture the evolutionary conservation and are used as indicators for the rate of conservation of genetic locations.

PhyloP considers single positions in the evolutionary development, while phastCons takes evolution of neighboring position into account and thus provides runs of conserved sites. GERP estimates the constraint that is put on a loci, derived from multiple alignments with other mammals.

3.1.5 *Functional context*

Missense variants that result in a amino acid exchange are the most abundant protein changing variant observed[39]. Yet it is clear that due to conservation, amino acid localization and other factors, not all of these amino acid mutations have a functional effect, let alone one that damages the protein function. *Loss of function prediction* tools such as PolyPhen[58] and SIFT[59] attempt to estimate the impact a missense variant has on the function of a protein. They use conservation scores (SIFT) or information about the protein structure (PolyPhen) to train models that allow the effect prediction of amino acid exchanges.

However, a substantial portion of protein altering mutations, such as those affecting splice sites, intronic or regulatory regions cannot be scored using these models. CADD[60] introduced a generic framework to score all positions of the human genome using synthetic variants to train their model. Although it is widely used even for variant interpretation guidelines and reviews[61,62], it's clinical validity has been questioned recently[63,64,65].

3.1.6 *The complexity of resources*

Peterson et al.[66] conducted a review of the available resources for the prediction of deleterious variants, including a comparison of publicly and commercially available variant databases. They found that *"The comparison of variant databases is a complex task due to differences in inclusion criteria, quality filters, amount and quality of annotation, and discrepancies in the reference sequences used"*[66] . They describe further details of the difficulties and details of the quanitative analysis of overlapping genes and variants. All in all a 62% recurrent disase-gene associations were found(2601 of 4164 genes recurrently associated [d]). Furthermore, they reviewed and compared 32 different in-silico tools that can be used to predict or grade the functional impact of a variant. Serveral conservation, loss-of-function prediction, splice site and structural stability tools are listed, highlighting the vast range of tools available. They conclude that: *"translating the analysis of human variants performed in-silico into the clinical setting is still one of the main challenges towards the goal of precision medicine."*[66] .

Niroula and Vihinen[67] conducted a meta review of state of the art prediction tools that can be used to support variant interpretation. They list a total of 23 variant databases that support interpretation, including meta databases that link to

d derived from Figure 1A of Peterson et al.

even more resources such as LOVD[6]. To assess the effectiveness and performance of in-silico tools for damage assessment, they compiled a list of eleven performance studies that overall tested the performance of 55 tools using different scenarios and datasets. They conclude that it is important to use computational tools for the are that they have been developed for. Thus they advice to use disease variant databases, such as Clinvar first before using other computational prediction tools.

Because so many tools exist for damage assessment, which themselves require multiple feature and training resources Liu et al. developed dbNSFP (database for nonsynonymous SNPs' functional predictions)[68] that compiles the result for multiple in-silico loss of function prediction tools for all possible missense variants of the exome. This provides a reproducible and integratable database for the various measures, which researchers can use.

3.2 ANNOTATION TOOLS

The numerous annotation resources, each possibly available in different versions with asynchronous update cycles makes it very difficult to provide accurate reproducible variant annotation. The different formats the resources are published in result in an integration challenge to access them in a unified way. Variant annotation tools aggregate different annotation resources and provide such a unified way to perform annotations on variants. They annotate VCF files with relevant information and are also part of the most clinical variant interpretation guidelines for cancers[62]. Additionally, these tools can provide databases that contain the integrated information and thus provide reproducibility of the annotation process.

A list of available variant annotation tools is available Table 3.1.

Table 3.1 lists available variant annotation tools. SNPEff[69] and Annovar[70] are the two most cited tools. They provide annotations from different sources and and organisms. 6 of 8 tools[e] provide methods to also filter annotated variants. However, not all tools are available for download and may require users to upload variants through a web interface. Such online tools are not feasible for automated analysis. subsection 4.2.2 will present how these tools are currently used by existing variant discovery applications.

e Variant Effect predictor was published in two articles

Name	No citations	Reference	Filter Capabilities	Resource Type
AnnoVar	2979	Wang et al. (2010)	Yes	Download/Online
Bystro	0	Kotlar et al. (2018)	Yes	Online
NGS-SNP	65	Grant et al. (2011)	Yes	Download
SeattleSeq	1065	Ng et al. (2009)	No	Online
SNPEff	1750	Cingolani et al. (2012)	Yes	Download
VARIANT	30	Medina et al. (2012)	No	Online
VAT (VAAST suite)	38	Habegger et al. (2012)	No (VAAST suite has)	Download
VEP (1st version)	823	McLaren et al. (2010)	Yes	Download/Online
VEP (updated version)	354	McLaren et al. (2016)	Yes	Download/Online

Table 3.1: List of variant annotation tools and their current number of citations.

Part II

STATE OF THE ART

4

VARIANT DISCOVERY

Querying and exploring variants should lead to the discovery of variants, which are relevant for genetic researchers.

The variant annotation tools (see Table 3.1) provide command line based filter scripts, to process and filter VCF files. Computer literate users can utilize these to filter lists of variants using the command line and the provided filter capabilities. However, these filter capabilities are limited to being used on the output of the respective tool.

Because of this, many tools were developed in the past that provide more general support for variant discovery for users. This chapter will give an overview, summary and categorization of existing tools to introduce the current software landscape currently available to researchers.

4.1 EXISTING APPLICATIONS

Several solutions for variant discovery tools exist(see Table 4.1), which can be divided into command-line-interface tools (CLI), graphical-user-interface tools (GUI) and web applications (WA). Command line tools are designed for experienced users, capable to work with a command line. GUI and WA are suitable for regular users, assuming that a user friendly interface is provided.

The target user group for CLI are bioinformaticians, who use these tools to process variant datasets as part of a larger workflows. These tools require expertise in basic programming and for users to manage datasets on their own. Examples for CLI tools are Var-MD, PriVar and GEMINI.

Var-MD[78] is a software designed to filter and rank variants for small projects, although discontinued in its development. It supported filtering by mendelian inheritance models and annotation using SeattleSeq[73]. PriVar[79] provides six static filter strategies to users and is able to rank the results based on the a logistic regression of functional impact predictors. GEMINI[80] allows researchers to query variants and their annotations from VCF files using Structured Query Language (SQL) expressions. It is designed to be used as command line tool, but also provides a simple web interface to send SQL queries to the database interface.

GUI tools have a graphical interface and thus are more suitable for users, who do not have extensive programming skills. They are more user friendly compared to CLI tools, and are designed to be installed as programs on a local work station computer.

SVA, VarSifter, myVCF, VCF-Explorer, VCF.filter, exomeSuite and VarAFT provide such interfaces. SVA[81], provides an interface to filter by variant consequence on the protein product. VarSifter[82] allows users to filter pre-annotated VCF files and visualize the results. Similarly, VCF-Explorer[83] allows users to filter variants from VCF files, but does not enrich the datasets with additional information (such as associated genes etc.). myVCF[84] provides a GUI through a local web server instance that stores its data in a local SQLite database. It does not perform annotation, but provides methods to make use of pre-annotated variants. Similarly, VCF-Miner[85] enables users to filter pre-annotated VCF files. The software exomeSuite[86] supports to filter input files by inheritance models and genes of interest. Afterwards annotation is performed enriching the result using several population and loss-of-function prediction tools. VarAFT[87] is a desktop tool that allows annotation and filtering on variants from VCF files. It features annotations that are performed on the datasets on the users desktop environment.

Web applications(WA), support a server-client structure by design. Here data is stored on a server and multiple users can interact with the data using a modern web browser, which makes these interfaces platform independent.

BierApp, wKGGSeq, VariantDB, Var2GO, VarApp and Mendel,MD are examples for web applications that can be used to annotate and filter variants.

BierApp[88] allows users to upload variants in VCF files and provides filters to take additional information such as inheritance into account. BierApp offers relational and non-relational (so called noSQL) databases to store variants.

wKGGSeq[89] is a web interface to KGGSeq[90], which provides a framework to perform pre-defined filtration and prioritization for whole exome data using a custom annotation and protein consequence prediction pipeline. This capability is utilized in wKGGSeq and results are presented to users in a user friendly fashion.

VariantDB[91] allows users to upload variants, which are then annotated using various resources and two variant annotation tools (Annovar and SnpEff). It allows users to customize queries and the annotations are displayed in the output.

Var2GO[92], allows to upload VCF files and provides an interface to filter the annotated variants. Although VCF defines how variants position and exchange are encoded, users have to specify this information as part of the upload. A new database is created for each dataset that is uploaded, including optional variant annotation.

VarApp[93] is a web application that makes use of reactive user interaction technologies that allows quicker updates to the web interface than common request-response patterns in most web applications. Utilizing this, users can upload VCF datasets and the application uses VEP[77] to annotate and GEMINI[80] to filter variants by user defined criteria.

Mendel,MD[94] is a web application that gives its users the ability to query, possibly multiple VCF files and provides parallel annotation. It is build on a relation

database and allows users to filter variants by their own criteria or filter them based on different modes of inheritance.

An advantage of web application architecture is that required software and data source do not have to be set up or maintained by users and heavy computations are performed on adequate hardware. However, when a web application is only available online and not individually installable (e.g. EVA[95], Annotate-it[34], wKG-GSeq[89], Var2GO[92]), users are required to expose private genetic data to foreign site. Because of the legal implications, such tools should not be used for genetic clinical patient data.

Maintenance and creating a sustainable framework and environment is a challenge for every software product. Consequently, 7 out of 28 listed tools (25%) are not available anymore, even after further research into finding a working copy.

Reference	Title	Year	No. citations	Available?	Type
Ge et al. [81]	SVA: software for annotating and visualizing sequenced human genomes.	2011	41	No	GUI
Sifrim et al. [95]	Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease.	2012	24	No	Online
Sincan et al. [78]	VAR-MD: a tool to analyze whole exome-genome variants in small human pedigrees with mendelian inheritance.	2012	20	Discontinued	CLI
Teer et al. [82]	VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer.	2012	85	Yes	GUI
Preston et al. [96]	VarB: a variation browsing and analysis tool for variants derived from next-generation sequencing data	2012	7	No	GUI
San Lucas et al. [97]	Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools	2012	56	Yes	CLI
Vuong et al. [98]	AVIA: an interactive web-server for annotation, visualization and impact analysis of genomic variations	2012	2	Yes	Online
Coutant et al. [34]	EVA: Exome Variation Analyzer, an efficient and versatile tool for filtering strategies in medical genomics	2012	11	No	Online
Paila et al. [80]	GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations	2013	116	Yes	CLI
Zhang et al. [79]	PriVar: a toolkit for prioritizing SNVs and indels from next-generation sequencing data	2013	16	Yes	CLI
D'Antonio et al. [99]	WEP: a high-performance analysis pipeline for whole-exome data	2013	28	Yes	Web application
Na et al. [100]	AnsNGS: An Annotation System to Sequence Variations of Next Generation Sequencing Data for Disease-Related Phenotypes	2013	2	No	Web application
Yao et al. [101]	FamAnn: an automated variant annotation pipeline to facilitate target discovery for family-based sequencing studies	2014	5	Yes	CLI
Alemán et al. [88]	A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies	2014	20	Yes	Web application

Reference	Title	Year	No. citations	Available?	Type
Maranhao et al. [86]	exomeSuite: Whole exome sequence variant filtering tool for rapid identification of putative disease causing SNVs/indels	2014	14	No	CLI
Vandeweyer et al. [91]	VariantDB: a flexible annotation and filtering portal for next generation sequencing data	2014	19	Yes	Web application
Li et al. [89]	wKGGSeq: A Comprehensive Strategy-Based and Disease-Targeted Online Framework to Facilitate Exome Sequencing Studies of Inherited Disorders	2015	5	Yes	Web application
Hart et al. [85]	VCF-Miner: GUI-based application for mining variants and annotations stored in VCF files.	2016	13	Yes	GUI
Granata et al. [92]	Var2GO: a web-based tool for gene variants selection	2016	3	Yes	Web application
Delafontaine et al. [93]	Varapp: A reactive web-application for variants filtering	2016	0	Yes	Web application
Salatino and Ramraj [102]	BrowseVCF: a web-based application and workflow to quickly prioritize disease-causative variants in VCF files	2016	4	Yes	CLI/Web application
Pietrelli et al. [84]	myVCF: a desktop application for high-throughput mutations data management	2017	0	Yes	GUI
G. C. C. L. Cardenas et al. [94]	Mendel,MD: A user-friendly open-source web tool for analyzing WES and WGS in the diagnosis of patients with Mendelian disorders	2017	2	Yes	CLI/Web application
Akgün et al. [83]	VCF-Explorer: filtering and analysing whole genome VCF files	2017	0	Yes	GUI
Müller et al. [103]	VCF.Filter: interactive prioritization of disease-linked genetic variants from sequencing data	2017	1	Yes	GUI
Desvignes et al. [87]	VarAFT: a variant annotation and filtration system for human next generation sequencing data	2018	3	Yes	GUI

Table 4.1: A list of existing tools to filter variant datasets and their usage in literature. Number of citations were obtained using *rcrossref* [104](Oct. 2018).

4.1.1 *Utilization of existing tools*

Looking at the citation records of existing solutions (see Table 4.1) one can see that a total of 497 citations refer to the mentioned variant discovery tools. This number indicates a low of usage when compared to the more than 6900 articles found through Pubmed including the exact term *"whole exome sequencing"* in title or abstract[a]. It suggests that most whole exome variant analysis are based on custom solutions or pipelines, possibly linked to the filter capabilities of the variant annotation tools which are cited more frequently (see Table 3.1).

a Number from Oct. 2018

The most cited variant discovery tool is GEMINI[80] with 116 citations. GEMINI is a command line tool and thus can be expected to be used by computer literate users as part of workflows or in the case of VarApp[93] as the database of choice to store variants.

The second most cited tool is VarSifter (85 citations), a GUI tool that allows to filter VCF files by genomic coordinates, gene name, consequence or information from the INFO field of a VCF . In general, such tools are a user friendly way to filter variant datasets. However, VarSifter is only able to load one VCF file at a time. This means that is has limited usability when large studies, with multiple individuals are sequenced as part of an institutionalized sequencing effort. The fact that it relies on pre-annotated datasets means that an even higher burden is put on users to manage these, possibly complex annotation tools as well, with all the implications it has for maintainability and feasibility.

The third most cited tool is "variant tools"[97] (56 citations), again a command line tool that can be expected to be utilized by computer experts. The output of this tool allows its users to generate reports in VCF format, which is not suitable for end users and thus requires post-processing.

Together these three tools make up for 51% of all variant discovery tool citations to date. However, for the reasons mentioned above, none of these tools provide a suitable environment for users with low computer literacy.

4.2 FEATURES OF EXISTING SOLUTIONS

The application and feature range of existing solutions is large, ranging from line-by-line filtering[83] to more complex technological applications[93]. To get a better overview I will present the recurring features and how existing solutions implement them. A tabluar overview of the existing solution and features can be found in table 10.5 as part of the results.

4.2.1 *Data management*

Generally two types of data management can be identified: First local storage in singular files, and second a centralized database structure.

Local storage usually makes use of indexes to access datasets more rapidly, either using custom solutions[105] or building onto existing ones (e.g. using indexes of local database solutions[80,84,97,102], such as SQLite). It has been recently proposed that index-free solutions may be better suited when it comes to large whole-genome datasets[83]. Restriction of this approach are the limited possibilities to share work between users and machines, which become a bottleneck when many users work with many datasets as part of a large cooperative studies.

Centralized database structures, when combined with a user authorization and authentication system are able to overcome this problem. Examples for these solu-

tions may use relational databases[b] or non-relational databases [c]. Some tools also offer additional user management[d].

Another difference between the data management approaches is how and which information is stored as part of the variant discovery process. This affects two types of data: One is the variant genotypes and the other is the variant annotation, necessary for the filtering. Some solutions require previously annotated input[e], others provide an on-the-fly annotation[92,100], and a third option stores annotations for subsequent filtering[87,91,93].

Variant and genotypes can be stored in a database, when multiple input files from various projects need to be integrated. Applications supporting this feature are EVA[34], Annotate-it[95] and VariantDB[91].

Not all tools require data management features, for example VCF-Explorer[83], SVA[81], FamAnn[101], VarSifter[82], VCF-Miner[85] or Var-MD[78] process input files and store the results without persisting additional data.

4.2.2 *Annotation*

Annotation and data augmentation are crucial parts in the variant discovery and interpretation process[62] because the ability to identify relevant variants depends on the features that can be derived from these annotations. The sources for variant annotation differ, but the most popular variant annotation tools are AnnoVar, SNPEff, SeattleSeq and VEP (see Table 3.1). Consequently these annotation tools are commonly used by variant discovery tools to gather information about variants, either by integrating an automated annotation or by using pre-annotated input files. Annovar is used by 8[f] ,SNPEff used by 5[g] and VEP used by 7 [h] tools. Interestingly SeattleSeq[73], although highly cited is only used as an annotation resource for one tool[78]. SeattleSeq data cannot be downloaded and does not provide a programmatic interface to facilitate its capabilities in an automated or integrated fashion.

Out of 14 tools that use variant annotation tools as a resource for unified annotations, only eight integrate these tools while the others rely on users to pre-annotate the datasets they want to analyze.

---

b Coutant et al., Li et al., Vandeweyer et al., Granata et al., Delafontaine et al., Sifrim et al., San Lucas et al., D'Antonio et al.

c Hart et al., Alemán et al., Salatino and Ramraj, Lopez et al.

d Coutant et al., Vandeweyer et al., Delafontaine et al., Sifrim et al.

e Sincan et al., Paila et al., Teer et al., Pietrelli et al., San Lucas et al., Yao et al., Salatino and Ramraj

f Coutant et al., Pietrelli et al., Desvignes et al., Vandeweyer et al., Granata et al., San Lucas et al., Vuong et al., D'Antonio et al.

g Paila et al., Vandeweyer et al., Granata et al., Yao et al., Müller et al.

h Coutant et al., Paila et al., Pietrelli et al., Delafontaine et al., Yao et al., Salatino and Ramraj, Müller et al.

Results of protein consequence prediction depends on the transcript set that is used[107]. RefSeq is referenced by 6 tools[i], the Ensembl transcript set is referenced by 4 tools[j] as well.

Population frequency data is available through 17 variant filtering tools published after the release of the 1000 genomes project data, highlighting its significance for the variant discovery process.

Similarly, loss-of-function prediction tools are also widely used. PolyPhen2[58] and SIFT[59] are among the most well established tools for this task, with 15 tools mentioning them as part of their filtering features.

Clinical and disease related information from ClinVar and OMIM are available for 9 solutions, only VarAFT supporting both.

4.2.3 *Query*

The presented tools support genomic scientists to varying degrees in their ability to find relevant variations.

Tools such as VarSifter, VarB, VCF-Explorer and VCF.Filter [k] are designed to explore VCF datasets in an easy way. This is useful, when users need to investigate only a limited set of locations or genes e.g. for validation sequencing. However, it does not enable users to investigate variants in many hundreds of samples and genes that are expected from whole-exome sequencing.

Nine existing solutions provide pre-defined query strategies, which allow users to filter datasets by a set of implemented rules, possibly parameterizing thresholds. This strategy is helpful, when samples need to be analyzed in-breadth and in the same way repeatedly. While these tools allow a broader view on a sample compared to explorative tools, only 3 of them allow users to also query based on user-defined criteria, which is necessary to allow unbiased and dynamic hypothesis testing (e.g. in case pre-defined rules do not yield results). This is especially relevant for research settings, where, by definition, new and unexpected findings are of interest.

13 variant discovery tools allow users to dynamically query datasets by defining feature of interest. This allows users to remove or retain variants based on genomic properties (e.g. gene name). However, only 4 of these tools provide an integration of variant annotation tools and other resources. All others rely on the user to perform the necessary augmentation of the datasets beforehand. Because the results of this filter process depends on the specific feature versions used, it is fundamental for reproducibility to provide such integration.

Only two tools are available to query multiple samples at once. Unfortunately, Annotate-it[95] is not available as an online resource anymore and the only other tool with this feature is the command line tool 'variant tools'San Lucas et al..

---

i Coutant et al., Ge et al., Desvignes et al., Vandeweyer et al., Vuong et al., Na et al.

j Desvignes et al., Alemán et al., Vandeweyer et al., Sifrim et al.

k Teer et al., Akgün et al., Preston et al., Müller et al.

Most tools empower users to query datasets by genes, consequence[l] and inheritance. These query features are fundamental to identify relevant variants. Among the most common other query features is the support to remove variants based on population frequency and loss-of-function predictions.

One example, of the abundance of variant features that users are faces with is given by VariantDB[91] which provides 105 different features per variant. Some of the features, such as gene names are easily comprehensible for end-users, more advanced properties such as conservation or computational prediction scores are not. No existing solution gives standard default cutoffs that users can rely upon, in order to set appropriate filters. Such defaults are necessary because computational predictions can yield arbitrary values, scales and thresholds that are not common knowledge among regular users at all. Thus it becomes impossible for them to perform abstract queries, such as *"only show variants in conserved regions"* when existing solutions require to filter variants by different unintuitive thresholds.

Remarkably, only VarAFT[87] enables users to work with copy number variations that can be derived from NGS datasets and provide important information to investigators. For all other tools, this means that user need to switch the annotation and query platform, if they need to analyze copy number variations derived from the same raw NGS data.

4.2.4 *System implementation*

Four existing solutions are only available online as web sites, which requires user to upload genetic patient data to a foreign site. A majority of 15 tools choose a local system architecture, targeting single users and smaller datasets to be investigated. Seven tools are database driven, although GEMINI, the most frequently cited one, uses local database files to transform VCF files into a relational structure.

Five tools implement features that allow collaboration between users in their software, while none of them allows multiple institutions to work on shared datasets. Only four existing solutions provide users with quality control measures that allow basic quality controls and only three give users the possibility to document uploaded datasets with comprehensive meta information. A lack of a comprehensive data management is detrimental to build large databases and manage large cohorts or studies.

4.2.5 *A note on commercial solutions*

This review leaves out existing commercial solutions that are provided by manufacturers and service providers (e.g. Illumina Basespace, Quiagen Ingenuity or Strand-NGS). Research facilities have to follow scientific rules of transparency, which commercial solutions due their competitive nature do not offer. Furthermore, it hinders

l The consequence is the predicted consequence a variant has on the protein biosynthesis process.

reproducibility when groups around the world need to purchase licenses, especially with regards to high license costs for developing countries. The work therefore focuses on tools that are readily available to the research community and where good scientific practice can be achieved.

4.3 INTRODUCING SNUPY

In this work I developed Single NUcleotide PolYmorphism platform (SNuPy) to address the short-comings of existing solutions and to meet the requirements that a comprehensive variant discovery platform has to meet (see Part iii). Several aspects are either missing in current solutions, or are only available sporadically or by combining several tools - which is not feasible for the target user group of genomic scientists.

SNuPy integrates commonly used variant annotation tools, databases and consequence predictions, allowing users to reach informed decisions about the impact and significance of variants. This approach is supported by the possibility to collaborate with other users making it possible to find consensus in their interpretation. A comprehensive data management approach allows to document datasets using controlled vocabulary, thus establishing the base for sustainable data foundation as more and more datasets are added.

Because of its user-friendly interface, SNuPy does not require users to combine tools in a complex tool chain in order to find the necessary features of variants.

Furthermore, several quality control measures allow the verification and validation of uploaded datasets, a feature necessary when thousands of samples are made available to users. Currently, no other tool offers such means to assess the correctness of variant datasets.

It allows to manage multiple projects, users and datasets independently from one another. A role-based user-authorization system enables fine-grained access permissions to the uploaded datasets, possibly sharing individual samples or full projects.

Pre-defined queries allow users to query all individuals in the database using the same query conditions and subsequently compare the results.

By collaboration, users can stay up-to-date in a fast moving research field, by sharing genes of interest with each other. Such gene of interest lists can represent biomarkers, pathways or known disease associated genes, compiled by experts and can be integrated into user-defined queries. Furthermore, standardized reporting of query results is possible for variant interpretation, druggable target and mendelian diseases.

VariantDB[91] allows to query variants through a web interface, but only retrieves 100 variants at a time, requiring multiple re-queries when more variants match the filters. SNuPy allows to query possibly thousands of samples through its web interface and retrieve large results sets, allowing users to investigate and follow-up

on these findings inside the web application, on external web sites, or exporting the results to spreadsheet processing programs for publication or further analysis.

Currently, to the best of my knowledge, no solution exists that provides an extendible or modular framework to integrate new annotation resources in an orderly fashion. Only GEMINI provides an API to their local SQLite database, allowing to send SQL statements to the engine through Python rather than through the command line. This lack of systematic architecture layout hinders the expansion of analysis features necessary for the adaptation to others fields of research.

SNuPy addresses this problem with the development of a framework (Annotation QUery and Aggregation framework (AQuA)) that allows programmers to add modules to the existing platform providing general or specialized annotation features for the variants in the database. User can access these annotations in a unified fashion, making the usability consistent and user-friendly. Developers set default values and group query conditions by their meaning, allowing users to focus on hypothesis testing rather than determining the correct threshold for an arbitrarily scaled feature. Furthermore, the aggregation layer allows to present variant features in an intuitive way using color coding and link-outs to additional information.

Part III

REQUIREMENTS

5

ANALYSIS REQUIREMENTS

To build a variant discovery system it is necessary to identify and clarify the user requirements, which need and can be addressed by such a system.

In general, next-generation sequencing technology can be used to study the genome of any organism. This work focuses on the usage of NGS technology to detect small genetic variants of genomic significance in humans and model organisms.

Genomics *"is defined as the study of genes and their functions, and related techniques"*[108] . This highlights the underlying complexity that genomic scientists face to explore and identify relevant genetic aberrations.

Human genomics in health aims to investigate the molecular causes, mechanisms and impact of genetic aberrations on diseases. Genomic researchers study these topics and focus on different levels of the protein biosynthesis process, thus a multitude of annotations are required to cover transcriptional, translational and regulatory impacts of variants. The low and dropping costs associated with NGS means that it is now feasible for research laboratories to sequence hundreds of samples, facing the computational and analytical challenge of this *"DNA deluge"*[32] . Consequently, a variant discovery platform needs to empower users to handle hundred or thousands of samples.

The scientific background of this work is based on research in pediatric oncology, hematology and immunology in a clinical setting. Because of this, there is a clear motivation to research the use of this technology in diagnostics[15,35,109,110,111], to determine possible treatment options[22] and to include medical doctors in the user group of genomic scientists. Tumor development, progression and even treatment options differ for pediatric and adult malignancies, but from an analytical standpoint the same requirements apply because the hypotheses tested do not depend on the age of a patient or sample. Age however is important for the interpretation of the results and the conclusion drawn from it, a task preferably performed by medical doctors and genomic scientists[36]. This requires a variant discovery platform to be accessible and usable by these user groups.

In addition to the clinical applications, research laboratories, such as the laboratory of the Clinic For Pediatric Oncology, Hematology And Clinical Immunology at the University of Düsseldorf, perform research e.g. on mice to study mutational landscapes and tumorigenesis[24,25]. The analysis requirements are the same for every diploid organism, but the availability of data resources differ greatly, depending on how comprehensively a model organism has been studied. Such research on multiple organisms needs to be supported by a variant discovery platform, requiring different annotation resource of varying complexity or completeness.

5.1 VARIANT FILTERING

In general, whole exome sequencing reveals between 20,000 and 80,000 variants per human individual[33,34,35]. A recent study of mice exomes revealed between 27.000 and 31.000 mutations in tumor and germline samples[25]. Showing that a whole-exome variant analysis platform needs to handle ten-thousands of variants per individual, each possibly sequenced multiple times from different tissues and disease states (e.g. diagnosis, remission, relapse etc.).

Any mutation that affects the function or abundance of a gene product directly or indirectly may be causative for a disease[35,112]. Protein-coding variants can be found in healthy individuals, even those where variant consequence can disrupt translation or transcriptional processes, such as by introducing stop codons, disrupting splice mechanisms or causing frameshifts during translation[113,114].

The study of more than 60,000 exomes from the ExAC project revealed that individuals carry 54 mutations on average that have been reported disease-causing[4].

In line with these findings is a study by Chen et al., who analyzed 589,306 healthy individuals and found individuals carrying disease-causing mutations for early-onset diseases. Although the study could not follow up on all identified individuals, due to missing recontact clauses and information, it indicates that incomplete penetrance may play a larger role than previously expected.

In order to narrow down the number of potentially disease-causing variants, a single patient sample is usually compared to control samples. These can either come from the individual (usually from germline/disease-free tissue), close relatives (parents, sibling etc.) or from large population studies such as the 1,000 genomes[3], ExAC project[4] or the UK10K project[2].

To narrow down the list of potential candidates, filtering steps from four categories (genetic scenario, variant properties, gene features and functional context) should be combined[35,61].

5.1.1 *Genetic scenarios*

Genetic scenarios consider the different genetic background and inheritance effects that may lead to a dysfunctional genomic environment, either by inheritance or spontaneous events (see Table 5.1).

	Description	Study requirements
Somatic	A disease-tissue specific mutation, not present in the germline of a patient	Germline control sample has to be available.
De novo	The mutation is not inherited by the parents, but is present in the germline.	Parents, possibly siblings and related family members
Germline	A mutation that is inherited from the parents.	Parents, possibly siblings and related family members
Autosomal dominant	A single variant allele causes a disease.	Parents, possibly siblings and related family members
Autosomal recessive	A homozygous mutation that causes a disease, which is heterozygous in both parents.	Parents, possibly siblings and related family members
Compound heterozygous	A patient inherits two defective copies of a gene, parents carry one functional copy.	Parents, possibly siblings and related family members
Gonosome-linked	A mutational burden that is imposed by the presence/absence of a X or Y chromsome.	Parents, possibly siblings and related family members
Single linkage	Only a single case can be identified within a family.	Parents, possibly siblings and related family members
Multiple linkage	Multiple members of the family are disease, possible spanning generations.	Parents, possibly siblings and related family members
Recurrent	Multiple patients carry the same mutation	Multiple patients with the same disease.
Known cause	Patients carry a mutation in a known disease gene	A list of disease associated, or relevant genes.
Full penetration	Carrying a genomic defect will result in the disease.	Family samples and clinical data
Partial penetration	Carriers of the mutation do not necessarily develop a disease, or may only show mild symptoms.	Family samples and clinical data

Table 5.1: A list of possible genetic scenarios that researchers may consider to identify disease causing mutations.

Some genetic scenarios require additional samples to be sequenced as part of the study, in order to test them. With this in mind it is clear that users need to query and compare multiple samples at the same time. This includes the usage of shared control samples that can be used to exclude variants of low significance.

5.1.2 *Variants, Genes and Functional Context Filters*

The properties of variants and genes are decisive when it comes to filtering variants and finding those that are of interest in the context of a hypothesis. A first step is to remove variants of low confidence[33,116] or those that violate the mendelian inheritance model[117a].

VARIANT QUALITY: Filtering by properties and measures of the variant detection process is a basic filter strategy to remove variations that are detected with little confidence[8,20,117]. These values depend on the sequencing process and are usually added during the variant detection steps as part of the variant calling pipeline. Filtering by these attributes during the variant discovery allows to adapt the quality criteria instead of hard filtering them. For example, a low confidence variant may still be regarded as significant and a candidate for further validation; if it meets other significant filter criteria (e.g. pathogenicity, disease-association . . .).

READ DEPTH AND TARGET ENRICHMENT: The sequencing read depth describes the number of DNA fragments that were identified by NGS in a genetic region, i.e. how often a position was observed in all fragments. A higher read depth gives more confidence in the variant and allows to identify low frequency variants[118]. The maximum possible read depth varies, but is limited by the fragments available as well as the sequencing technology used. When a whole genome is sequenced, the possible number of reads should ideally be evenly distribute across the whole genome. Whole exome sequencing on the other hand involves a targeted amplification step that elevates the sequencing probability in exonic areas of the genome, thus generating higher read depth in these areas. Which areas of the genome are amplified depends on the use case and manufacturer providing the amplification kit. They may include, additionally to the exonic regions, regulatory regions, microRNA sites, but may also exclude specific known genetic regions because of technical difficulties targeting these sites.

All of these factors play a role for the filtering[8,20], evaluation and interpretation of variants, thus adjustable read depth and capture region filters are an important requirement.

MENDELIAN FILTERS: The genotype of a variant is especially relevant when the mode of inheritance has to be determined for patients and to evaluate the impact

a Violation of the mendelian inheritance may also be of interest, for example when looking for de novo mutations that manifested spontaneously, so caution is necessary when removing variants based on these criteria.

of a mutation. Filtering by genotype is necessary for some inheritance models to be checked, thus it is required that users can filter by this property.

POPULATION BASED DATA: Population based variation data is used to filter variants that are common in the general population and thus are unlikely to be disease causing[35,65,117]. Several projects compiled datasets to assess the frequency of variants over large sets of individuals, namely e.g. the 1,000 genomes project, the exome variant server and ExAC are valuable resources[3,4]. Disease specific databases that record the variant frequency in tumors, such as the COSMIC database[53], are relevant for cancer studies. They are part of the recommended guidelines for the interpretation of sequence variants[38], highlighting their necessity for users to filter using these data resources.

CONSEQUENCE PREDICTION: A variants consequence is the predicted effect it has on the transcriptional and translational process. The consequence of a mutation can be predicted using transcript sets, which allow to perform an in-silico transcription using exon boundaries. Two well established transcript sets are available from GENCODE[45] and RefSeq[46]. Possible consequences include missense mutations that lead to an amino acid exchange, or the effect of a frameshift caused by an InDel . McCarthy et al. demonstrated the differences of the consequence predictions of different tools and transcript sets, thus multiple tools for the consequence prediction should be used such that different interpretations of the same variant can be presented to users[119].

GENE PANELS: To make hypothesis testing more consistent, checking for mutations in sets of genes is an important tool. Using such lists allow several users to query the same genes of interest, which may be compiled from pathways, known disease associations, publications or external databases. Ideally these lists are created by and shared between users and institutions, to facilitate reproducibility.

FUNCTIONAL IMPACT: A multitude of tools exist that allow the in-silico assessment of the probable functional impact of a variant[61,113]. But especially missense variants, the most common protein altering consequences, can be hard to interpret correctly unless structural, conservational and molecular factors are taken into account[113]. Loss-Of-Function-Prediction tools calculate scores to help users with the interpretation, although careful interpretation and follow-up analysis is required[64,65] (e.g. functional and other laboratory validation).

Other functional impacts may be identified from specific amino acid exchanges, such as losses of cysteine or lysine which are known to affect post translational modifications. Enabling users to identify and select these types of exchanges aids in hypothesis verification because unlike the output of loss-of-function-prediction tools, these effects are supported directly by molecular biological functions, rather than an numeric score.

FUNCTIONAL CONTEXT: One way to assess the functional context of multiple variants is to embed the mutations into the protein-protein interaction graph of

their protein products. This allows to display the functional landscape of multiple variants of interest and supports the interpretation process. Protein-protein interaction network integration has been used to describe new potential disease causing mechanism for autoimmune lymphoproliferative syndrome-like syndrome[120] and to identify a hub of proteins that may can be used as a potential biomarker for drug-resistance[23].

Thus, functional context integration is required as it allows to discover significant variants that impact protein function.

DISEASE ASSOCIATIONS: Variants, or the genes they affect, which were previously linked to a disease or phenotype are important for classification, even when the association is not to the same phenotype or disease (see Richards et al., Sukhai et al.). Databases that record variant-disease and gene-disease association, such as ClinVar[7] and OMIM[54] respectively, are valuable resources on inherited diseases.

INHERITANCE MODEL: When samples from family studies are available, selecting variants that match an inheritance model (as described in Table 5.1) is crucial. Some inheritance models, such as recessive inheritance of single nucleotide variants, can be checked directly by comparing the genotype at a single location for parents and their children. Other models, such as compound heterozygous defects can only be found when transcript sets are integrated & and taken into account.

SET OPERATIONS ON SAMPLES: Combining, intersecting and subtracting the variant sets of individuals is required to support basic set operations on samples. This allows an intuitive and straight forward way to e.g. to exclude variants from a control sample or find intersections between patients. Furthermore, when these capabilities are integrated into the filtering process it allows complex filter strategies to be implemented that are necessary to study complex disease progressions (see section 11.3 for an example).

INFORMED DECISIONS: Informed decisions are crucial for diagnostic applications and when researchers need to investigate every possible scenario. Consequently, the integration of multiple annotation sources from possibly multiple versions has to be supported so users can get the full picture and make informed decisions.

INTEGRATION INTO EXTERNAL WORKFLOWS Variant filtering and the ability to do so is not an isolated process, but is part of a larger workflow. To facilitate the integration of the variant discovery process into larger projects with multiple experiments, or complex computational tasks, it is essential that a variant discovery platform provides an external interface. Such interfaces allow integrated study of a multitude of next-generation-sequencing technologies.

BATCH QUERIES Variant interpretation most often requires dynamic filter strategies with user-defined parameters when datasets are investigated in a explorative fashion, especially when the phenotype is rare and the study is small. Larger stud-

ies that investigate e.g. a shared genotype among many individuals benefit from applying the same filtering logic to all datasets in order to have comparable results.

META ANALYSIS The decrease in sequencing costs lead to an increase of available sequencing datasets. Consequently, even small- and medium sized laboratories now have sequenced hundreds or thousands of samples. A variant discovery platform should allow users to harvest this data abundance and allow meta queries, that are performed on all available datasets. This requires an easy way to select the samples of interest. For example a user may be interested in all somatic variants of the second tumor from samples with a specific cancer subtype. When thousands of samples are available users have to have a user friendly way to automatically select the subset of datasets, that match such criteria. Furthermore, the variant discovery platform has to be able to query hundreds or thousands of samples that such a subset might include.

5.2 VARIANT INTERPRETATION

Most genetic disorders are rare[122], thus most genetic diseases have to be studied on a case-by-case basis, making interpretation crucial for understanding the genetic mechanisms[36,123].

In general, the categories to study genetic disorders can be classified into mendelian and somatic groups. Mendelian variants are explained by inheritance or their inheritance violation (e.g. by spontaneous events) and generally attempt to identify pathogenic *germline* variants present in all cells of the body. *Somatic* variants on the other hand, are variants that occur in the diseased tissue, but are not present or functional in germline tissue[124*b*].

A distinction between germline and somatic analysis scenarios has been proposed by Sukhai et al.: mendelian disorder variant interpretation aims to find *pathogenic* variants. In contrast somatic variant interpretation aims to identify *actionable* variants[121]. Pathogenic variants try to explain the phenotype, while actionable variants are direct or indirect targets of a drug for disease treatment[121].

5.2.1 *Germline Variants*

In an attempt to standardize variant interpretation for mendelian disorders the American College of Medical Genetics (ACMG), the Association for Molecular Pathology (AMP) and College of American Pathologists (CAP) have released joint guidelines to classify variants in the context of mendelian diseases. The guidelines consists of a catalog to assign variants into a five-tier system that rates their pathogenic-

b In general a mutation should occur in the diseased or non-diseased tissue. There are scenarios e.g. mosaicism, which are special cases of somatic and germline mutations.

ity[38]. It was later adopted by the UK based Association for Clinical Genetic Science (ACGS)[125].

The ACMG/AMP guideline defines 27 criteria and five rules, to combine the criteria into *pathogenic*, *likely pathogenic*, *benign*, *likely benign* and *uncertain significance* classes. It incorporates hard criteria from information that can be queried statically, such as population frequencies, predictive tools results or even disease association. Other criteria are based on soft evidence, such as the PVS1 criteria: *'null variant in a gene where [loss of function] is a known mechanism of disease'*[125]. Such soft criteria are subject to expert opinion and may vary between experts.

Biesecker and Harrison[126] have recently proposed to remove the criteria for a 'reputable source' from the ACMG/AMP variant interpretation schema. This criteria allows the integration of pathogenicity predictions of specialized databases into the interpretation schema. However, the authors of the interpretation guideline assume that the removal of this criteria is unlikely to impact the the variant classification[127]

The formalization of soft criteria has been focused on recently[128]. Abou Tayoun et al.[128] developed a decision tree that helps to streamline and standardize such interpretations. It gives researchers a better understanding how to apply the criteria. Nevertheless, Abou Tayoun et al. decision system requires 'biological relevant transcripts', which in turn will require further clarification and context-based interpretation by experts.

While soft criteria have been one source of interpretation variance, Ghosh et al.[129] focused on the interpretation of benign criteria BA1(*"Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium"*[38]) in the context of common disease alleles, but show incomplete penetrance (e.g. hemochromatosis)[129].

The Sequence Variant Interpretation Working Group (SVI WG) [c] of the clinical genome project (ClinGen)[130] is working towards a standardized variant interpretation and criteria refinements, both for disease and gene-level contexts that can be adapted in practice.

Another approach was developed by Nykamp et al.[131] with the Sherloc framework, a refined version of the ACMG-AMP guidelines that is based on a numeric scoring system for categories and criteria. Just as ACMG-AMP guidelines, Sherloc requires users to include soft data such as clinical reports and to evaluate functional experiments and databases, highlighting the significance of case-by-case interpretation.

c https://www.clinicalgenome.org/working-groups/sequence-variant-interpretation/

5.2.2 *Somatic Variants*

Similar attempts exist for somatic variant interpretation. Either based on gene panel targeted sequencing that relies on a pre-defined set of known disease-causing genes[121] or an unbiased whole-exome sequencing approach[62,132].

Sukhai et al. uses a 5-tier classification system that rates the actionability with regards to patient care. It uses information whether a variant is 'actionable' based on histology, recurrence in literature, previously reported information and predictive tools.

Van Allen et al. developed a rating system for variants resulting in 4 criteria: (1) potential clinical, (2) biological, and (3) pathway relevance, as well as (4) synonymous variants. In addition to this, the variants of interest are summarized in a structured annotation form, using five evidence levels in four categories (approved therapies, predictive therapies, prognostic and diagnostic) to generate standardized reports.

Li et al. developed *"A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists"*[62] that allows the categorization of malignant variants into four tiers based on their clinical significance and therapies approved by the U. S. Food and Drug Administration (FDA) .

Common to all methods and guideline approaches is that, the information that goes into the interpretation process are highly context and disease specific, thus users need access to all necessary information about variants.

The subjectiveness of variant interpretation was studied by Amendola et al., who compared the result of nine laboratory internal variant assessments to the standardized ACMG/AMP guidelines[37]. They report a 79% concordance for intra-laboratory variant assessment compared to the standardized guidelines, which was not significantly different between laboratories as well. However, when it came to the concordance between different laboratories, only 34% of variants were classified identically from all participating laboratories, which shows a high variance in the interpretation of variants.

Interestingly, this concordance rose to 71% after consensus discussion between the involved laboratories. This is in line with a previous clinical challenge study, by Brownstein and Others that identified a *"de-facto consensus of experts for interpretation of NGS"*[36] , demonstrating a general agreement among experts about the severity and role of variants.

This study highlights the necessity for genomic scientists to work together across laboratories and to be able to interpret the variants in a disease specific context[27,124]. This has been realized by other authors as well: *"Its(the variant interpretation - editors note) early steps are highly automated, but the final, most critical aspects are not. Instead, they rely on expert review and human interpretation."*[113]

Standardized reporting supports the streamlined interpretation of variants by experts. A variant discovery platform should therefore allow to present the results of a filtering process in comprehensive reports. These can provide a base for communication between users and external users.

To keep discordance between variant interpretation low, it is necessary to include consequence predictions based on different transcript sets to counteract the reported differences explained[119] and thus allowing the users to reach an informed decision. There is also a need to integrate multiple data sources, because the transcript consequence predictions alone are not sufficient to predict the actual consequence of a variant on the protein product[122].

The most comprehensive data resources are available for the human genome. Model organisms, such as mus musculus, also play an important role in the researchers ability to study genetic diseases as exemplified by Martin-Lorenzo et al. [25]. Thus it is necessary for a variant discovery platform to provide filter capabilities that do not depend on a specific organism.

In summary: A variant discovery platform should provide features to allow cooperation between laboratories. It should also let users select which criteria they think to be relevant and enable interpretation, possibly using refinement of queries to narrow down the mutational landscape of an organism to the significant variants. "*As a result of its complexity and impact on patient diagnosis and treatment, this process remains largely one of expert interpretation and literature review.*"[113] .

6

PLATFORM REQUIREMENTS

The primary goal of the variant discovery system is to enable genomic scientists to query and explore whole-exome datasets. The main focus is to reduce the large number of variants to the most relevant variants by attribute based filtering.

An information system, like a variant discovery platform developed for the setting described above, can be categorized in three parts[133]: (1) Memory - a function to store and represent data, (2) Active - a function to modify the current state of the memory, and (3) Informative - a function that provides information about the systems memory.

The memory function is implemented by a database and by the data representation of the variant data. The active function augments the variant datasets with the necessary information from external resources. In the context of this work the informative function allows to query and work with the variant datasets.

Parallel to these definitions, the platform requirements can be split into three categories:

1. Data management
2. Variant annotation
3. Variant filtering

There are three areas that need to be addressed in order to meet the above defined requirements for a variant discovery platform. First, primary data has to be added to the system, where it can be managed in a user-friendly fashion. Second, the system needs to integrate multiple external resources to augment the variant data and provide supporting features for the filtration and interpretation processes. Third, user have to be enabled to filter and explore the datasets in a user-friendly fashion, while developers need to be able to extend the capabilities and integrate more external resources.

6.1 DATA MANAGEMENT

As described above, variant filtering and interpretation are the crucial parts to leverage information in NGS datasets. With tens of thousands of variants per individual, possibly requiring multiple study and control individuals per study, it is clear that data management is a central task for a variant discovery platform.

Three aspects are relevant for this topic. First, a feasible underlying database management system has to be identified. Second, is to look at the requirements for

the meta-data which has to be stored and organized alongside the plain variant data. Third, will be to find the necessary annotations to add and assess the data resources and features which are provided to the users.

6.1.1 *Relational vs. noSQL Databases*

In recent years, so called noSQL database systems have been developed that allow key-value, graph or document oriented database systems. Such database systems promise higher performance, while breaking with traditional reliability concepts in relational databases such as ACID (Atomicity, Consistency, Isolation, Durability), in favor of more flexible concepts such as BASE (Basically Available, Soft state, Eventually consistent)[134,135].

NoSQL systems outperform traditional RDBMS systems when it comes to performance (insert, update, query) regularly[136,137], but not consistently[138].

Zeng et al. reported a performance advantage of noSQL over relational databases when comparing data insertion and retrieval of clinical patient dataset [a]. Due to the high number of attributes per record and the column count limit of the MySQL database system, they chose to split each record into multiple tables. Results for other data management approaches, such as using long lists instead of wide lists or using other normalization approaches are missing[136].

Schulz et al. demonstrated that for genotype data noSQL solutions provide higher insertion and better query performance, when building a SNP-locus database[137]. However data on advanced techniques, such as direct file import that can be used to increase the performance in SQL system are not part of the study. This highlights the necessity to use database-specific methods to increase performance.

Parker et al. showed that while noSQL solutions may outperform SQL systems when it comes to insert, update and delete operations, complex queries are better suited for SQL systems[139]. Such complex queries are expected to be the most common use in variant discovery systems.

These cases highlight an important aspect when considering traditional relational databases over noSQL solutions: Comparing the two systems solely on their performance is a complex task because the data models should be independently optimized towards the tasks the systems should be used for (also see Sahatqija et al., Lourenço et al.). A poorly designed model in one database system, may perform better in another. A key-value store is expected to show best performance when retrieving the available keys. Vice versa, a relational database should outperform other systems when it comes to relational operations.

Consequently, based on the performance benchmark by Cooper et al., the authors suggest that both traditional and noSQL solutions have their own benefits. For this reason, practical applications may chose to combine the two technologies

a The data analyzed does not include large variant datasets, but regular clinical patient data

and use them as tools for an integrated database system. The integration of SQL and noSQL database systems is the target of current research (e.g. Liao et al.).

To analyze the requirements it is therefore better to compare the two systems based on conceptual differences rather than benchmark performances, as exemplified by Mohamed et al.[135].

REQUIREMENTS FOR A DATABASE MANAGEMENT SYSTEM

While insertion of data records is important for dynamic systems that need to handle user generated input frequently, this performance measure is not meaningful for query systems that do not update data frequently. A variant discovery platform is such as system because data is added only once for each study, while multiple queries are performed afterwards.

When security is a concern, as it is for genetic data, RDBMS provide better overall security features[135].

Another difference between relational and non-relational databases are their flexibility when it comes to changes in the data models[134,139]. Relational databases require static database schemas, allowing well defined data models, while noSQL databases are more flexible. A variant discovery platform benefits from well defined data models, when standardized input formats are imported and features from possibly unstructured external sources have to be made available in a structured fashion.

Furthermore, relational databases have proven capable to store genotype data in a relational way[143,144]. These more traditional relational database systems can be considered the industry standard[139], thus they are considered more maintainable.

There is a also a known lack of a standardized query language for noSQL database systems. Vendors implement their own language or dialect, but it requires custom programming to perform join or aggregation operations on such systems[134], making them harder to develop and maintain.

To fulfill the requirements for a variant discovery platform, a relation database system is not the only, but the best maintainable, secure, mature and accessible solution. Existing solutions that use database backends also mostly use relational databases (most notably GEMINI), showing that feasible performance is possible.

6.1.2 *Sample Access & Meta Information*

Managing variant datasets has to include storage of meta-information about sample state and relevant context specific information that should be stored alongside the datasets to allow reproducibility and re-usability. Because the clinical and biological research field may focus on very different aspects of an individual, it is necessary that a multi-purpose variant discovery system provides a flexible way for users to integrate the information, which is necessary for them.

META INFORMATION

Organization of datasets and adding annotations necessary for documentation and meta-analysis is always an important aspect of data management. Any data system that allows cooperative research should also impose mandatory minimum informations about datasets. This makes reproducible research easier and supports the structuring of possible hundreds of datasets, eventually enabling inter-laboratory interpretation and cooperation. There are three reasons for this requirement. First, a basic form of documentation is necessary as a basis for collaboration between colleagues and institutions. Fluctuations in employees means that knowledge of samples and their background changes with them and might be hard to recover once knowledge-holders are not available anymore. Second, it is crucial for NGS studies to differentiate between case and control samples in order to build up groups of shared control samples, which can be confidently reused in other projects. Third, implementing mandatory minimum information enables automated analysis that can be based on a minimal set of available information.

The mandatory minimum information for each whole-exome sample are:

SAMPLE STATE - describes the disease state of the sample.

TISSUE - describes which tissue was used to retrieve the specimen.

SAMPLE CLASS - describes if a sample is a shared control sample, or can be categorized into a specific disease class (malignant, immune-deficiency, rare etc).

DISEASE - describes the disease or phenotype the sample presents.

ACCESS RESTRICTIONS

To support cooperation across laboratories, the system has to allow multiple colleagues and laboratories to access the relevant information and keep other datasets private.

According to De Capitani di Vimercati et al., an access control system should support eight features to be feasible. I will adapt these features to fit into the context of a variant discovery platform:

ACCOUNTABILITY AND RELIABLE INPUT - This is necessary to make sure that users are authorized to use the system. In a variant discovery system this should be as strict and stringent as possible, only allowing access to the system after authentication.

SUPPORT FOR FINE- AND COARSE-SPECIFICATION - If possible this makes sure that only specific actions can be granted access to. Given that multiple user groups access a variant discovery system, it is useful to define which user group, or user can perform which action.

CONDITIONAL AUTHORIZATIONS - To protect data, it is necessary to apply conditional authorization, based on the context of the action. For example the owner of a dataset may modify it, but others may only display the entity.

LEAST PRIVILEGE - This feature mandates that users should perform their actions with the least privilege necessary, to minimize the possible damage by errors.

SEPARATION OF DUTY - *"refers to the principle that no user should be given enough privilege to misuse the system on their own"*[145] . This concept refers to managing potential conflicts of interest and fraud. This is especially relevant when the system is used by different institutions, possibly within a competitive field. The most detrimental action would be deletion of datasets, which should be prevented for users and seen as an administrative action. Because users upload the data on their own, they can be expected to have a copy of the data, thus it would never be truly lost. This loss of information has to be seen in contrast to other information systems, e.g. sales-systems where customers can't be asked to redo their orders to recreate a database state. The same is true for altering the minimal documentation, required for datasets, which should only be editable by specific users or small administrative user groups. Another aspect is an overloading of the system with datasets or complex queries. Restrictions on the ability to upload datasets and limit query result complexity are therefore required.

MULTIPLE POLICIES AND EXCEPTIONS - usually policies can be divided into two classes: closed and open. Closed policies do not allow any action, unless authorization is present. Open policies allow every action, unless a rules prevent it. A variant discovery system should follow a closed policy to prevent any potential data misuse.

POLICY COMBINATION AND CONFLICT-RESOLUTION - This feature accepts that rule based system can be incomplete or inconsistent, but requires the access restriction system to handle such cases, e.g. with default behavior. To protect the datasets from misuse the default behavior in a variant discovery platform should always be to enter a safe state that does not allow access to the datasets, thus following the closed policy paradigm.

ADMINISTRATIVE POLICIES - This feature describes who is able to set the aforementioned policies and modify them. For a variant discovery system only privileged users should be able to modify coarse-grained access to datasets (e.g. *access to all samples of an institution*), but give users the ability to decide on the fine-grained access (e.g *a specific other user should be able to use this specific sample*).

6.1.3 *Variant Data*

Variant data can be stored in a sample-variant-genotype relational association[143,144].

However, before analyzing datasets users should perform a quality control step, to verify the correctness and validity of the samples, to prevent a 'garbage in, garbage out'[146] scenario.

GENOTYPE DATA

When storing genotype data in a relational database, several database normalizations are available from a database design point of view. Such normalizations aim to minimize redundant information and split up the attributes of an entity over several tables using foreign-key associations. These associations are merged during a query by *JOIN* operations to build or rebuild the attributes of an entity.

However, when working with large databases of genetic data, it has become best practice to abandon rigorous normalization in favor of performance increase, which is feasible for databases that are not regularly updated. Examples for reduced normalization (or denormalization) are the StringDB[147] and Ensembl database[148].

INPUT FILE FORMAT

The de-facto standard[93] format for variant datasets is the VCF format (described by Danecek and Others). This format uses absolute genomic coordinates to locate variants unambiguously. Coordinates in such an encoding can be translated into coding and protein nomenclature[b](e.g. HGVS nomenclature) more easily than the other way around, are mappable to other organisms and are the recommended way to handle variants for clinical interpretation[38,62]. Therefore, this format should be supported for data upload and management, possibly with the ability to handle tool-specific output and additions to the flexible format.

QUALITY CONTROL

Generating NGS datasets is a complex task, which involves a multitude of protocols and wet-lab work by different personal. Standard operating procedures are employed to prevent mistakes and these processes are carried out carefully to provide reproducibility and traceability. However, as the number of sequencing experiments increases, even with low error rates there is a statistical limit to when the first error occurs with significant certainty.

Quality measures of raw NGS data are mostly focused on the validation of the sequencing process and its success. However, variant data quality analysis is also necessary to gain confidence in the correctness of the pipeline, which is carried out to create the NGS datasets. For example, a sequencing run can meet all the quality

b A nomenclature where variant positions are given with respect to the gene, transcript or protein then affect

criteria, but if the sample was mistakingly swapped during wet-lab processing this would not stand out and may result in false reporting.

Several quality measures, such as variation call statistics, variant frequency and count distribution must be implemented to enable quality control. This should be carried out before any analysis to check the integrity of NGS process and datasets[146].

Many potential sources for errors arise during the process of sampling a biological specimen and the actual upload of the datasets into the variant discovery system. The main reason for this are human errors during handling of the specimen, incomplete or inaccurate documentation as well as errors during in-silico processing of the next-generation sequencing results. Sources of errors are:

SAMPLE SWAPS - A sample was mislabeled during any stage. This can happen during handling of the physical as well as the in-silico sample.

WRONG RELATIONSHIP ASSOCIATION - This is a special case of a sample swap, which can be recovered if relationship information and clinical information are available.

SAMPLE CONTAMINATION - impureness due to foreign genetic material can be caused by mistakes during wet-lab processing or erroneous data processing.

Validating candidates is a laborious time consuming task and a variant discovery platform should support users to identify sample contamination events before engaging in this work. Such events are not apparent from the query result that user work with and potentially lead to false reporting in publications or worse a misdiagnosis in a clinical setting.

Another prospect of quality control is to spot unusual samples, e.g. those that carry an exceptionally high number of variants or have exceptionally less variants on a specific chromosome. These information are used as a control measure, e.g. a female sample should not carry a significant number of mutations on the Y-chromosome. They may also serve as a control of expectation, for example when a patient received allogeneic stem cell transplantation, the mutational profile should significantly differ from a sample before transplantation.

Quality control can also lead to a phenotypic observation, e.g. a sample with a high mutation rate might suffer from a dysfunctional DNA-repair mechanism, which is especially relevant information in cancer studies and may have implications for treatment options.

Several measures are possible to e.g. track potential sample swaps, or detect unusual/unexpected results. Gender prediction and sample similarity measures allow to verify that the sequenced samples were not swapped by mistake. Variant frequency distribution plots can further help to check if a sample was potentially con-

taminated [c]. Sample swap and contamination indicators are necessary in a variant discovery platform so users can take appropriate actions to resolve such issues.

6.2 VARIANT ANNOTATION

Augmenting the variant datasets with relevant and comprehensive annotations is necessary to meet the analysis requirements, described in the previous chapter.

In general, there are two sources for annotations: primary and aggregated resources. Primary resources are databases with information about variants and genes, while aggregated resources provide multiple primary resources from a single location.

While primary resources provide the most up-to-date data, the aggregated resources usually provide a fixed collection of primary data that is updated regularly. Examples for such aggregated resources are Variant Effect Predictor(VEP)[77], Annovar[70], SNPEff[69] and dbNSFP[68], which provide rich variant annotations. Many primary resources, e.g. PolyPhen2[58], require a multitude of other primary data sources and the result may differ when versions differ. Using aggregated resources is therefore advantageous because of the increased maintainability and reproducibility.

Nevertheless, setting these tools up requires significant effort, to install the software dependencies, acquire and store the necessary data sources, as well as organize and maintain such an environment in a sustainable and manageable fashion. This becomes even more challenging, when different versions of data sources need to be handled. Overall, these tasks are not suitable for users with low computer literacy, such as genomic scientists, hence they are usually taken on by bioinformaticians.

A variant discovery platform should allow to integrate primary and aggregated resources and make them available to user. Furthermore, such as platform should be extendible to integrate new data resources in the future, to react to new developments and to allow flexible customization for different areas of research. To achieve this a unifying framework that provides developers the environment to integrate annotation resources is necessary. Such a framework implements common tasks, such as input data generation and annotation tracking allowing developers to focus on tool specific tasks.

---

c Here the term contamination is used to describe impureness of a sample on DNA level, which may be caused by e.g. biologically processes (e.g. in-vitro growth), medical treatment (e.g. transplants), accidental or unexpected mixing of diploid DNA samples of different individuals of the same organism. Contamination does however not refer to contaminations e.g. of hygiene that may refer to contamination with fungi, bacteria or viruses, or radioactive contamination.

6.3 VARIANT FILTERING

SEMANTIC QUERIES

As outlined before, variant filtering is fundamental when exploring variant datasets. Using external annotation resources means that users get a variety of possible filter options for their queries. The same annotation attribute by different annotation tools may have a different or a similar meaning for a query, putting a burden on users to identify the annotation feature name they need to use. One common example are *gene symbols*, which may be called *gene name*, *hgnc*, or simply *gene*.

When designing queries, users should be empowered to formulate them in a user friendly way and should not have to think in terms of feature names, but rather in terms of which variants they want to identify. For example, users should be presented with an input for a gene identifier and the query system should use the necessary feature names (e.g. symbol, gene name, HGNC, MGI, ...) to test against. I call this *abstract query* to emphasize that features, although they can have different names, can be semantically equivalent on an abstract level.

COMPLEX QUERIES

When a relational database is used, SQL is the standard query language to answer queries on these databases. These queries are complex and not user-friendly, therefore user defined conditions have to be translated into a SQL -query that can be executed and answered by the database system. However, some queries are very complex to formulate in a relational system, especially when filter conditions require self-joins to utilize multiple rows of the same table. Thus, variant filter methods should also provide means to filter query result sets programmatically, to ensure that such complex queries can be performed in an efficient, sustainable and non-relational fashion. One example for this type of complex query are searches for compound heterozygous mutations, which can only be executed with regards to transcript boundaries and family genotypes because multiple variants need to hit a single transcript.

PRE-DEFINED QUERIES

It is important for users to be able to test their hypotheses dynamically in an explorative way. However, to facilitate reproducibility and comparability it is also important to support standardized, pre-defined queries that filter a dataset with fixed parameters. When large studies contain hundreds of samples, such pre-defined query criteria provide a valuable starting point for a more detailed and individualized analysis or overview.

6.4 TECHNICAL REQUIREMENTS

Because a centralized database is required, encryption and authenticated communication is necessary when data is transfered over public networks.

When input data is transfered over a network, a validation of the data integrity should be performed to avoid data corruption. This is especially relevant when data is transfered through a strict and possibly intrusive firewall that clinical networks employ.

A quick turnaround time is necessary for an explorative variant discovery process. The complexity of the query is essential for the overall filter time. Displaying mutations in a single gene should be possible within a few seconds, whereas more complex queries involving many control samples and filter criteria may be expected to return results within minutes. Processing of raw data is expected to take longer because this step mostly depends on the external tools and their performance to annotate datasets.

Because of the unpredictable complexity of a query and the required runtime, it is necessary to handle long running processes. Such a feature allows users to send a query to a queuing system that decides which jobs to execute and stores the results to be retrieved later. Furthermore, due to potentially restrictive computer network policies in a clinic, the application should support usage through a firewall.

Because variant discovery is usually done alongside other biological experiments, features to export the datasets and relevant information about variant- and sample-annotation are required to integrate the platform into external workflows. The absence of external interfaces significantly obstructs the reuse in new platforms, e.g. SeattleSeq that although highly cited is only used by one variant discovery tool (see 4.2.2). The interface should support different response formats to allow flexible integration for future developments.

Part IV

CONCEPT

7

A VARIANT DISCOVERY PLATFORM

7.1 DATA AND SAMPLE MANAGEMENT

7.1.1 *Upload*

The de-facto file format to encode lists of variant is the VCF format. It defines a set of attributes and sample information that can be used to describe the quality, support, genotype and additional annotations for a variant. Each file may contain multiple samples that are the result of a variant calling process (e.g. genotypes in tumor and normal tissue, or genotypes from parents and their offspring).

The format is flexible, thus it is possible for the same information to be encoded in different ways and under different attribute names. For this reason, different parsers have to be supported to extract the relevant information from standard VCF files, or customized output by standard tools such as GATK[150], VarScan2[151] or EXCAVATOR[152].

Data is uploaded as individual files, possibly compressed using the well established block gzip format of tabix[153] or as a ZIP archive that contains multiple files and a meta-information file. The upload is based on the Hypertext Transfer Protocol (HTTP) protocol allowing programmatic access to integrate the software into workflows.

When the upload is successful, the variants are extracted into the database and a index is generated that allows developers to trace the variants in the database to the file location in the raw-data. Based on these information, the variants for individual samples are extracted into a table containing the genotype, as well as the minimum information provided by the VCF standard.

7.1.2 *Sample organization & meta information*

Sample data and meta information are organized in a hierarchical structure that represent varying degrees of abstraction. It is complemented by a dynamic tagging system, that allows each instance in the hierarchy to be annotated. The minimal information necessary to document variant datasets is realized using this annotation, with pre-defined tags. Tags provide a flexible way to add controlled vocabulary to the system without having to restricting users.

SAMPLE ANNOTATION SCHEMA

The hierarchical sample schema includes several levels of abstraction from a specific sample to the analysis of complete pedigrees.

Furthermore, a structured sample organization allows to compile focused studies from previous experiments and the establishment of a shared control sample pool. These features are fundamental when samples are queried as part of a meta-analysis.

The sample hierarchy (figure 7.1) allows to view a dataset on different levels of abstraction. The most abstract level are entity groups, which can be used in multiple projects, e.g. when they contain entities carrying multiple phenotypes of interest. Each entity group is comprised of at least one entity, which represents an individual (e.g. patient or cell line) that is part of a study group. Each entity is associated with at least one specimen, which represent the biological specimen (e.g. aliquot) that are sampled from one individual and for which next-generation-sequencing analysis is performed. Because several analysis pipelines may process the same specimen, each one is linked to several VCF files, containing samples from the different analysis workflows.

In practice,entity groups represent the smallest group of entities sufficient to perform variant discovery on an individual (e.g. patient or study subjects like mice). In case of complete family trios, parents and child are part of one entity group. In case of tumor-normal studies the entity group only contains one entity. This allows to compile new projects with different research questions from existing entity groups, while being flexible to be applied to a wide range of study setups.

A note on the terms specimen and sample: In an interdisciplinary field, such as bioinformatics, the same term may have different meaning. The term sample *is an example for this. For wet-lab applications this usually refers to a biological specimen that was sampled from an organism. For variant calling from a NGS pipeline it refers to the result of the analysis stored in a VCF file. The structured sample hierarchy uses the term specimen to describe the biological sample taken and the term sample in the way used by the VCF standard.*

TAGS

To classify the elements in the sample hierarchy (see Figure 7.1), controlled vocabulary in the form of tags are used. This is necessary to establish a unified terminology that can be used for automated analysis and support long-term documentation in a field where different users work with the same datasets over a long period of time. Tags are a flexible way to allow adaptation to new projects. User are able to add new tag categories, which adds new terminology making documentation of research project from different disciplines more feasible.

While the tags are dynamic and new tag categories can be added, some tags are mandatory to provide basic analysis functionality. I call these tags *mandatory minimal information*, that are necessary to perform the required variant filtering strategies (see Part iii). There are mandatory tags for entities, specimens, samples and

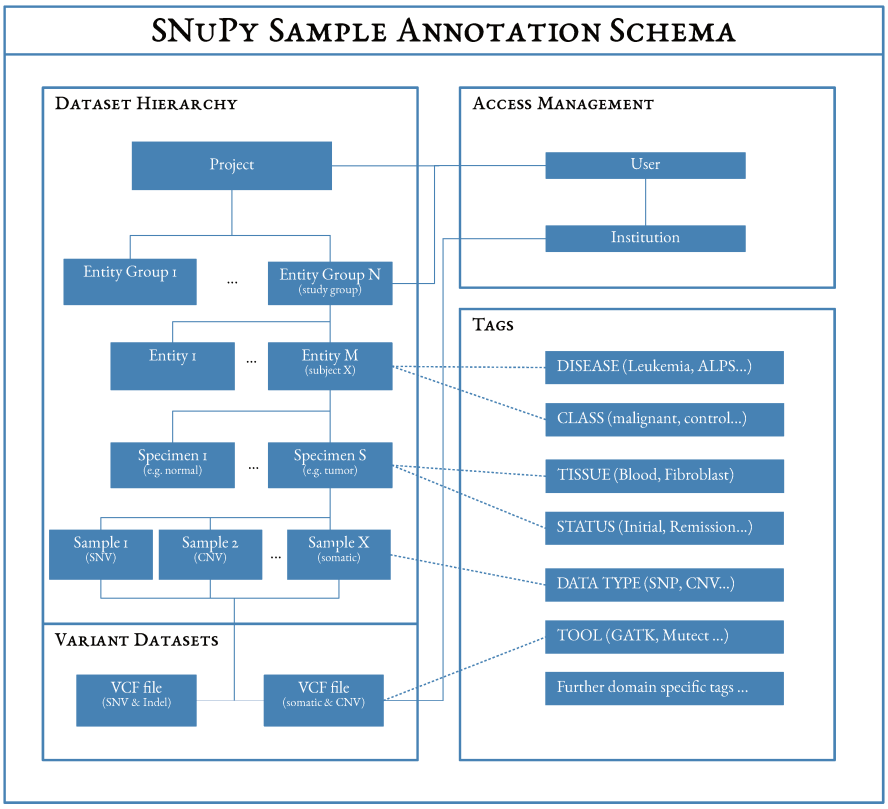


Figure 7.1: The sample annotation schema shown in this figure is used to organize and structure datasets in SNuPy. The hierarchical structure allows to change the abstraction level on which users view the mutational profile of a study. Access management controls access to entity groups and VCF files through affiliations of users to institutions to ensure that users are not allowed to view, use or modify datasets from alien institutions. Tags allow documentation with a customized and domain specific controlled vocabulary.

vcf files. First, entities are tagged with a class tag, which is used to divide datasets into cases and controls. Disease entities are required to carry a disease tag, based on Medical Subject Heading (MeSH) terms.

Second, specimens require tags for the state of the specimen (initial, remission etc.) and a tissue tag, which corresponds to a MeSH terms. The specimen status tag hierarchy (see figure 7.2) allows to represent complex disease progressions such as multi-stage tumor progressions. This helps to clarify which disease or health state a probe is associated to and unifies different terms. The naming follows a lexicographical order which improves the overview in complex analysis scenarios. The hierarchy is set up as follows: All control states are prefixed with *C*, thus they are lexicographically smaller than all the terms after the first initial diagnosis, prefixed with *I, R* or *T*. Each relapse adds a *R*-prefix and each transplant adds a *T*-prefix,

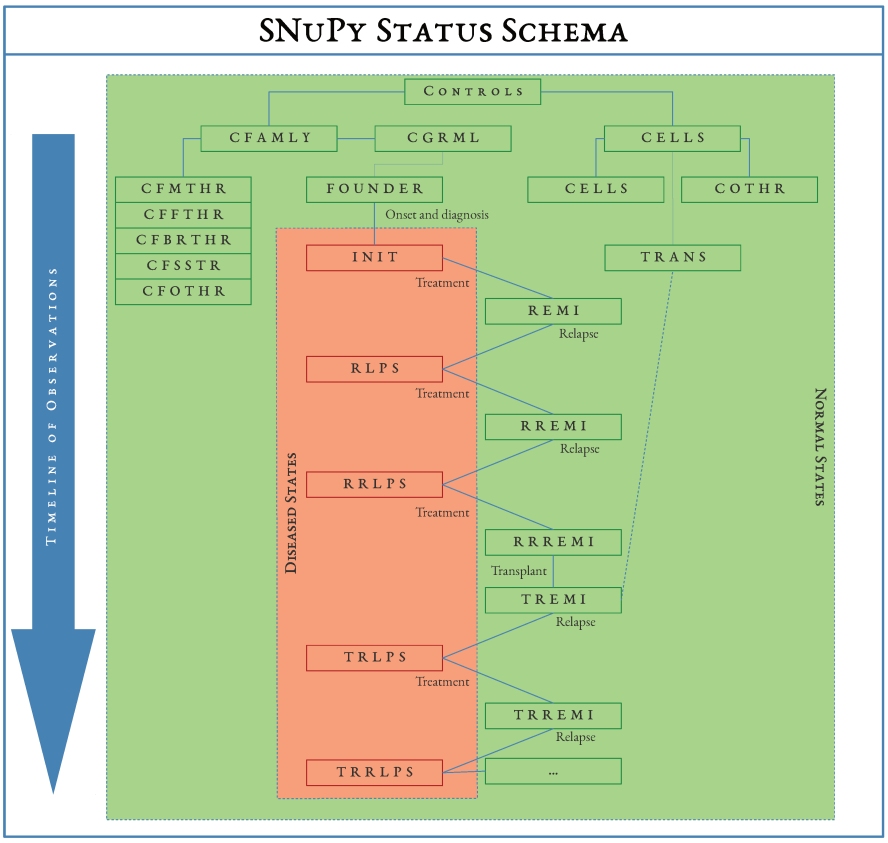


Figure 7.2: The specimen status schema in SNuPy is a compromise between readability, shortness and comprehensibility. The disease state progression can be flexibly extended to incorporate complex progressions, while following a lexicographical order that makes it easier for users to keep an overview.

so that the timeline of events keeps a lexicographical order through disease stages. Applying these rules leads to a human readable and convenient naming convention. The ordering helps to keep an overview over possibly many stages of complex disease progression.

Third, samples are tagged with a data type tag, such as somatic, germline, denovo, CNV and others to indicate what kind of analysis was carried out to produce the variant calls.

Forth, VCF files are tagged with the tool name, that was used to generate the data. This allows to document the version in the tag name as well.

Users are not able to query datasets if these information are not present, enforcing the minimal documentation.

Other tags are readily available too and provide valuable information for investigators, such as karyotype classes or anti-body presentation properties that distinguish cell states. Because of this, the application spectrum is flexible and can be used for many research field dealing with next generation sequencing data.

7.1.3 *Reports*

User need to be empowered to filter variants to find variants of significance. This is most often achieved through multiple query refinement iterations and context based variant interpretation. For this task users can export lists of variant in different formats (e.g. Excel or VCF format) to work on result lists in a offline fashion which is suitable for a practical solution.

Once users have identified variants of significance they have the option to mark variants of interest and generate a standardized summary report as a unified template to communicate their findings to others.

For this purpose developers utilize a reporting template engine to create Open Document Format files, that users can view and edit using standard text processing software (e.g. LibreOffice, Microsoft Word etc). The possibility to modify a report allows users to add smaller modifications (e.g. highlighting) by themselves if necessary, increasing productiveness.

After modifying the reports, users can upload the documents back into SNuPy to document their finding and keep documentation centralized.

7.1.4 *Quality controls*

The quality control measures in SNuPy are based on variants and data, which is generally available from VCF files based on the format reference[149].

Several measures are computed to support the discovery and possibly subsequent recovery of errors:

- Detecting sample swaps is a crucial step and the most basic way to do this is to look for coherent overlap between different specimens of the same patient or between family members. If the overlap of variants is too small a sample swap is possible and users can take action to verify this further.
- Another straight forward method is to check the variant patterns on the gonosomes to predict the gender of the sample, which can be used to verify if the result is coherent with the study documentation.
- Some sample contamination events can be detected on a variant basis, for example by investigating the variant frequency patterns of heterozygous mutations. In the event of a contamination of two individuals the frequency distribution becomes a mixed distribution, which becomes obvious in a distribution plot.
- Using basic statistics about the variant count distribution and the allele exchange rates also provide valuable information about the mutation rate.

All of these measures are made available to the users for investigation and to empower them to detect exceptional datasets.

For more information on the calculated measures see section 8.6.

7.1.5 *Role-based authorization system*

A role-based user management system is used as a manageable and flexible solution to the user management and access regulations. Each user has direct access to the datasets he is assigned to, while he can access other datasets based on institutional association and his role. This allows to share datasets in a fine and coarse grained fashion.

The access rules can be divided into four different actions and four different roles: The four actions are:

QUERY - Query datasets and filter their mutations by submitting a query.

MODIFY - Modification of a dataset and its meta-data. Only few members of project have the competence and authority to document and modify a dataset.

REVIEW - Use samples in a review process, that is used to filter rare mutations in the database or exclude possible sequencing artifacts. This implies read-access to the samples of a users institutions.

UPLOAD - Add new VCF files to the platform and start processes to augment the datasets.

These actions are used to categorize the abilities users have to interact with the system.

In order to categorize the users, three different roles and an administrator are used to organize the user access.

REGULAR USER - is able to query datasets and modify only a specific set of datasets.

RESEARCH MANAGER - is able to modify all datasets as well as review all datasets of the associated institution.

DATA MANAGER - is able to modify all datasets and has the ability to upload new data. This requires basic training for users to identify VCF file format violations that might lead to misleading variant calls in the database.

ADMINISTRATORS - have access to all datasets and are able to add and modify users & institutions, as well as their roles.

To increase the integrity of data access, the elements that users are authorized for are retrieved through a single SQL statement. This prevents developers from mistakingly granting unauthorized access when extending the platform.

7.2 ANNOTATION, QUERY, AGGREGATION FRAMEWORK (AQUA)

The AQuA framework defines three module types which developers use to extend the variant discovery capabilities. Annotation, query and aggregation modules are independent from each other and enable a configuration based approach, which allows developers to add new modules with reduced programming efforts. This principle allows developers without much experience in the underlying web application framework to contribute new features and filters.

All modules are configured to support a set of one or more organisms and variation types (i.e. SNV, Indel and copy number variation).

AQuA is divided into three module categories

ANNOTATION , which executes external annotation tools and store the result in the database.

QUERY , which merge a set of filter conditions into feature-based queries. For example, a query for a gene identifier merges multiple filter conditions to match against different gene identifier names (e.g. gene name, gene symbol etc.).

AGGREGATION , which provides variant annotation attributes and presentation options for the query result.

All modules are independent from each other, although in practice each annotation tool will require a set of query and aggregation modules to provide the annotation features to users.

7.2.1 *AQuA annotation modules*

Annotation modules provide the functionality to setup, manage and execute external variant annotation tools as well as storing the result of the process. Their role is to set up the requirements to run an external annotation tool, check for system dependencies and ensure that execution of the tool is possible. When this is implemented fully, automated tasks are run to apply the annotation to all datasets in the database and ensure that every variant is annotated. How and which annotations are stored is defined by a database model, which is also used further downstream to define filter dependencies and generate the appropriate query statement.

One annotation module consists of the following components:

RAKE FILE Rake is a make-like system that is used to define setup and clean up tasks for an annotation resource. Several general tasks are available, which are available for all tools. This includes an activation and deactivation tasks, which is used to run the annotation on all datasets currently in the database and make sure the annotation is available for every one of them. Manual

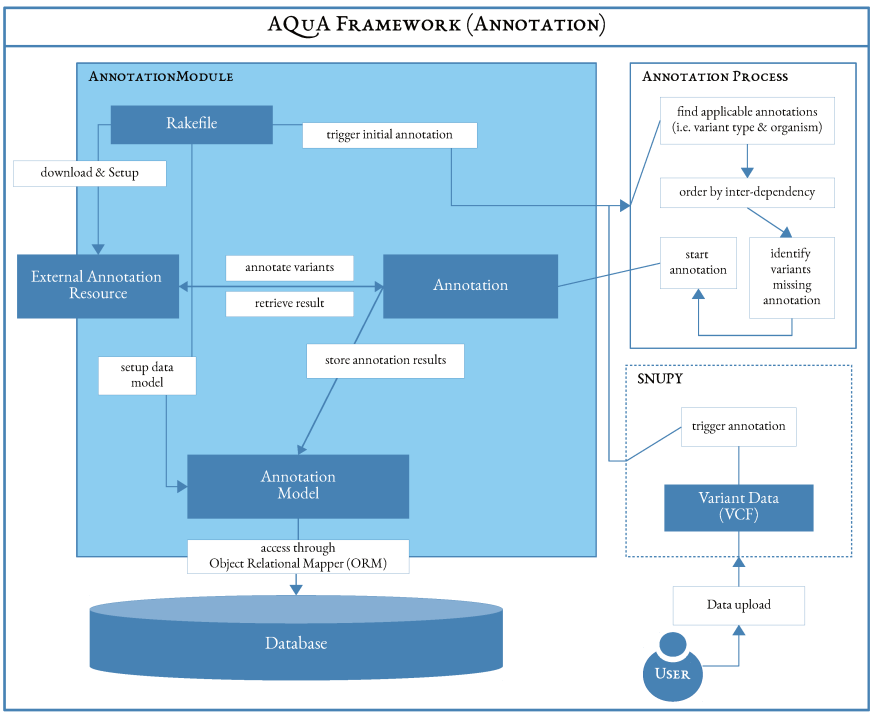


Figure 7.3: This figure shows the components of AQuA annotation modules and the process, that coordinates these annotations. The process is started after data is uploaded to SNuPy . The first step identifies applicable annotation tools by their supported organism and variant types, the second stage orders the annotation tools in a way that fulfills interdependencies between modules and determines variants that have not been annotated. The annotation module then annotates these variant using an external annotation resource (e.g. program or database) and stores the result using a model definition that abstracts the necessary database operations.

annotation of a VCF file from the file system, enabling advanced users to integrate the annotation modules into custom workflows. Further customized tasks can be implemented and executed when necessary.

DATABASE MODEL The database model defines the tables, objects and relations which are used to store and provide the annotation data. This includes a database migration task that sets up the table definition and the model, which defines methods and relations on other models that can be used in a later step to define dependencies. It has to contain at least two attributes that are used to identify the variant and organism that a record is referencing. All access to the annotated variant features is managed through these models, utilizing a object-relational mapper.

AQUA ANNOTATION IMPLEMENTATION Here methods to annotate an input file and storage of the results in the database is implemented. Standard file formats for variant annotation, such as VCF files are generated as input for the module by the annotation process, according to the specifications in the implementation. Annotation implementations are also configured to support specific variant types (snp, indel, cnv) and organism.

Figure 7.3 shows an overview of the process that is used to annotate variant datasets. During the setup phase, an administrator uses the task definitions in the Rakefile to download and setup the dependencies for the external annotation tool, as well as use it to trigger the creation of the necessary data model. After completion, existing datasets are annotated as part of the initial annotation. Afterwards an annotation process will always trigger the annotation, when it is applicable to the uploaded dataset.

The first step in the annotation process is to identify all annotation modules, which are (a) suitable for the variant type (SNV , InDel , CNV) and (b) able to provide annotation for the organism the data is based on. To support more complex annotation scenarios annotation modules can contain inter-dependencies, this way a tool that works on protein interactions can rely on the protein consequence prediction of another annotation resource. The annotations are then processed in the order determined by the topological order of the dependencies and will fail if circular dependencies are detected.

Before annotation, the missing variants are determined and a minimal input file is generated, eliminating redundant annotations. The minimal input file can have different format, currently supporting CSV and VCF . An AQuA annotation module implementation will then annotate the input file and use the annotation model to store the results.

It is also possible to use the annotation capabilities directly on an input file, without storing information in the database. This can be used to integrate the annotation capabilities of SNuPy into a larger workflows.

7.2.2 *AQuA query modules*

Query modules implement the requirements for abstract queries (see section 6.3) and merge filter modules, which provide the concrete filter conditions.

Abstract queries allow natural problem formulation for genomic scientists in a way that is comprehensible and matches the knowledge domain of the field. Detailed filter details, such as feature names and meaningful value ranges can be specified by developers, while users are enabled to translate their hypothesis into a query unobstructed from specific feature names. Fine granular modifications to the query are made possible by activating and deactivating specific filters, if necessary combining them with different logical operators.

A query module has the following components:

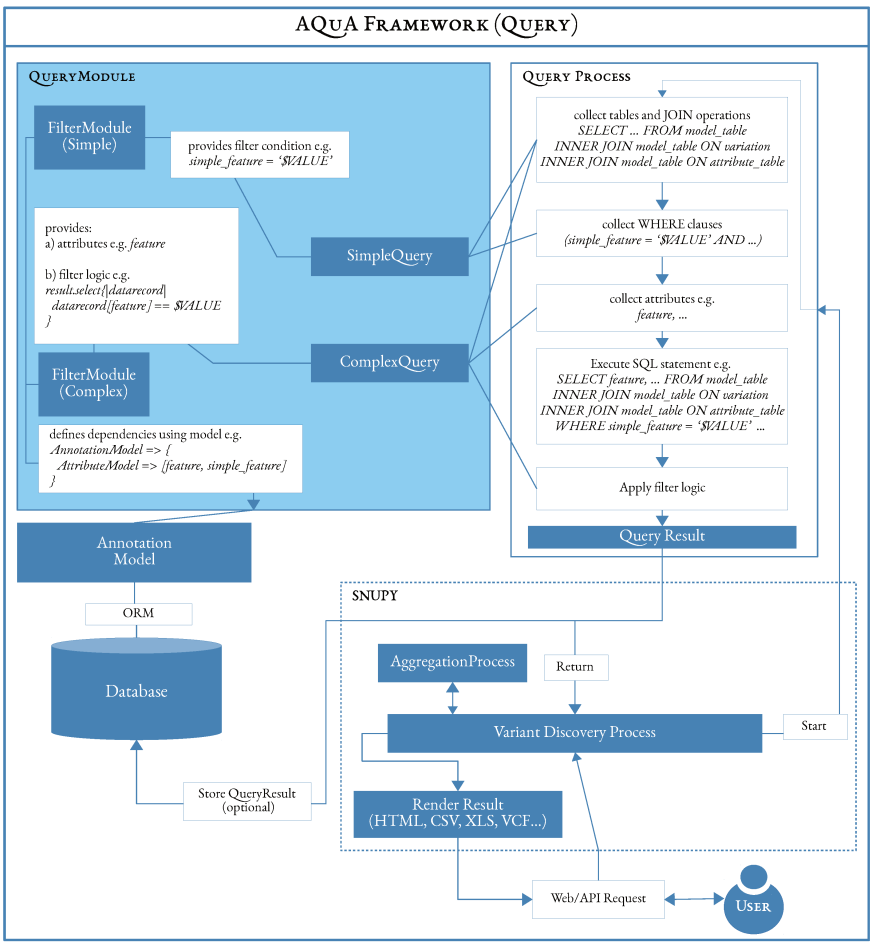


Figure 7.4: This figure shows a query module and the process, that utilize them to provide the filter functions to users. A query is triggered by a variant discovery process, that is configured by an external request. It starts with collecting the required table and attribute dependencies, compiles them into a SQL statement, which is executed by the database management system. The result is then further filtered programmaticically. The final result is returned to the variant discovery process and optionally stored in the database. The response for the user is generated by a subsequent AggregationProcess (see Figure 7.5).

AQUA QUERY IMPLEMENTATION which specifies abstract queries. This definition contains the data type (e.g. number, range, text etc), the default value, a tooltip to explain what the query is used for and a list of organism this query is applicable to.

AQUA FILTER IMPLEMENTATION provide methods to generate a partial SQL condition or define a computational procedure to filter a list of

variants. Such a filter also includes a specification of requirements, which have to be retrieved from the database.

Each abstract query is supported by a set of filters that provide the necessary filter conditions. The set of filters is combined either using logical *OR* or *AND* operations.

Figure 7.4 shows the processes that are executed when utilizing query modules. The query definition is submitted using HTTP, either through the web interface or by an API request (see subsection 7.2.4). A variant discovery process in SNuPy passes the query parameter on to the QueryProcess, which first builds the SQL statement, then executes programmatic filters and returns the result. To built the SQL statement three information are necessary: 1) the attribute names that are required for the programmatic filters, 2) the SQL conditions that perform database-side filtering and 3) the tables which are used in the process. The latter also accounts for relations among depending models, introducing necessary JOIN operations to the SQL statement. The other information are obtained from the active filter modules, which utilize the Object Relational Mapper (ORM) ActiveRecord to provide database specific partial statements.

7.2.3 *AQuA aggregation modules*

Aggregation modules decouple the features required during the data filtering process from the results, which are displayed. An aggregation module consists of a AQuA aggregation class implementation, which defines the features that need to retrieved from the database and a method to enrich them(presentation logic), e.g. with links to other websites or resources. Optionally, developers can define color gradients or assign colors to regular expresion matches to highlighting important features.

Two aggregation types are available: attribute and grouping aggregations . The first retrieves annotation attributes for a variants, the second groups records of a query based on an attribute. Attribute aggregations define which attribute is retrieved and how it is presented. Grouping aggregations are required to denormalize the output of relational table queries, which is convenient for data analysis, but is confusing for users who are confronted with redundant information. Thus, users may chose to group their aggregated results by exact location, or region overlap. Just like query module implementations, aggregation module implementations specify their dependencies using the annotation data model and associations to generate a SQL statement, that is database system independent.

Figure 7.5 shows the aggregation process, which is triggered after a query process was either executed or its result loaded from the database. The colorizing capabilities are only applied to the result when users decide to render it in a format, that allows coloring (e.g. HTML websites).

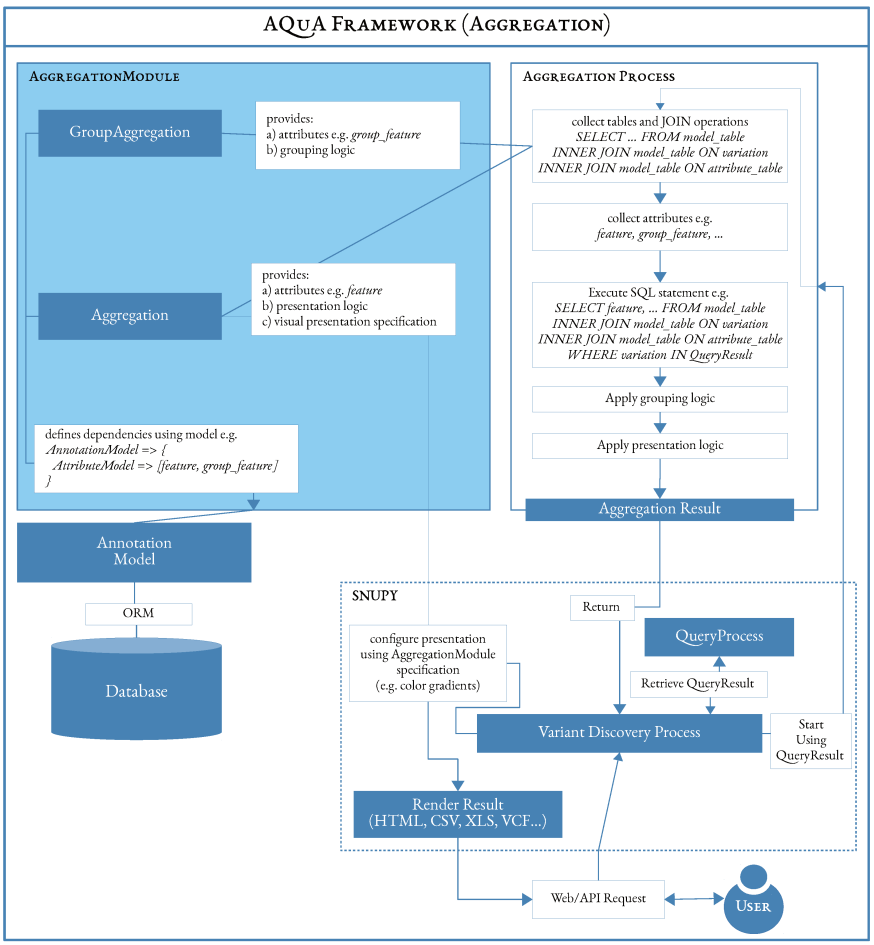


Figure 7.5: This figure shows an aggregation module and the process, that attribute compilation for a variant discovery process. It starts with an external request that defines the desired variant attributes. The variant discovery process executes the aggregation on the result of a query process and enriches the genotyped variants with information from the annotation resources. For this purpose the dependencies and attributes are compiled into a SQL statement, which is send to the database management system. The result is then grouped by an attribute, or grouping logic (e.g. overlapping regions). After the records are grouped the presentation logic is applied, which may reformat the variant attribute for better comprehension of usefulness (e.g. link to an external resource, abbreviate long descriptions). The result is returned to the variant discovery process which, when applicable, uses the aggregation presentation definitions (e.g. colors) to generate the desired output format. The latter may not be necessary when coloring is not required (e.g. VCF output format).

7.2.4 *Query Capabilities*

SNuPy implements three different ways to query datasets. First are parameterized queries that filter a user-defined set of datasets by user-defined criteria and thresholds. This allows users to explore datasets and refine queries in a top-down fashion that incrementally applies refined filter conditions to find candidate variants.

Second are per-defined queries that can be used to query datasets in a reproducible manner without the possibility for users to change the parameters individually. This is used for large studies that need to run the same query on many individuals and require lesser exploration.

Third are meta-queries that allow users to query all samples of a specific organism that he has access to. When hundreds or thousands of datasets are available and spread over multiple projects, this empowers user to perform a bottom-up approach. Here candidate genes are already known (e.g. from other studies) and users are interested in which samples carry mutations in these genes.

All of these ways utilize the query process defined above in order to make sure that results retrieved through one are reproducible using another method.

External Interface

An external advanced programming interface (API) enables the data management and query capabilities of SNuPy to be integrated into larger workflows and into libraries for other applications. Two output formats are supported: JSON (JavaScript Object Notion) and TSV (tab-separated values). The first is a state of the art way for object serialization and has wide support in other programming languages and programs. The TSV format is a human readable format, that is supported by spreadsheet programs so users can view the tabularized data and makes processing of data easier.

Automated data management and sample retrieval can be based on a RESTful HTTP[a] interface, allowing programmatic retrieval of sample groups and additional information in JSON or TSV format. The same RESTful interface allows to display the HTML-based web interface, that users interact with when HTML out is requested. The latter is the defalt behaviour when SNuPy is access through modern web browsers.

Such an approach allows any information and action that is available through the user interface to be accessible through a programming interface too. Additionally, the same access rules apply, as they do when users access the platform. This means that any restrictions defined by the role-based authorization approach is abided when data is accessed programmaticically.

Users of the web interface are shielded from retrieving excessive amounts of variant genotypes, because of the practical reasoning that large number of variants

a RFC 7231: https://tools.ietf.org/html/rfc7231

(e.g. more than 15,000 by default) are not feasible for regular users in most circumstances. However, the API allows complete datasets and large volumes of variants to be analyzed and downloaded programmaticically.

Part V

IMPLEMENTATION

8

SNUPY

8.1 OVERVIEW

SNuPy is a Ruby-on-Rails web application that uses a MariaDB database to empower genomic scientists to query their own variant datasets. This chapter will give a brief introduction to the web interface and how users interact with it to query and interpret variants.

The web application provides access to the functionality and processes, that users can execute to manage and explore whole-exome datasets.. This includes the basic model definitions for the data and user access management, as well as the sample hierarchy, the variant data and VCF files. Furthermore SNuPy provides methods to visualize data in dynamic tables, network graphs or plots.

The data management and processing concepts presented before need to be organized and made accessible for genomic scientists, both on a technical and usability level. A web application solution is system independent, scalable and can be used even with limited hardware resources on the client side. Its interface allows flexible, dynamic and user friendly interaction, while the underlying data model allows validation which is crucial when custom data is uploaded.

Figure 8.1 shows an overview of the platform and its components. Users interact with the system through the web interface or HTTP API, that requires authentication before directing the request to the actual web application. Performing authentication on the web server allows SNuPy to be used in conjunction with external authentication services (e.g. LDAP and Kerberos). This allows seamless integration into existing user management systems, that are commonly used in institutional IT infrastructures. Additionally, such a setup ensures that users are prohibited from accessing any information that is stored inside SNuPy without proper access permissions.

After users have been authenticated, the authorization process checks if the user has the permission for the requested action (e.g. uploading a VCF file). This stage makes use of the role-based authorization system described earlier. It grants access to the tasks that users require to manage variant datasets and users, as well as perform quality control, variant discovery and reporting.

Access to the database, for basic data models (see section 8.2) and the variant annotation is performed through the ActiveRecord object-relational mapper making the system database independent. Tasks that require to work with variant annotation use AQuA modules to access annotations or filter datasets (described below

section 8.3). Annotation modules are executed for every uploaded variant dataset and starts external tools (see section 9.1 for details).

Long running tasks can be executed in a separate process in the job queue (see section 8.3). This job queue runs a separate SNuPy instance from the web application and works off jobs that are submitted to the database via the web interface or API.

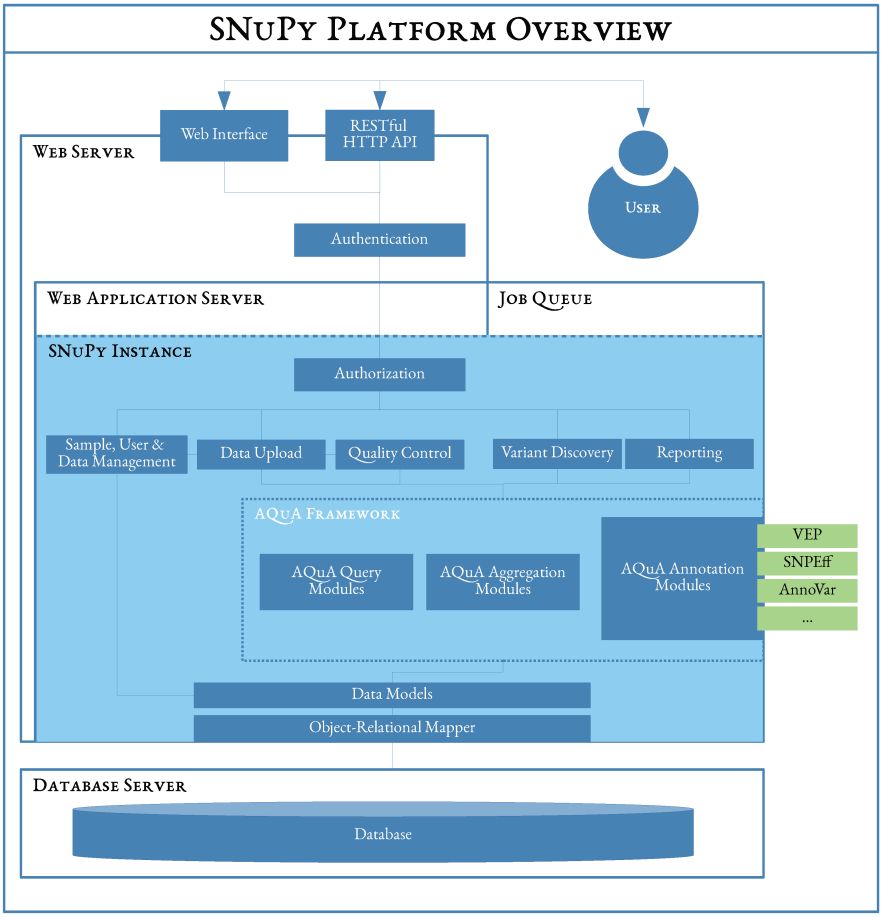


Figure 8.1: SNuPy has three major components, first there is a web server and a web application server that provide access to SNuPy over the web and HTTP interface. Second there is the job queue, that is a separate service that processes long running tasks and third is the database server, that manages the relation database. Both the job queue and the web application server use separate SNuPy instances and communicate through the database only. The AQuA framework is used to integrate external tools into the platform.

8.2 DATA MODELS

This section will present the basic data models used by SNuPy to provide the necessary functionality (see figure 8.2). Its purpose is to give a broader overview over the models and their interactions.

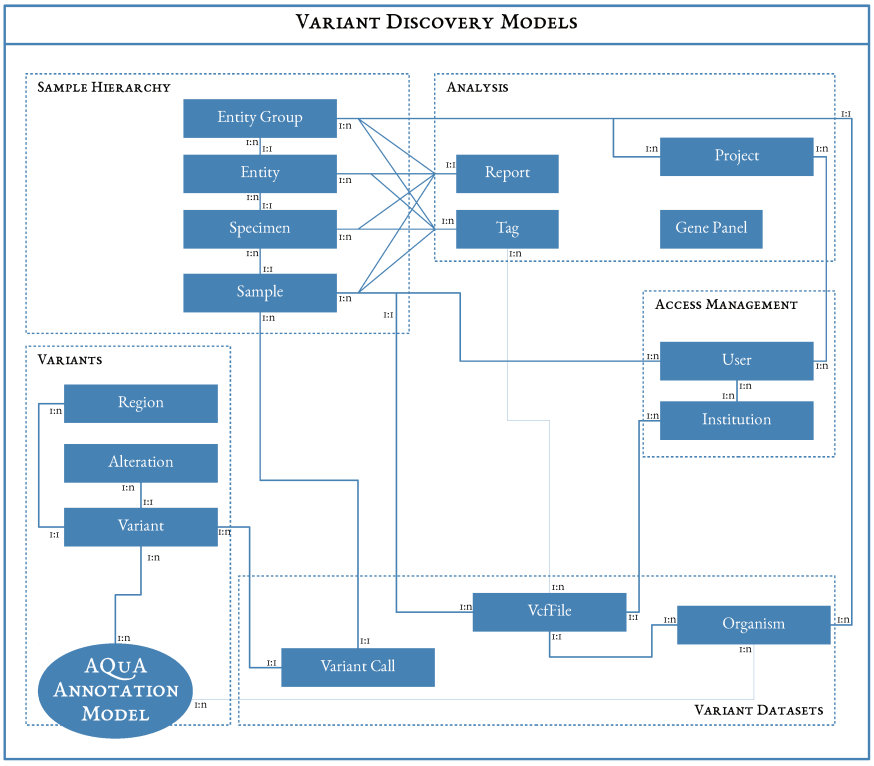


Figure 8.2: This figure shows the basic data models, which are available in SNuPy . They model the required data models, that allow working with variant datasets in SNuPy . The models are placed in five categories for better representation of their purpose. Data managers upload new VCF files and extract the genotypes from the samples contained in them. These are linked to the variants and annotated by the AQuA annotation process, if no annotations has been computed previously. Access management to projects, samples and datasets is defined by users and their roles in different institutions.

The base models can be categorized into five different categories:

SAMPLE HIERARCHY - holds all information regarding project and dataset organization, including gene panels and the tags used to document datasets.

VARIANTS - includes models that are used to collect annotations and information regarding a single mutation.

VARIANT DATASETS - uploaded raw data and the set of mutations per sample are defined here.

ACCESS MANAGEMENT - is used to manage authorization to different parts of the web application and the actions users are allowed to perform on datasets.

ANALYSIS - Contains all relevant information to an analysis, including the generated reports.

Information about variants is divided into two parts: One are the variant data models, which hold information regarding the genomic position and nucleotide acid exchange of a variant, the second are variant dataset models, which link variants to observed genotypes in individuals (variant calls). These genotypes depend on the analysis the NGS sequencing data is based on and is uploaded as a VCF file. The variant dataset is also where the multi-organism support in SNuPy is rooted. Each variant call is associated with an organism through its VCF file association, that is necessary for the AQuA annotation processes to determine the correct organism-dependent database, such as transcript sets.

As described above, AQuA modules use these standard data models to annotate variants with external information and define associations between annotations and variants.

The sample hierarchy allows the organization and categorization of variants datasets using a flexible tagging system.

Gene panels can also be added dynamically to provide re-usable lists of genes of interest. This is useful when the same set of genes is either of interest to larger user group, or a specific study setup.

Additionally, users can associate static and dynamic reports with their datasets. Static reports are files, that are associated evidence and is uploaded through SNuPy . This can include reports on the variant calling process for VCF files or additional documents relevant to the dataset documentation. Dynamic reports are generated from a query result and are associated to diseased entities. Such reports provide standardized summaries that users store when they have found a variants of interest.

8.3 PROCESSING INFRASTRUCTURE

The basic processing infrastructure consists of two elements: One is the web server, which receives query requests and renders the results into the output format. The second is a daemon that polls a job queue for new jobs, executes them and stores the result in a database to be delivered to the user through the web interface later. The latter is required to handle long running or computationally intensive jobs.

A central job queue also provides basic load balancing and scaling capabilities, when multiple worker processes execute jobs from the job queue. Two different

queues are used to handle sample annotation & extraction in one and query tasks in another.

Long running web requests are often blocked by firewalls, which are common in clinical IT infrastructures. This is problematic for computational or read-write intensive jobs, because they usually take longer than the firewall timeout allows. Such tasks are can be submitted to the queues for asynchronous processing. The central database is used to place jobs in the queue, which in turn are processed by job queue workers. This concept allows scaling with any number of workers, which have access to the central database.

8.3.1 *Annotation queue*

The annotation queue processes jobs for variant annotation, computing quality measures and storing genotype data from VCF files in the database. These tasks are read-write and computationally intensive. The runtime depends on the amount of data uploaded and the annotations that are computed.

8.3.2 *Query queue*

The queries which are submitted use flexible filter values. Thus runtime can vary greatly between queries, because of the attributes which need to be matched and the number of datasets that are included. The query queue processes are dynamic read-only operations and might be processed by any number of workers, thus providing scaling for a growing user base. Furthermore the usability is increased by caching of long running jobs that allow to load results from previous complex query.

8.4 DATA MANAGEMENT

8.4.1 *Vcf File & Sample import*

The basic workflow to add new variant data consists of four steps:

1. Upload variant data in VCF format
2. Annotate the variant data using AQuA
3. Extract genotype information from samples in the VCF file.
4. Link the genotypes to the specimens

In the first step, data manager upload VCF files and link them to an institution and an organism. The institution is required to make sure the data can only be accessed by authorized users, the organism is required to apply the correct annotations. The integrity of the upload can be ensured by providing a MD5 checksum for each VCF file.

In the second step the annotation of the VCF file is performed, using the applicable AQuA annotation modules (see section 9.1). A tool is not be applicable, if the given organism or variant type is not supported. This step only annotates variants, which were not annotated before to avoid unnecessary annotation computation.

Once the annotation was successful for all applicable annotation tools the genotype information can be extracted into samples from the VCF file.

A standard VCF file contains variant coordinates as well as the state of a variant in a list of samples. This format allows to store arbitrary information alongside a variant (in the *INFO* field) and the individual samples, but also defines a core set of attributes, which should be present for every record. These include:

GENOMIC LOCATION - chromosomal position of a variant.

REFERENCE & ALTERATION - two columns that carry the reference allele and a list of alterations. Multiple alterations are possible, if both alleles deviate from the reference or multiple samples were sequenced with different mutations.

FILTER STATE - a filter field that indicates if a variant has passed all necessary filters during variant calling. *PASS* and . suggests that a variant passed the filters, all other values indicate that a filter has failed.

GT - the genotype of a variant in a sample, representing homozygous and heterozygous mutations.

DP - the read depth of a variant in a sample.

AD - the allelic depth of a variant in a sample, allowing to calculate frequency of a variant in a sample.

GQ - genotype quality, a measure that is used by variant callers to rate the confidence in the genotype prediction.

FS - short for fisher score, used to measure the strand bias of a variant call. Some variant are only called on one strand of the sequenced fragment, which suggests a systematic error in the sequencing process. Variants with a hight strand-bias score are usually removed or discarded.

The information that is encoded in these files depends on the variant caller, I have implemented VCF parsers for the standard format, the variant caller GATK[154], the somatic and trio variant caller VarScan2[151] as well as the copy number variation caller EXCAVATOR[152].

If other tools deviate from the standard format definition developers can implement their own parsers, which extract the minimal requirements defined in the VCF standard. The parser allows to define rules for records and header, to make sure the user uploaded data in the correct format. VarScan2 for example produces

a VCF-like format, that requires careful attention and does not comply fully with the standard format.

Because projects may contain many VCF files, which in turn may contain many (possibly hundreds) samples the upload and sample extraction procedures allow batch processing. VCF files can be uploaded as ZIP archives, which include a configuration file associating variant datasets to an institution, organism and the name to use inside the platform. Sample extraction sheets allow batch extraction of samples from VCF files. They are uploaded as tab-separated-value (TSV) files, containing information of the sample type, name, tags and associated users.

The forth and last step is to link a sample to the sample annotation hierarchy. This is done by research managers and also allows to create large batches using the web interface. In general, research managers have access to all relevant information about a NGS run, its history, the phenotype and additional information that are relevant to the study. Once a sample is categorized, it can be queries through the web interface.

8.4.2 *Database & ActiveRecord*

A relational database is used to store all datasets and other information. ActiveRecord maps the data models to the database. This object relational mapper allows to instantiate objects from database record and define direct and indirect relations between models. All relations are kept consistent by triggers that are executed before, during or after an object is created, updated or destroyed. The AQuA framework uses these relations to determine the requirements for a module and compile query statements during the query and aggregation processes (see below). ActiveRecord provides several database adapters to the most common database management systems, allowing the discovery platform to be deployed on a wide range of database systems.

8.4.3 *Access Management*

The access management is implemented as a role-based user system, that grants users different permissions in different institutions. The roles and the permissions they give to users were described before in 7.1.5.

A role-based authorization system (*cancancan* library) is used to grant access to each action users can perform through the web interface. This includes actions to query, create, edit or removal of elements.

Developers may use the three different methods to access objects on either level of the sample hierarchy (7.1.2, or any other object that has a relation defined to an institution and user.

`USER.VISIBLE(MODEL)` - retrieves all instances of a model, which are associated with any of the users' institutions, or where the object is associated to the user through a direct association.

`USER.REVIEWABLE(MODEL)` - retrieves all objects of a model, which are associated with institutions, where the user is a data or research manager, or the object is associated to the user through a direct association.

`USER.EDITABLE(MODEL)` - retrieves all objects of a model, which are associated with institutions, where the user is a data manager, or where the object is associated to the user through a direct association and belongs to one of his associated institutions.

The *model* parameter can be any model, which can be associated to an institution and user (see section 8.2). This includes indirect associations, that are defined through other models. This ensures that access to *specimens* is always consistent with the access to the respective *entity group*, because the institution and user association is established through the sample annotation hierarchy.

Granting users access to elements that they are directly associated to, allows a very flexible sharing of datasets with users. It enables specific objects to be shared with users from external institutions without exposing other private datasets.

The usual workflow does not require datasets to be removed from the database. However, if it becomes necessary, this task is only authorized for administrators to prevent any accidental or deliberate wrongdoing when handling the datasets.

8.5 WEB INTERFACE

Data and user management interfaces are available to upload VCF files, extract samples and integrate the data into the sample hierarchy, that is used to organize the data (see Figure 8.3). Users are shown extracted samples that have not been placed into the sample hierarchy to improve organization. Long running jobs, such as annotation and complex queries are send to the job queue that is displayed at all times, allowing users to access the result conveniently.

Details for datasets and variants are shown in JavaScript-driven, dynamic Hypertext Markup Language (HTML) tables. Developers colorize elements of the table by defining rules as exact or pattern based matches against texts, color scales for numeric values or a discrete coloring for factors (for an example see Figure 8.7).

8.5.1 *Query interface*

To query datasets, users first select which samples they want to query (see Figure 8.4). It is possible to pre-select all datasets that match specific tag (e.g. tissue, disease etc.), to make the query design easier when hundreds of samples need to be queried. This

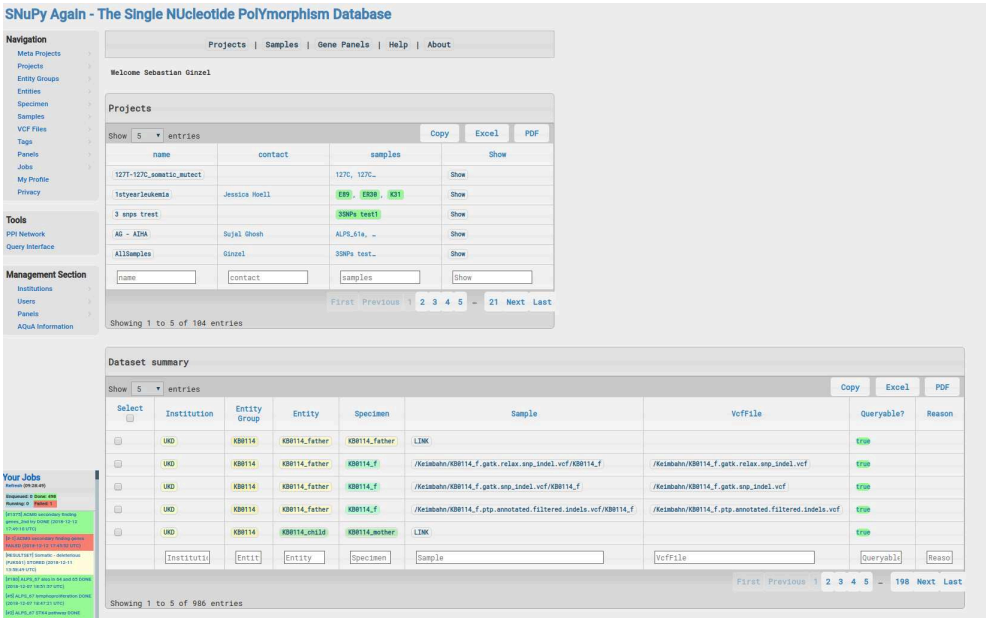


Figure 8.3: The basic SNuPy interface.
Active projects, a list samples that need to be integrated in to the sample hierarchy and of previously or currently running jobs are displayed when users access the platform. Tables are supported by a JavaScript library (dataTables) to a allow dynamic filtering and sorting.

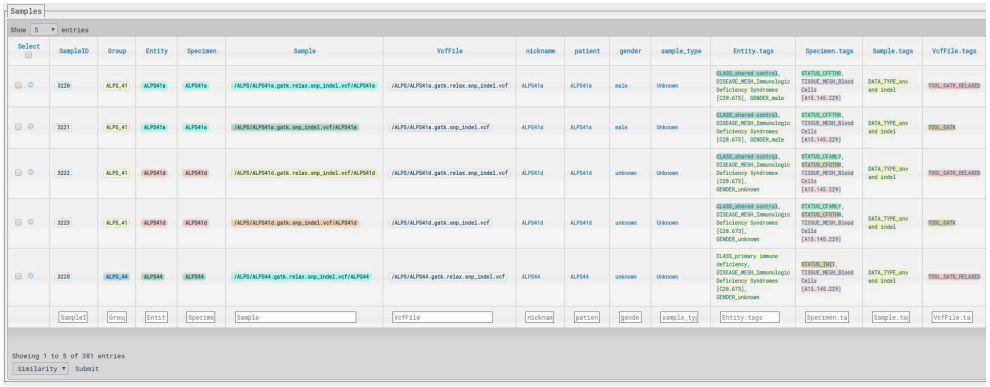


Figure 8.4: An example for an annotated query output.
The dynamic tables allow developers to use different coloring rules to help guide users to find relevant variants more easily. Developers can add methods to the bottom of a table, that allows users to perform actions on selected elements. In this example, users can calculate sample similarities between selected samples.

feature also allows users to apply set operations to the selection of samples, e.g. to select all datasets that are tagged as relapse and were built using the somatic caller Mutect[155]('mutect & RLPS').

After users have established the samples of interest, the filter conditions are specified through a variety of categorized query criteria (Figure 8.5). If necessary users can

use different filter conditions and combinations, for each filter criteria. Figure 8.5 shows an example, where the user has chosen to find homozygous variants that are covered by at least 10 reads, have a frameshift or otherwise fatal consequence for the protein product, is heterozygous in two other samples and is present in not more than 15% of the general population. The setting for the last criteria are shown in the figure and demonstrate that users have a choice how the abstract query condition *present in not more than 15% of the general population* is applied. The example shows that the condition takes the variant frequency of the 1000 genomes and the ExAc project into account, as annotated by the Variant Effect Predictor (VEP).

Users also chose which aggregated attributes should be displayed for each variant (Figure 8.6), allowing focused assessment of variant impact or a more explorative approach. Each variant attribute is then displayed as a column in the result table (Figure 8.7), possibly color coded.

The color coding highlights important or unusual attributes for a visual inspection by users and to provide a reference point for its significance. This is a necessary step during the variant interpretation process, because of the multitude of unscaled variant scores that are available. Without such interpretation guidance, users need to research the meaning of scores and how to interpret them themselves, leading to a fragmentation of the way the scores are interpreted.

User also have the option to download the results directly, instead of viewing them in a web browsers. This is necessary e.g. when results are stored for publication or later use in an external program.

8.5.2 *API Interface*

SNuPy provides access to its data management and query capabilities through a RESTful HTTPS interface. The model-view-controller (MVC) framework, provided by Ruby-on-Rails, allows to provide such an interface in conjunction with the regular HTML-based user web interface. This approach reduces redundant code because the API returns the same information as the user interface does, only in machine readable format. Because the access is based on HTTPS protocol, a secure and encrypted connection is established between the API user and the server.

There are two use cases that users of the API will face. First is the data management and sample annotation feature of SNuPy , that is available through a pure REST interface. This allows the following actions on objects of the sample hierarchy and its information, by using the appropriate HTTP verbs: listing all or a specific object (GET), create an object (POST), update an object (PUT) and delete an object (DELETE). This allows developers to identify samples of interest that they want to query (e.g. samples derived from tumors). Second is the usage of the query and aggregation capabilities, which can be utilized through the /aqua endpoints of a project (e.g. projects/1/aqua) and returns the results in JSON or TSV format.

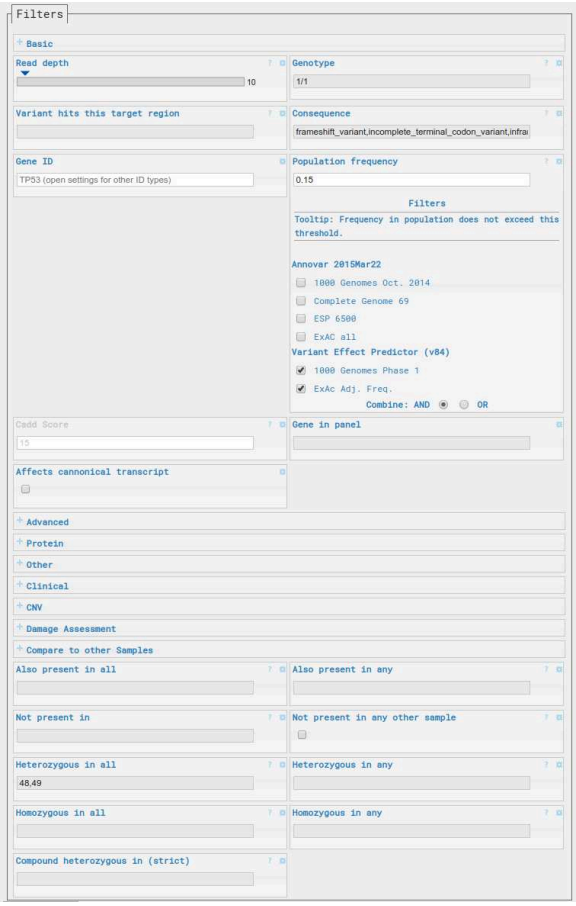


Figure 8.5: Query filter configuration.
The filter conditions and procedures are grouped into categories for easier sorting. Each user-defined value is passed on to the selected filters, that are used to reduce variant datasets. When multiple options are available users are displayed dialog that allows them to select specific elements.

This API interface has been used to develop a R-library[a] that allows access to variants and annotations in an R environment. R is a well established platform for statistical analysis, machine learning and visualization. The R library uses R6 classes to represent the objects of the sample hierarchy and their relation, as described earlier. This solution allows call-by-reference, which is critical because variant datasets can become large and data duplication puts a heavy burden on the working memory. To retrieve variant data it also uses a streamlined download from SNuPy , allowing the data to be parsed while data is still send from the database. The library was used for custom analysis workflows that integrate additional data sources into a study (e.g. see chapter 11), as well as making use of the advanced statistical methods to analyze datasets with respect to a specific study question, that cannot be answered with the SNuPy web interface.

a available at https://github.com/sginzel/snupyR

Figure 8.6: The list of grouped aggregations.
Each tool provides its own set of aggregations that users are displayed as part of the query result. Some aggregations (e.g. Loss Of Function Prediction) will add multiple columns to the result in order to make it easier for users to select consistent sets of attributes.

8.6 QUALITY CONTROL MEASURES

Quality control measures can be calculated on-the-fly (e.g. sample similarities) or are pre-computed for individual samples. When multiple samples are assigned to a project the quality controls are presented in an aggregated fashion to make comparisons easier. Quality control summaries are displayed either as plots or color coded tables.

SAMPLE SIMILARITY

Several measures of sample similarity are available for users that allow to assess the mutational similarity between samples.

The similarity between samples may be calculated from:

Figure 8.7: The result of an example query.
Variants and genotypes in samples are displayed in rows and using color coding to highlight features that are important for users. For example, the population frequency ranges from red to green, depending on how common a variant is. This example also shows that users see the frequency of a variant in the database, helping users to judge its importance. Categorical values, such as the amino acid class of the affected amino acids are highlighted in different colors to support quick visual assessment of differences.

ABSOLUTE OVERLAP - although this is not a similarity in a metrical sense, the absolute overlap between two samples lets users compare samples and investigate possible sample swaps.

RELATIVE OVERLAP - as the absolute overlap this measure is not a mathematical similarity measure. But it enables a normalized view on the size of the overlap between different samples.

COSINE SIMILARITY - The cosine similarity between two samples is based on the cosine similarity between their b-allele frequency vectors a and b.

$$cosim(a,b) = cos(\overrightarrow{a}, \overrightarrow{b}) = \frac{\overrightarrow{a} \cdot \overrightarrow{b}}{|\overrightarrow{a}| \cdot |\overrightarrow{b}|}$$

Using allele frequency vectors results in higher similarities between related members compared to unrelated members. Because all frequencies are > 0, the similarity can be transformed to a distance measure ($codist(a,b) = 1 - cosim(a,b)$), which in turn can be used for sample clustering.

The sample similarity measures allow users to identify possible sample swaps or sample contamination, when combined with information from lab-books and other documentations.

B-ALLELE FREQUENCY PLOTS

The b-allele frequency is defined as the frequency of an alternative read among all reads at a given position[1].

[1] *I do not use the total number of reads at a position for the frequency, due to low quality base calls some reads are not taken into account when calling a variant.*

$$baf(var) = \frac{|altreads(var)|}{|altreads(var)| + |refreads(var)|}$$

The b-allele frequency distribution plots helps to identify possibly unexpected contamination of samples with foreign DNA material. This quality control measure plots the frequency distribution of heterozygous variants for each sample. It can be expected that heterozygous mutations are evenly distributed around the expected frequency of 0.5. *Note that multi-allelic variants are ignored for this kind measure.*

In case of a contamination this distribution becomes a mixed distribution with additional spikes to the left and right of the 0.5 peak, because of homozygous and heterozygous mutations introduced from the contaminant.

Such a contamination may be expected if the sample is impure or if donor material is expected to be present in a sample that underwent allogeneic stem cell transplantation at some point. Interpretation of the sample context by the user is therefore always necessary.

VARIANT COUNT & QUALITY SUMMARY A summary of the variant counts is provided as a measure that users can include in their sample statistics. It includes the total number of variants, deletions and insertions with different genotypes (homozygous, heterozygous and heterozygous without reference), for all variants and those covered sufficiently (20-fold). Mutational profile patterns may be extracted from the nucleotide exchange statistics that list the number of specific nucleotide exchanges for each sample

To identify abnormalities regarding the read depth and variant count distribution a detailed chromosome specific overview is calculated as well. This include the average quality of the variants, the average read depth and number of variants for each chromosome. This may help to identify large deletion events, which can result in a reduced number of reads on a chromosome.

SEX PREDICTION A sex prediction was implemented to suggest to users whether a sample is male or female, which can be matched against the documentation of a sample and may be used to identify possible sample swaps. The calculation is based on the average ratio of homozygous mutations over all chromosomes, compared to the same ratio on the X-chromosome.

For female samples the ratio of homozygous mutations on the X-chromosome should approximately be the same as on all other diploid chromosomes, thus the ratio between the average portion of homozygous mutations is close to one.

Because all variants on the X-chromosome for male samples are homozygous the expected value for this group is close to 2. Larger deviations may be due to chromosomal aberration, such as XXY karyotpyes, contamination or other genetic effects and should be investigated by experts.

8.7 REPORTING

Static and dynamic reports allow users to add additional information to query results and facilitate standardized reporting for clinical application (see 7.1.3). This allows users to communicate findings with external partners, share results and support reproducible variant interpretation. To generate a report, users select a list of variants of interest from their query results After the report is generated, it can be download as .docx file, which allows users to work with the result using text processing software (e.g. LibreOffice, Word). After editing the report, possibly adding additional information for its interpretation, users can upload the modified report to document their findings.

Three dynamic reporting templates have been developed, based on a plug-in mechanism that lets developers add additional templates (e.g. project- or disease specific).

First a general report, that users may use to summarize their findings (see supplement 3 for an example). They select variants of interest and optionally a set of gene panels to configure the report. If no panel is selected all selected variants are part of the report, otherwise the report is split into sections for each panel. A section of this report contains three subsections: (1) a transcript based information about the affected genes and transcript consequences, (2) a table of allele frequencies in the specimen of the patient and relatives (if available) (3) a table of clinically relevant information including the associated phenotypes and loss-of-function predictions. The last page shows the query parameters that have lead to the discovery of the variants in the report to provide basic documentation and allow to refine the query later.

Second, is report for gene-drug interaction targets (see supplement 2 for an example). It adds information about the druggability to the gene that variants affect. The information for drug-gene interactions are retrieved from DGIDB[156] and updated regularly to provide most up-to-date information. This database aggregates interactions from various resources cumulating to more than 50,000 documented interactions for more than 3,000 genes[b].

Third, the ACMG/AMP evidence framework for pathogenic variants[38] is made available as a report (see 1). The report includes an overview table for each variant, that users select from a query as well as a description of the criteria, as published by Richards et al.. For each variant of interest, a single-page table of the evidence framework(see Figure 1 by Richards et al.[38]) is added to the report, where the criteria are substituted for the data which is available in SNuPy . This allows users to perform standardized variant interpretation and allows the results of a sequencing project to be reported in a standardized fashion across projects. The ACMG/AMP template is a proof-of-concept for standardized variant interpretation in SNuPy .

b April 2019

As the recommendation progresses and new disease-specific criteria will emerge I expect these to be integrated in SNuPy as well.

9 AQUA MODULES

9.1 ANNOTATION MODULES

Table 9.1 lists the available annotation modules that have been implemented in SNuPy .

The input and output specification defines which annotation modules can be used independently from SNuPy using the command line. The supported organism and mutation type is part of the module configuration, because not every tool generates annotation for all organisms and variant types. Dependencies can be added that make sure another module was successfully before another tool is available.

Other tools, such as StringDB, DIDA, ClinVar and OMIM are not executed during the annotation workflow, because these tools do not annotate variants directly.

For example, the OMIM module adds the information from the OMIM database to a table. It is then used in conjunction with the gene annotations derived from VEP to provide useful query and interpretation capabilities.

Source	No. Features	Supported Organisms	Supported Mutation types	Depends on
Annovar 2015Mar22	69	homo sapiens, mus musculus	snp, indel	
CADD	2	homo sapiens	snp, indel	
Capture kit	1	homo sapiens, mus musculus	snp, indel	
ClinVar	30	homo sapiens		
DIDA v2	6	homo sapiens		VEP
OMIM	33	homo sapiens		VEP
snpeff 4.1 db_75	21	homo sapiens, mus musculus	snp, indel	
StringDB v9	14	homo sapiens, mus musculus		VEP
Variant Effect Predictor (VEP v84)	47	homo sapiens, mus musculus	snp, indel, cnv	

Table 9.1: A list of annotations implemented as AQuA annotations modules.

9.2 QUERIES

Table 9.2 table shows which queries were implemented as AQuA query modules. The query classes implement the abstract queries described in 7.2.2 and 6.3. In total

users can utilize and combine more than 200 different filter conditions for more than 70 abstract queries from 10 categories.

Category	Query class	No. abstract queries	No. filters
Variant Properties	Target region	2	14
	Population frequency	1	6
	Read depth	1	1
	Genotype & quality	2	2
	Genetic coordinates	1	1
	Recurrence in Sample/Entity	2	2
Genetic Impact	Transcription consequence	1	9
	Genetic identifier	1	12
	Gene panel	1	11
Protein Impact	Affecting canonical transcript	1	1
	Amino acid exchange	2	13
	Protein domain	3	13
Clinical Association	Digenic disease association	1	1
	Clinical significance	2	13
	Phenotype	4	17
	Tissue expression	1	1
Functional Impact	Loss-of-function predicted	1	15
	Conservation	1	6
	Protein Protein Interactions (incl. Interaction to panels)	2	5
Inheritance	Presence/Absence in other sample	8	15
	Single/Combined Inheritance pattern	6	12
	Compound heterozygous	1	1
Sample comparison	Presence/Absence in previous query	2	4
	Restrict search using set operations	2	4
	Allele difference between samples	1	4
CNV	Gain/Loss by CNV	1	1
	Absolute/Relative overlap of CNV with feature (gene, protein, regulatory)	2	4
Regulatory Impact	Affects transcription factor binding site	1	4
	Hits microRNA or its binding site	2	3
API access	Sample details	11	9
	Variant & Genotype details	8	9

Table 9.2: A list of queries implemented as AQuA query modules.

VARIANT PROPERTIES Queries of this category filter variants by position and variant specific features. Some of them are flexible between samples, such as the read depth that depends on the sequencing process. Others are static, such as the population frequency with which a variant has been recorded.

GENETIC IMPACT Includes queries that allow users to remove or retain variants that affect a specific gene.

PROTEIN IMPACT These queries focus on filtering variants by their consequence on the protein product, including the focus on variant that affect protein domains.

CLINICAL ASSOCIATION Allows users to formulate queries that require information from clinical associations and phenotypes.

FUNCTIONAL IMPACT These queries are relevant when loss- or gain-of-function events are associated with a disease.

INHERITANCE Different modes of inheritance can be applied to the filtering process. Because it is an important feature, users have a wide range of options available to combine the inheritance models. They can either specify specifically which datasets to include/exclude or use the dataset hierarchy to automatically detect parental variants.

SAMPLE COMPARISON Because some query scenarios require complex filtering scenarios, this category can be used to query dataset using set operations such as union, intersection and set differences.

CNV Copy Number Variations provide different attribute types than smaller variants, such as SNV and InDel . For example, smaller variants usually only affect a small region of a transcript, while CNV can overlap with large portions.

REGULATORY IMPACT These filters allow users to reduce the result list to variants that have a regulatory, e.g. by affecting microRNA binding sites.

API ACCESS Special queries are available through the API accees, which allow its users to query single properties of the data models and variants.

9.2.1 *Pre-defined Queries*

SNuPy implements a total of seven pre-defined queries, which are used to provide reproducible and systematic queries on all entities of a project. This empowers users to query hundreds of patients, using the same parameter and to compare the results.

COMPOUND HETEROZYGOUS - Compound heterozygous mutations that occur when each parent inherits a defective copy of a transcript to their offspring, leaving the offspring with no working copy.

DE-NOVO MUTATIONS - Mutation not occurring in the parents.

RECESSIVE VARIANTS - Homozygous mutations that are heterozygous in the parents.

ONCOGENOME (GENERAL) - Somatic mutations from any tumor.

ONCOGENOME 1 (INITIAL) - Somatic mutations from an initial tumor.

ONCOGENOME 2 (RELAPSE) - Somatic mutation of the first relapse after treatment.

IMPACTFUL SILENT MUTATIONS - Silent mutations with a considerably high CADD and conservations score.

The great benefit is that each pre-defined query can be configured to exclude parents, siblings or other control sample that are automatically determined by the sample hierarchy (see figure 7.1). Rendering manual selection of controls or parental samples unnecessary, which is a time consuming and error-prone factor when manually searching for the correct parent samples in thousands of datasets.

Additionally users can also configure which variant analysis tool (such as GATK or MuTect) they want to base their analysis on. Developers are free to give users power over additional parameters, for example a population frequency cutoff.

9.3 AGGREGATION MODULES

Table 9.3 lists the available aggregations that users can use to enrich their query results.

As presented in table 9.1, some annotation sources provide more than 60 different features for variants that are stored in the database. The features are summarize and categorized for each annotation resource, to unburden the user and present the features in a user-friendly fashion that allows users to focus on the relevant aspects of variant interpretation.

For example, the 21 loss of function prediction annotations, provided by AnnoVar are summarized into one aggregation module that users can activate to show loss of function predictions as part of their output, if necessary.

Lastly more than 280 variant annotation features are made available through the API access that workflow developers can use to extract specific features of a variant.

Annotation Source	No. aggregation modules
Annovar 2015Mar22	12
CADD	1
Capture kit	7
ClinVar	1
DIDA v2	1
OMIM	1
snpeff 4.1 db_75	8
Variant Effect Predictor (VEP v84)	28

Annotation Source	No. aggregation modules
Other (e.g. inheritance, variant coordinates etc)	15
API access	281

Table 9.3: Number of AQuA aggregation modules derived from annotations. Each aggregation may utilize and summarize multiple available features from the annotation source.

Part VI

RESULTS

10

SNUPY

10.1 DATA MANAGEMENT

SNuPy supports a comprehensive data management that allows to store millions of genotyped variants and as well as their associated datasets and documentation using controlled vocabulary.

Currently the platform manages more than 5,800 raw VCF files, and holds annotations for more than 19 million variants. These result in more than 320 million genotyped associations in more than 5,100 samples available for users[a].

The flexible tagging approach allows the adaptation of SNuPy to different fields of research, currently holding 1,200 tags as controlled vocabulary. Although it was developed in the context of pediatric oncology and immunology, it was used in other studies as well (e.g. drug-resistance).

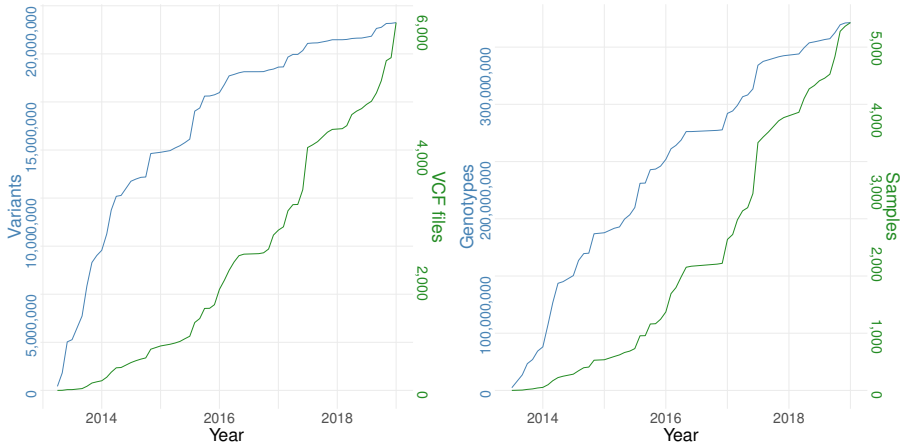
The role-based authorization system facilitates cooperative research and enables users from different institutions to work together on specific user-defined projects. To this day 53 users have used SNuPy to query their variant datasets.

10.1.1 *Integration into external workflows*

The access to data management allows to programmatically upload new variant datasets in VCF format, wait for their annotation to be finished and extract the variants into sample objects. Additionally, elements of the sample hierarchy can be created externally as well, allowing large and complex datasets to be added automatically as part of an external workflow.

Spike[157] is a reference implementation that processes raw sequencing output files, generates the necessary intermediate data, extracts variants and add these to SNuPy is available at https://github.com/sjanssen2/spike/. This custom pipeline is based on the python framework SnakeMake[158], which is a scalable workflow engine. It illustrates how the external SNuPy interface can be used in conjunction with other programming languages and as an endpoint for variant sequencing workflows to deliver the results to genomic researchers and experts.

a The disparity between uploaded raw data and extracted samples can be attributed to incomplete processing of datasets at the time this statistics was recorded from the production system.



(a) Variant growth with VCF file

(b) Genotype growth with samples

Figure 10.1: SNuPy variant data growth statistics. There is a steady growth in datasets over the years, with a stagnation phase during 2016, when technical problems prevented samples to be sequenced.

10.2 QUERY PERFORMANCE

To demonstrate the performance of our system, I analyzed the query log files for queries done between between August and December 2018. This revealed 708 queries send by users, excluding queries made through the programmatic interface. During this time, 12 queries exceeded the threshold of 15,000 genotypes set to prevent overloading the user with unnecessarily large result sets. The 696 successful queries are further divided into 605 regular and 91 meta queries. Meta queries allow users to query all samples of a specific species in the database. Details of the analysis are shown in table 10.1 and table 10.2[b].

A total of 605 normal queries were analyzed and during this time 43% were executed in less than 1 second, 72% within 10 seconds, and 91% returned within one minute. Only 6% of queries took longer than two minutes, which in 74% of cases (28) resulted in up to 1000 variants or more.

Table 10.1 shows that 20% of queries send by experienced users lead to empty result sets. Without SNuPy such queries would require additional coordination with bioinformatic staff to modify the criteria and rerun a query. The explorative query capabilities offered by SNuPy allows users to do these modifications themselves and retrieve a new result set.

With SNuPy users were able to focus on small result sets. When queries were successful (meaning they actually return a result > 0, 483/605), 36%(177/483) of the result sets contained less than 10 genotyped variants and 74% contained less than 250 genotyped variants, demonstrating the filter and reduction capabilities to

b Please note that when refering to relative fractions, the percentages are rounded and may not sum-up to the total of 100%.

deal with whole-exome datasets that regularly contain 80.000 variants. Still a large fraction of 25% (123/483) returns 250 and more genotypes. However, the retrieved genotypes may be detected by multiple tools in the same individual, thus the actual result size after aggregation is smaller. This can be attributed to SNuPy 's capability to query multiple samples at the same time, allowing different analysis of the same specimen to be compared and find consensus.

Interestingly, 65% of meta queries return within one second. This is likely explained by stringent filters that users apply, when performing meta queries. A typical meta query would be to identify rare variants within a single candidate gene, or a small subset of genes. As expected however, the fraction of queries returning an empty result set is smaller (12%) compared to regular queries (20%). This can be explained by the high number of samples that are queried as part of meta queries, making it less likely to return a completely empty result set. This is supported by the fact that 77% (62/80) of non-empty meta queries return 50 or less genotyped variants.

Additionally, I recorded 2,111 queries made to the query interface through the API. As expected, queries made through the API took longer on average, yet 67% finished within 10 seconds and 98% of queries finished within one minute. The fraction of empty result sets is much larger. However, of the 1076 successful queries 53%(581) returned less than 50 genotyped variants.

This analysis of queries made through a variant discovery platform reveals that users are enabled to focus on manageable and specific result sets, compared to the up to 80.000 variants, which can be expected from a whole exome sequencing analysis. Additionally, the average number of 8.7 queries per working day recorded during this time frame suggests a constant utilization of SNuPy and the need and ability for users to perform their own variant discovery.

No. genotypes | Duration	1s	10s	60s	120s	+120s	Count
∅	82	28	12	0	0	122(20%)
10	82	63	23	4	5	177(29%)
50	40	32	23	0	2	97(16%)
100	8	12	4	0	2	26(4%)
250	13	14	29	2	1	59(10%)
1000	25	7	24	6	10	72(12%)
15000	7	9	14	3	18	51(8%)
Count	257(43%)	165(27%)	129(21%)	15(2%)	38(6%)	605(100%)

Table 10.1: A summary of 605 queries recorded between August-December 2018, analyzed for the number of returned genotyped variants and filter duration. An empty list of genotypes is returned, when the filters do not yield results.

No. genotypes | Duration	1s	10s	60s	120s	+120s	Count
∅	4	5	1	0	1	11(12%)
10	29	5	0	0	0	34(37%)
50	22	1	5	0	0	28(31%)
100	2	0	2	0	0	4(4%)
250	1	0	3	0	0	4(4%)
1000	1	1	3	0	0	5(5%)
15000	0	1	2	0	2	5(5%)
Count	59(65%)	13(14%)	16(18%)	0(0%)	3(3%)	91(100%)

Table 10.2: A summary of 91 meta queries recorded between August-December 2018, analyzed for the number of returned genotyped variants and filter duration. User send meta queries to query all available samples of a species.

No. genotypes | Duration	1s	10s	60s	120s	+120s	Count
∅	125	637	267	6	0	1035(51%)
10	11	247	140	2	0	400(20%)
50	8	107	65	1	0	181(9%)
100	8	67	17	6	0	98(5%)
250	7	26	21	1	0	55(3%)
1000	5	81	26	5	0	117(6%)
15000	4	72	54	21	3	154(8%)
Count	168(8%)	1237(61%)	590(29%)	42(2%)	3(0%)	2111(100%)

Table 10.3: A summary of 2,111 queries made through the API recorded between August-December 2018, analyzed for the number of returned genotyped variants and filter duration. These queries have been sent through the API as part of larger analysis workflows.

10.3 VARIANT INTERPRETATION

As described in the requirements analysis, variant interpretation is a context and case specific process that is best performed by experts. Software platforms, such as SNuPy aid these experts in the process.

To demonstrate its utility for this task, I will look at the variant interpretation criteria for the pathogenicity of variants recommended by the ACMG-AMP[38]. More specifically, at the criteria relevant to variant datasets because some categories require other experimental setups (e.g. functional validation and transcriptome analysis).

Table 10.4 gives an overview of the criteria, that users of the guideline follow to reach a conclusion about the relevance of a variant. The following paragraphs use the same categorization summary that Richards et al. published to show how SNuPy allows its users to perform the categorization necessary for the task.

PVS1 NULL VARIANTS

Specifically damaging variant consequences e.g. nonsense, frameshift, canonical and

splice site regions, are often associated with a functional loss of the gene product. SNuPy provides variant consequences by three state of the art variant annotation tools that calculate impacts on multiple alternative transcripts. With the integration of disease-association from ClinVar and OMIM, as well as links to literature through PubMed, users have the ability to identify variants falling in this category.

PS1 SAME AMINO ACID CHANGE

Because a different nucleotide exchange can lead to the same amino acid exchange investigators should not only look for the exact same exchange, but also for the same consequence on the gene product. This is possible in SNuPy through the ClinVar module, that integrates known pathogenic variants that are co-located or neighboring to variants in patients.

PS2 PM6 DE NOVO VARIANTS

De novo variants are those that cannot be explained by inheritance and are not found in the paternal lineage. When parental samples are available, SNuPy allows to look for de novo variants and find variant falling into these categories.

PS3 BS3 FUNCTIONAL STUDIES

Functional studies require additional experiments to be carried out that do not focus on variants, but rather on the functional characteristics of a sample and are outside the scope of a variant discovery platform.

PS4 PM2 BA1 BS1 BS2 VARIANT FREQUENCY AND USE OF CONTROL POPULATIONS

SNuPy provides features to filter variants by population frequency from the population databases proposed by Richards et al. [38]. Additionally, users can obtain the variant frequency based on all samples from the database. This allows to identify possible artifacts or recurrent mutations.

PM1 MUTATIONAL HOT SPOT AND/OR CRITICAL AND WELL-ESTABLISHED FUNCTIONAL DOMAIN

When users look for variants that are part of ClinVar, they can also look for variants in the vicinity of these variants using various thresholds (direct hit, 10, 30, 100, 150 base pairs up or downstream). This allows users to identify variants that fall within the region of clinically associated variants. Furthermore, users can identify variants that fall within the functional region of a protein. Additional interpretation about the specific site and its impact on the protein is necessary and provided through linking out to generic resources (e.g. USCS genome browser), that further assessment.

PM3 BP2 CIS/TRANS TESTING

SNuPy is able to work with phased genotype data that allows investigators to determine whether a heterozygous mutation was inherited from the mother or father. Utilizing this feature enables users to establish if a working copy of a gene may still be available or not. This is closely related to identifying compound heterozygous variants, that leave a patient without a working copy of a gene, because both copies carry mutations fatal to the gene product.

PM4 BP3 PROTEIN LENGTH CHANGES DUE TO IN-FRAME DELETIONS/INSERTIONS AND STOP LOSSES

SNuPy is able to classify InDel variants as inframe or frameshift events, as well as identify stop losses through a single nucleotide variant.

PM5 NOVEL MISSENSE AT THE SAME POSITION

This is a case similar to PS1, where the same amino acid is encoded by another nucleotide. PM5 however allows categorization of novel amino acid exchanges, which is interpretable by users in the same fashion.

PP1 BS4 SEGREGATION ANALYSIS

Segregation analysis is possible in SNuPy , when multiple family members are available and the necessary controls were sequenced. However, the authors of the guidelines note that, *"If appropriate families are identified, clinical laboratories are encouraged to work with experts in statistical or population genetics to ensure proper modeling and to avoid incorrect conclusions of the relevance of the variant to the disease"*[38] . While SNuPy allows the identification of potential candidates and checking them against all samples in the database, further external case-specific analysis is necessary.

PP2 BP1 VARIANT SPECTRUM

This category is mainly based on literature review, expert knowledge and clinical data. SNuPy provides links to clinically relevant resources, such as OMIM, ClinVar, COSMIC and Pubmed, to aid users in the literature review process.

PP3 BP4 COMPUTATIONAL (IN SILICO) DATA

More than ten loss of function prediction tools are made available to users through SNuPy . This includes tools to identify non-coding regions of interest through CADD and rate splice site variants through MutationTaster.

PP4 USING PHENOTYPE TO SUPPORT VARIANT CLAIMS

This category builds upon literature review, expert knowledge and clinical data when comparing the observed and described phenotypes, that may or may not over-

lap. OMIM, ClinVar and Pubmed are useful resources for this task and provided by SNuPy .

PP5 BP6 REPUTABLE SOURCE

The use of reputable resources allows users to integrate information from external databases into their variant interpretation schema, that do not share primary data directly. ClinVar provides a transparent resource for variants that have been scored by experts with regards to their pathogenicity. Optionally, SNuPy allows the integration of specialized annotation resources.

BP5 ALTERNATE LOCUS OBSERVATIONS

This category requires expert knowledge based on literature review, clinical data and experience.

BP7 SYNONYMOUS VARIANTS

SNuPy provides computational prediction about conservation and the potential damaging affect on splice sites. Additionally, CADD scores were developed to score the impact on pathogenicity of silent mutations, which is provided by SNuPy . However, additional experiments may be necessary to make sure the criteria for this category is met (e.g. transcriptome and splicing analysis). Variant interpretation is case specific and a manual task of literature review and expertise. Thus a stand-alone variant analysis is not suitable as a singular source for a final variant interpretation. However, SNuPy empowers users with limited computer literacy, to utilize variant data and use it as a starting point for evaluation in 25 of 28 categories of the ACMG/AMP guidelines.

A total of 15 categories can be identified using only SNuPy , when the data is available as variant datasets. In further 10 categories, SNuPy empowers users to either identify variants of interest for further analysis (e.g. PP1 and BS4) or supports experts in finding relevant resource (e.g. for literature review). Three other categories are outside the application scope of SNuPy .

10.4 FEATURE COMPARISON

SNuPy is not the only tool to perform variant discovery, as already described in the state of the art. However, it is the only tool that is flexible, versatile and comprehensive to perform variant analysis in parallel on multiple samples and meet all requirements necessary to perform genomic variant analysis in a clinical research setting. Table 10.5 shows a comparison of the features supported by SNuPy and other tools that allow users to filter variant datasets.

SNuPy integrates all state of the art variant annotation tools, targeting the known problem of different transcript sets and annotation tools calculating possibly conflicting variant consequences (see Li et al., Miosge et al., McCarthy et al.).

Evidence Type	Category	ID	SNuPy	Description [38]
Pathogenic				
	Very strong			
		PVS1	+	null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease
	Strong			
		PS1	+	Same amino acid change as a previously established pathogenic variant regardless of nucleotide change
		PS2	+	De novo (both maternity and paternity confirmed) in a patient with the disease and no family history
		PS3	na	Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product
		PS4	+	The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls
	Moderate			
		PM1	*	Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation
		PM2	+	Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium
		PM3	+	For recessive disorders, detected in trans with a pathogenic variant
		PM4	*	Protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants
		PM5	+	Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before
		PM6	+	Assumed de novo, but without confirmation of paternity and maternity
	Supporting			
		PP1	*	Cosegregation with disease in multiple affected family members in a gene definitively known to cause the disease
		PP2	*	Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease
		PP3	+	Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.)
		PP4	+	Patient‘s phenotype or family history is highly specific for a disease with a single genetic etiology
		PP5	*	Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation

Evidence Type	Category	ID	SNuPy	Description [38]
Bening				
	Stand-alone			
		BA1	+	Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium
	Strong			
		BS1	+	Allele frequency is greater than expected for disorder
		BS2	+	Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected at an early age
		BS3	na	Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing
		BS4	*	Lack of segregation in affected members of a family
	Supporting			
		BP1	*	Missense variant in a gene for which primarily truncating variants are known to cause disease
		BP2	+	Observed in trans with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in cis with a pathogenic variant in any inheritance pattern
		BP3	na	In-frame deletions/insertions in a repetitive region without a known function
		BP4	+	Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.)
		BP5	*	Variant found in a case with an alternate molecular basis for disease
		BP6	*	Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation
		BP7	*	A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved

Table 10.4: A list of which ACMG [38] criteria are available through SNuPy . *SNuPy* : (+) directly available from the provided features, (*) derivable through features provided by SNuPy – may require additional resources or databases (e.g. expert knowledge, raw data, clinical record, wet lab verification), (na) not applicable for variant discovery platform, or whole exome data (e.g. transcriptome analysis)
The availability assumes, that all relevant individuals were sequenced and variants called appropriately. The table shows the categories and descriptions as published by Richards et al., the author of this work does not claim authorship of the description texts.

Its ability to also continuously integrate new, context-specific and updated resources through a modular annotation and query system ensures the software is sustainable, versatile and adaptable to other fields of research in the future. This includes annotation and links to resources associated to population and disease-specific databases, a wide range of in-silico loss of function prediction tools as well as the consequences as calculated by several state of the art variant annotation tools. Currently, no other tool defines and offers a framework that allows such modular extension of annotation resource.

This wide range of features that are already implemented and made available to users empowers them to interpret and classify variants according to ACMG/AMP[38] (for details, see above) the similar ACGS[125] guidelines for mendelian diseases. When the classification for actionable variants[121,132], somatic variant interpretation[62] and for disease specific guidelines[159] become institutionalized, the variant features in SNuPy and its extendability will allow their integration as reporting templates.

From the categorization criteria above, it becomes clear that users not only need access to a large number of annotations, but also they need to query many samples and check the presence of variants in control samples. Currently SNuPy is the only tool in the list of scientific software solutions, that is able to empower users to work with hundreds of samples in a user-friendly fashion. Users are supported by comprehensive interpretation guidance, that guides their interpretation efforts to identify important variant attributes, e.g. it makes use of colored classifications and gradients that guide attention and provides reporting templates. Also, it enables users to check the variant distribution in other samples of the database as well as its ability to display protein-protein interaction to check the functional context of a set of variants.

VarAFT was recently published, which allows to work with copy number variations, but only works locally, not allowing multiple users to cooperate and to utilize the stronger computational power provided by a server-based setup, such as SNuPy . The platform is, to the best of my knowledge, the only web platform that allows users to query copy number variation (CNV) and variant data in the same interface. Copy number variation have been associated with neurological diseases, such as Parkinson and Alzheimer, mental diseases such as Autism and Schizophrenia, Diabetes as well as infectious diseases and cancer[160,161,162,163,164,165,166]. This shows that SNuPy is of wide utility to NGS based copy number variation analysis of disease in theses contexts.

Table 10.5 displays that SNuPy is currently the only web application tool that allows to query multiple samples in parallel and aggregate the results. The variant discovery tools varianttools and Annotate-it also allow to query multiple sample, but are only usable through a command line or were only available as an online platform[c], requiring users to upload genetic patient information to a foreign site. Mendel,MD and VCF-Miner also allow to query multiple sample at once, but nei-

c Annotate-it is one of seven tools not available anymore.

ther of these tools provide the user with any possibility to document the datasets. In consequence, while users can query multiple samples it is not implemented in a fashion that allows continuous interactive analysis by multiple, cooperating users.

Although users of Mendel,MD are supported with a web interface to query their datasets, core features such as annotating uploaded variants still require a command-line interface. Thus, requiring its users to be computer literate and capable to use a command-line interface. The import and annotation process in SNuPy is triggered automatically and maintenance of the server instance and setup is performed by trained personal. VCF file format checks can be extended by developers, allowing the import of VCF data that deviate from the standard file format description and make use of its intended flexibility.

SNuPy is currently the only tool that allows the integration of its variant discovery capabilities into larger workflows. An example for this is Spike[157], which makes its results available to users of the Clinic for Pediatric Oncology, Hematology and clinical Immunology at the university clinic of Düsseldorf via SNuPy .

Name	Annotation Tools	Transcriptsets	Population Freq.	LOF Prediction	Clinical	Query	Filter Features	Variation types	Data Management	Interpretation Support	System features	Organisms
SVA		r	1o			f	g				l	
Annotate-it		e	1eo	pso	o	fm	gcio	s	sc		o	h
VAR-MD	(q)		1	ps		p	gcio				l	
VarSifter						e	gc				l	
VarB						e	g				l	
variant tools	(as)	r		pso		fm					l	
AVIA	a	r		ps		ep	g				d	hm
EVA	av	r	1eo	psc	o	ef		si	cq		o	h
GEMINI	sv		1e		c	f		si			la	h
PriVar			1e	psco		p		si	q		l	
WEP	a		1e	ps		fp	gcio	si			ld	h
AnsNGS		r	1		o	p		si			d	
FamAnn	(sv)		1e	pso		p					l	
BierApp		e	1e	ps	c	ef	gcio	si			d	h
exomeSuite			1e	pso		p		si			l	
VariantDB	as	re	1e	psc	c	ef	gcio	si	sc		d	h
wKGGSeq			1e	pso	o	fp	gcio	si	q	c	o	
VCF-Miner				p		efm		si			ld	
Var2GO	s		1e	s		f	go				o	hm
Varapp	v		1e	psc		efp	gcio	si	scr		d	
BrowseVCF	(v)					f	gci	si			l	
myVCF	(av)		1e	ps	c	f	gcio	si	cq		l	
Mendel,MD	sv	re	1e	psco	co	efm	gcio	si			d	h
VCF.Filter						e					l	
VarAFT	a	re	1eo	psco	co	f	gcio	sic			l	
SNuPy	asv	re	1emo	psco	co	efpmd	gcio	sic	scqr	cgoa	daf	hm+

Table 10.5: A list of variant discovery tools and their features compared to SNuPy, based on a literature review. Column description: *Annotation Tools*: a (Annovar), v (VEP), s (SNPEff), q (SeattleSeq), brackets indicates that pre-annotion is required *Transcriptsets*: r (RefSeq), e (Ensembl) *Population Freq.*: 1 (1000 genomes), e (ExAc or ESP), m (mouse genome project), o (other) *LOF Prediction*: p (PolyPhen), s (SIFT), c (CADD), o (other) *Clinical*: c (ClinVar), o (OMIM) *Query*: e (explore specific regions/genes), f (feature of interest), p (pre-defined), m (multi-sample), d (defaults) *Filter Features*: g (gene), c (consequence), i (inheritance), o (other) *Variation types*: s (SNV), i (Indel), c (CNV) *Data Management*: s (sample documentation), c (collaborative), q (quality control), r (reporting template) *Interpretation Support*: c (classification), g (gradients), o (other), a (ACMG-AMP) *System features*: l (local), o (online only), d (central database), f (framework), a (API) *Organisms*: h (human), m (mouse), + (expandable)

11

CASE STUDIES

11.1 ALPS

The autoimmune lymphoproliferative syndrome (ALPS) is a genetic disease of T-cells, caused by a dysfunctional FAS death receptor signaling pathway, which triggers cell apoptosis in its normal state[120]. The lack of this function causes defects during lymphocyte development and results in an accumulation of T-lymphocytes in lymphoid organs, which can develop into wide range of hematological diseases, including malignancy[167].

Currently five ALPS classes (Table 11.1) are defined, which affect the FAS death receptor[a], its ligand FASLG[b] or the disruption of the downstream protein caspase-10 (CASP10)[c], responsible for catalyzing the apoptotic signal[120].

Although the disease is characterized by well defined criteria (see Oliveira et al., Table 11.1), affecting three gene products, the true incidence rate remains unknown[169]. This may be explained by a high rate of ALPS-like cases that present similar phenotypes as ALPS, but have to be classified as *ALPS-U*(*ALPS III*), or are classified differently because of RAS mutations, known to result in similar phenotypes[120].

Consequently, for 20-30% of ALPS-diagnosed cases the genetic cause remains unknown[170]. This makes interactors and regulators of the known candidate genes a potential target for cases that carry an ALPS-like phenotype, but no mutation can be found in any of the three ALPS candidate genes (FAS, FASLG and CASP10)[120]. The rationale is that the dysregulation of interactors produces effects and phenotypes similar to those of a dysfunctional primary protein.

11.1.1 *Deregulation of Fas ligand expression as a novel cause of autoimmune lymphoproliferative syndrome-like disease*[120]

Nabhani et al. (2015) investigated 20 patients, previously diagnosed as ALPS-U, using whole-exome sequencing and employing SNuPy to perform the analysis. The authors identified a patient carrying a truncating nonsense mutation in Interleukin 12 Receptor Subunit Beta 1 (IL12RB1) resulting in loss of crucial IL12 signaling[120].

95,794 variants were identified in the patient using whole-exome sequencing of which 5,394 affected the FAS network and its interactors in a protein-protein interaction network (STRINGdb[147]). This number was further reduced by filtering

a protein name: Fas or CD95
b protein name: FASL or CD95L
c protein name: Casp-10

Previous nomenclature	Revised nomenclature	Gene	Definition
ALPS type 0	ALPS-FAS	FAS	Patients fulfill ALPS diagnostic criteria and have germline homozygous mutations in FAS.
ALPS type Ia	ALPS-FAS	FAS	Patients fulfill ALPS diagnostic criteria and have germline heterozygous mutations in FAS.
ALPS type Im	ALPS-sFAS	FAS	Patients fulfill ALPS diagnostic criteria and have somatic mutations in FAS.
ALPS type Ib	ALPS-FASLG	FASLG	Patients fulfill ALPS diagnostic criteria and have germline mutations in FAS ligand.
ALPS type IIa	ALPS-CASP10	CASP10	Patients fulfill ALPS diagnostic criteria and have germline mutations in caspase 10.
ALPS type III	ALPS-U	Unknown	Patients meet ALPS diagnostic criteria; however, genetic defect is undetermined (no FAS, FASL, or CASP10 defect).

Table 11.1: ALPS classifications according to Oliveira et al.. Six classes are currently used to classify ALPS cases. Suspected cases are commonly classified as ALPS type III.

for variants affecting the protein products (534), as these variants are most likely to have an effect on the protein interaction partners. 16 homozygous variants were detected that are not detected in the general population, utilizing data from the 1000 genomes project[3], HapMap[171], Exome variant server[d], alongside 300 in-house whole exome datasets as shared controls. Further filtering for *autosomal recessive* genetic inheritance scenario, using variants from the two parents reduced the number to 6 variants. The functional context of these last six variants revealed a stop gain mutation affecting IL12RB1, a direct high-confidence interactor of FASLG, the known cause for ALPS-FASLG. Further functional studies were conducted that confirmed IL12RB1 as a relevant candidate for an ALPS-like phenotype[120].

d http://evs.gs.washington.edu/EVS last visit 5th Sep 2018

This rigorous systematic filtering schema was developed by genomic experts during multiple iterations and query refinements, adding increasingly specific criteria of increasing complexity using SNuPy . Criteria were combined using consequence predictions, variant quality, genotype and its quality, inheritance patterns taking multiple samples into account, custom shared controls and population frequencies as well as protein-protein interaction data. It showcases how genomic researchers are enabled to find relevant and novel disease mechanism if they are empowered to refine their initial query by themselves.

11.1.2 *STAT3 gain-of-function mutations associated with autoimmune lymphoproliferative syndrome like disease deregulate lymphocyte apoptosis and can be targeted by BH3 mimetic compounds*[172]

Using SNuPy , Nabhani et al. (2017) analyzed 30 children with ALPS-U diagnosis for pathogenic variants and possible treatment options. After classical ALPS mutations in FAS, FASLG and CASP10 could not be detected by Sanger sequencing, whole exome sequencing was performed on the patients as well as parents and siblings (where it was possible and consent was given)[172].

Two patients in this cohort, one with consanguineous parents (patient 1) and one with non-consanguineous parents (patient 2), harbored variants in the signal transducer and activator of transcription 3 (STAT3), which regulates the gene expression of apoptosis, proliferation and cell growth factors. Consanguinity status is an important heritage factor, when analyzing samples for possible Mendelian disorders, due to increased loss of heterozygosity. Consequently, SNuPy was used to filter variants of the patients, with regard to this clinical information.

The identified mutations were found to have a gain-of-function (GOF) effect on STAT3, leading to hyperactivity of the protein. In case of STAT3 this leads to a decrease of FAS expression, therefore mimicking a loss-of-function of FAS that is the cause of ALPS-FAS phenotype[172].

With regards to the gain-of-function detection, by loss-of-function prediction tools: As described by Flanagan et al. gain-of-function may have more subtle effects and therefore are not expected to be detected with the same sensitivity as loss-of-function mutations by PolyPhen2 and SIFT. However, the functional validation of the STAT3 mutations highlights the importance of in-silico loss-of-function prediction tools and show how they can aid in the search and interpretation of gain-of-function mutations, when they are thoroughly and independently validated.

For both patients the analysis was focused on variants covered by at least 10 reads, the variant base quality was larger than 30[e], the frequency in the general population does not exceed 10^{-5}, the missense consequence is predicted to be deleterious by PolyPhen2[58] or SIFT[59].

e The variant base quality is a phred-scaled measure to indicate the confidence in the presence of a variant

A key difference in the filtering strategy of the two patients, lies in the incorporation of available controls into the filtering criteria. For patient 1, parents and healthy siblings were used as controls, resulting in two candidate mutations. For patient 2 on the other hand, whole-exome data for the father was not available and only variants from the mother and another healthy sibling could be used, resulting in 27 potential candidates.

The follow-up analysis of the protein-protein interactions from STRINGdb[147] further showed that FASLG and STAT3 have a described interaction that further strengthened the interest to validate the role of STAT3 as potential cause for ALPS-like phenotype.

Compared to the discovery of IL12RB1 (described above) this study displays the necessity to adapt filter strategies with regards to the actual available data. A flexible variant discovery system, such as SNuPy allows its users to make the best out of possibly sub-optimal starting position, such as incomplete parent-offspring trios.

Criteria:

11.2 A NOVEL APPROACH TO DETECT RESISTANCE MECHANISMS

So called 'epidrugs' have emerged as possible treatment tools over the last decade. These drugs "are defined as drugs that inhibit or activate disease-associated epigenetic proteins to ameliorate or cure the disease"[23] and are commonly used either in mono-therapy or in combination with other anti-cancer drugs during treatment.

Suberoylanilide hydroxamic acid (SAHA) (Vorinostat, Zolinza©(Merck and Co., Inc.)) is one of these drugs and the first to be approved by the American Food and Drug Administration (FDA) in 2006, as part of the treatment for patients with cutaneous T-cell lymphoma[174].

The use of this drug for the treatment of Non-Hodgkin lymphoma (NHL) is currently an area of active research[23]. NHL is a group of lymphoproliferative neoplasms that contains two aggressive subtypes: Diffuse large B-cell lymphoma (DLBCL) and Burkitt lymphoma (BL). Using SAHA monotherapy, remarkably 30% of patients reached complete remission[23], leaving however 70% of patients without response during treatment. Identifying genetic factors that can act as potential biomarkers for tumor resistance against SAHA was the focus of our study[23].

To identify possible genetic factors of drug-resistance towards SAHA we designed a drug efficacy testing with exome and captured target analysis (DETECT) method. This method combines (1) SAHA sensitivity test using flow cytometry (Annexin V/propidium iodide staining), (2) whole-exome variant analysis, (3) Capture Compound Mass Spectrometry (CCMS, *see below*) with (4) an integrated network analysis of the identified candidates.

The analysis used 26 commercially available B-cell lymphoma cell lines that were sequenced using whole-exome sequencing and tested for their drug sensitivity, which is meased as an *IC50* value. *IC50* indicates the *half maximal inhibitory concentration*, a molecular concentration measure that needs to be exceeded to inhibit a molecular process or function in-vitro. Higher values mean that a higher drug concentration needs to be achieved during treatment, which is restricted by the average plasma concentration feasible for patients (e.g. due to toxicity). An achievable plasma concentration of 2.5μM was reported for SAHA[175] and build the base of the classification of the 26 cell lines into resistant, intermediate and sensitive cell lines (Table 11.2).

Based on this SAHA resistance distribution over the 26 cell lines, 8 sensitive and resistance cell lines were chosen for CCMS measures with SAHA.

Capture Compound Mass Spectrometry (CCMS) uses a trifunctional small molecular probe called Capture Compounds (CC)Köster et al.. Using replication, competition and control samples this method allows to identify SAHA interaction partners. Due to the possibility of capturing protein complexes with this CC-based approach it is also possible to identify indirect binding proteins.

SAHA	Cell line	IC50 (μM SAHA)
Resistant	CA-46*	470.40
	Daudi*	282.60
	DG-75*	117.70
	Raji*	19.74
	Blue-1	11.72
	Namalwa	7.00
	DND-39	5.76
	HT	4.63
	BL-70	3.93
	Carnaval	3.56
	WSU-DLCL2	3.52
Intermediate	Kis-1	2.49
	SU-DHL-4	2.27
	OCI-Ly7	2.15
Sensitive	OCI-Ly2	1.98
	WSU-FSCCL	1.82
	HBL-1	1.62
	U2932	1.61
	SU-DHL-6	1.48
	Granta-452*	1.28
	BL-2	1.18
	OCI-Ly3	1.14
	BL-41*	1.13
	TMD8*	0.96
	OCI-Ly10	0.61
	OCI-Ly1*	0.50

Table 11.2: Cell line classifications based on their resistance towards SAHA exosure (by Joosten et al.). Cell lines that were analyzed using CCMS are marked with *.

In total we compiled a list of 315 candidate proteins interacting with SAHA directly or indirectly. Among the candidates, previously described SAHA targets such as HDAC1, HDAC2, HDAC3, HDAC6, HDAC8, members of the HDAC1/2 complexes, as well as ISOC1 and ISOC2 were identified. HDAC proteins are epigenetic modifiers and expected targets of SAHA. Of the 315 proteins, 117 and 12 were enriched in resistant and sensitive cell respectively.

To visualize the functional context of the significant SAHA binders we extracted their protein-protein interactions (STRINGdb[147]), revealing 125 proteins that interact with at least one other protein in the group.

The largest connected component (62 proteins) of this SAHA activity network harbored membrane associated and cytoplasmic proteins, mainly consisting of Src kinases (FGR, HCK, LYN), G-Proteins and their regulators (GNAI1, GNAI2, GNAI3, GNG7, GNAZ, GNAS, GNB1, RGS14 and RGS19). The association to nucleus-associated HDAC complexes is established through the STAT-pathway.

Using SNuPy external API we integrated the consequence predictions into the network to visualize proteins and interactions. Subsequently proteins whose genes are hit by a more severe mutation in resistance compare to sensitive cell lines, were highlighted (severity was derived by the consequence impact rating of the Ensembl analysis group[76]) (see Figure 11.1).

This analysis revealed the genes FGR, RGS14, and PAG1 to be mutated more severely in resistant cell lines, while also showing higher binding affinity compared to sensitive cell lines. In total 4 of 11 resistant cell lines (Blue-1, DG-75, HT and WSU-DLCL2) harbored mutation in these three genes. FGR showed the highest affinity for SAHA binding ($p = 9.6e-07$), affecting three resistant cell lines (Blue-1, DG-75 and HT) and was identified as the most likely candidate to characterize the SAHA binding and to use its expression as potential biomarker for SAHA resistance.

Interestingly FGR showed higher expression patterns in SAHA resistant cell lines, compared to sensitive cell lines and also higher expression in B-cell lymphoma compared to non-lymphoid entitiesJoosten et al.. We then performed CRIPR/Cas9 knockout of FGR in three SAHA resistant cell lines (Raji, Daudi and DG-75). For cell lines not carrying a FGR mutation in their exome, this knockout reveled an increase in SAHA sensitivity. DG-75, carrying a c.1459/C>A mutation in its tyrosine kinase domain, remained resistant despite of the knockout and 85% integration of Indels by CRISPR/Cas9.

FGR expression analysis of 1200 B-cell lymphoma patients demonstrated a high variance, while FGR mutations were rare, suggesting that FGR expression is more suitable to identify SAHA resistant cell lines. In conclusion, this study successfully used the DETECT method to unravel FGR as a potential biomarker to stratify SAHA resistance in B-cell lymphoma patients.

Using SNuPy we first evaluated if a systematic difference in the mutational profile can be found between SAHA-resistant and sensitive cell lines. Multiple iterations and filter strategies were developed. This included to identify overlapping and distinct mutations and genes between the two sample groups. We further incorporated additional shared control samples to better identify sequencing artifacts. The use of gene panels allowed us to also focus on genes that are know to modify the epigenetic landscape of a genome. However, these strategies did not yield conclusive results from stand-alone whole-exome variant analysis.

This study demonstrates the need for multiple complex experimental protocols and verifications steps that are necessary to identify potential biomarkers for drug resistance. My personal contribution to this work was the analysis of whole-exome datasets, the development of an data analysis strategy as well as the integration of the results to a protein-protein interaction network. The workflow incorporates IC50 drug sensitivity values and variants from whole exome analysis for 26 cell lines, with CCMS measures from only eight cell lines. Integrating whole exome variants

with the findings from CCMS and resistance measures allowed the identification of interacting protein networks that we believe contributes to SAHA resistance.

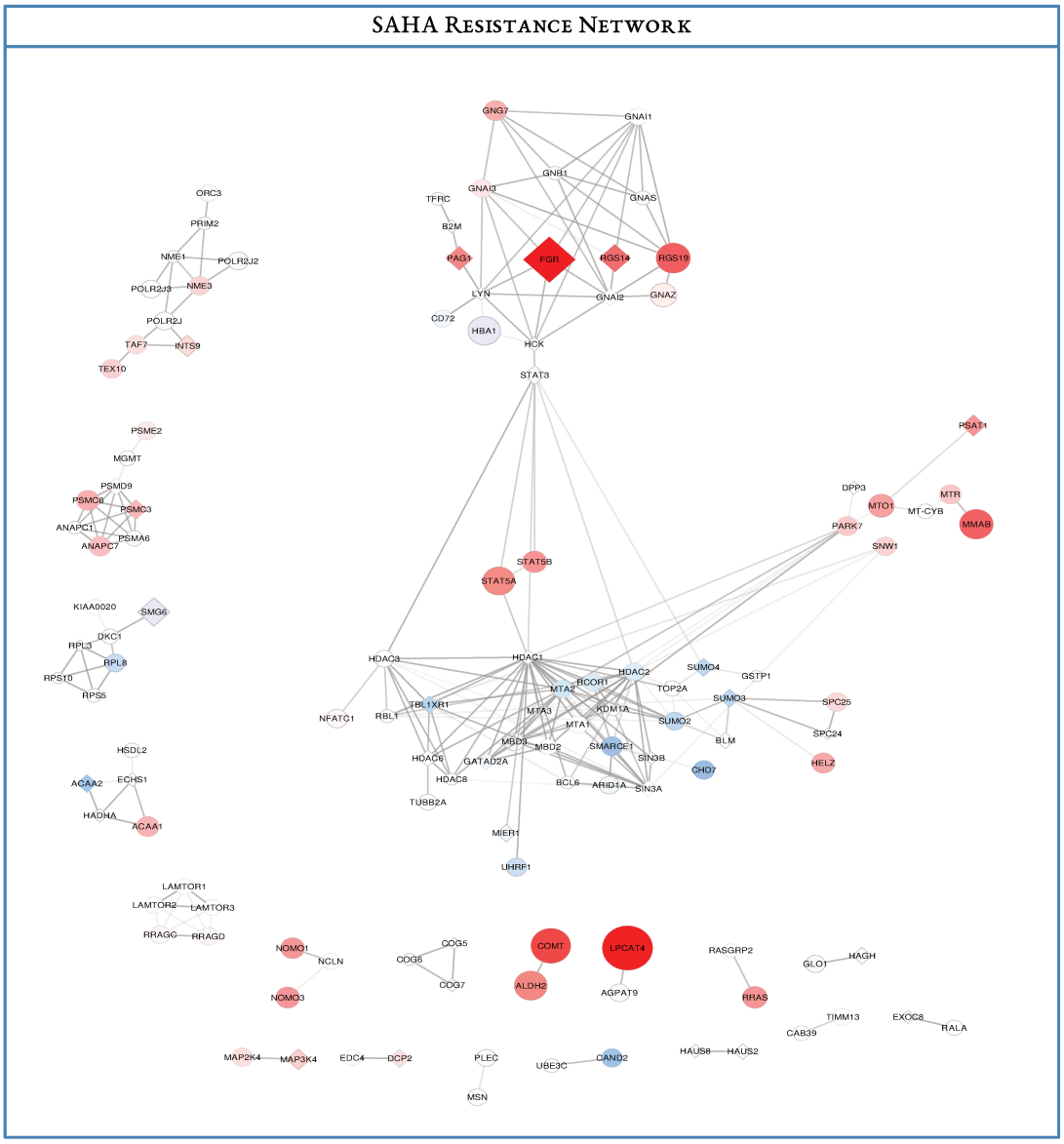


Figure 11.1: *"The integrated protein-protein interaction network based on CCMS capture proteins and whole exome sequencing data reveals 20 connected components. Node color and size indicate fold changes towards SAHA resistant (red) or sensitive (blue) cell lines in the CCMS capture experiment, white nodes indicate no relevant fold change (logFC < 1), the transparency expresses the p-value. Diamond shaped nodes mark gene products that carry a more detrimental mutation in any of the SAHA resistant cell lines. Edge opaqueness expresses protein interaction confidence. Data were visualized using Cytoscape."* [23] (Figure from supplemental material S3* Joosten et al.)

*A novel approach to detect resistance mechanisms reveals FGR as a factor mediating HDAC inhibitor SAHA resistance in B-cell lymphoma/Joosten et al./Molecular Oncology/Vol. 10/Issue 8 Published under a Creative Commons Attribution (CC BY) License [23]

11.3 EXOME SEQUENCING OF PEDIATRIC ALL RELAPSES AFTER ALLOGENEIC SCT

Acute lymphoblastic leukemia (ALL) is the most prevalent pediatric cancer and with a 5-year survival rate of approximately 90% in developed countries[177], can be treated effectively. However, when patients do not respond to the first treatment and relapse they have much poorer prognosis. One treatment option for these children are allogeneic hematopoietic stem cell transplantation (allo-SCT), which show 8-year survival rate of up of 70%[178].

There are however, no standard protocols for cases of early post-allo-SCT relapses available, and treatment options are limited, due to cumulative toxicity. For some children disease-free survival rates of 30% after two years could be achieved with subsequent SCT, when remission and low or negative minimal residual disease (MRD) level was achieved[179]. For this reason novel approaches are urgently needed to achieve complete remission from a post-allo-SCT relapse that evaded the chemotherapeutic and immunologic pressure from previous treatment.

Gröbner et al. have recently shown that 52% pediatric cancers (n = 675) harbor potentially druggable events (PDE), and 37% (n = 41) retain these during progression to relapse. This highlights the potential for druggable targets and possible individualized treatment options, for cases with relapse after SCT.

To address the clinical need for identification and description of patient-individualized treatment options, we used whole-exome sequencing on ten pediatric patients, who suffered from post-allo-SCT relapses and were enrolled in a multicenter pediatric ALL trial as part of the "Individualized Therapy for Relapsed Malignancies in Childhood" (INFORM)[10] project. The average age at first diagnosis was 4.6 years (range: 0.3-10.2), average time to first relapse was 2.2 years (range: 1.1-3.4), and average time to post-allo-SCT relapse was 0.7 (range: 0.2-1.7 years). Seven patients received transplants from matched-unrelated donors, two from siblings and one from his mother. At the time of the analysis only two patients were alive, illustrating the dismal prognosis of post-allo-SCT relapses.

As previous approaches have shown, it is important to use more than one somatic mutation identification pipeline when analyzing WES samples[181]. Thus we combined the variant calls of MuTect[155] and VarScan2[151] to find the mutations comprising the oncogenomes during tumor progression. These oncogenomes[f] are defined as:

ONCOGENOME1 : INIT-REMI [OG1]

ONCOGENOME2 : RLPS-REMI [OG2]

ONCOGENOME3 : (TRLPS-TREMI) ∩ (TRLPS-REMI) [OG3]

[f] For a description of the semantics behind the state identifiers see Figure 7.2

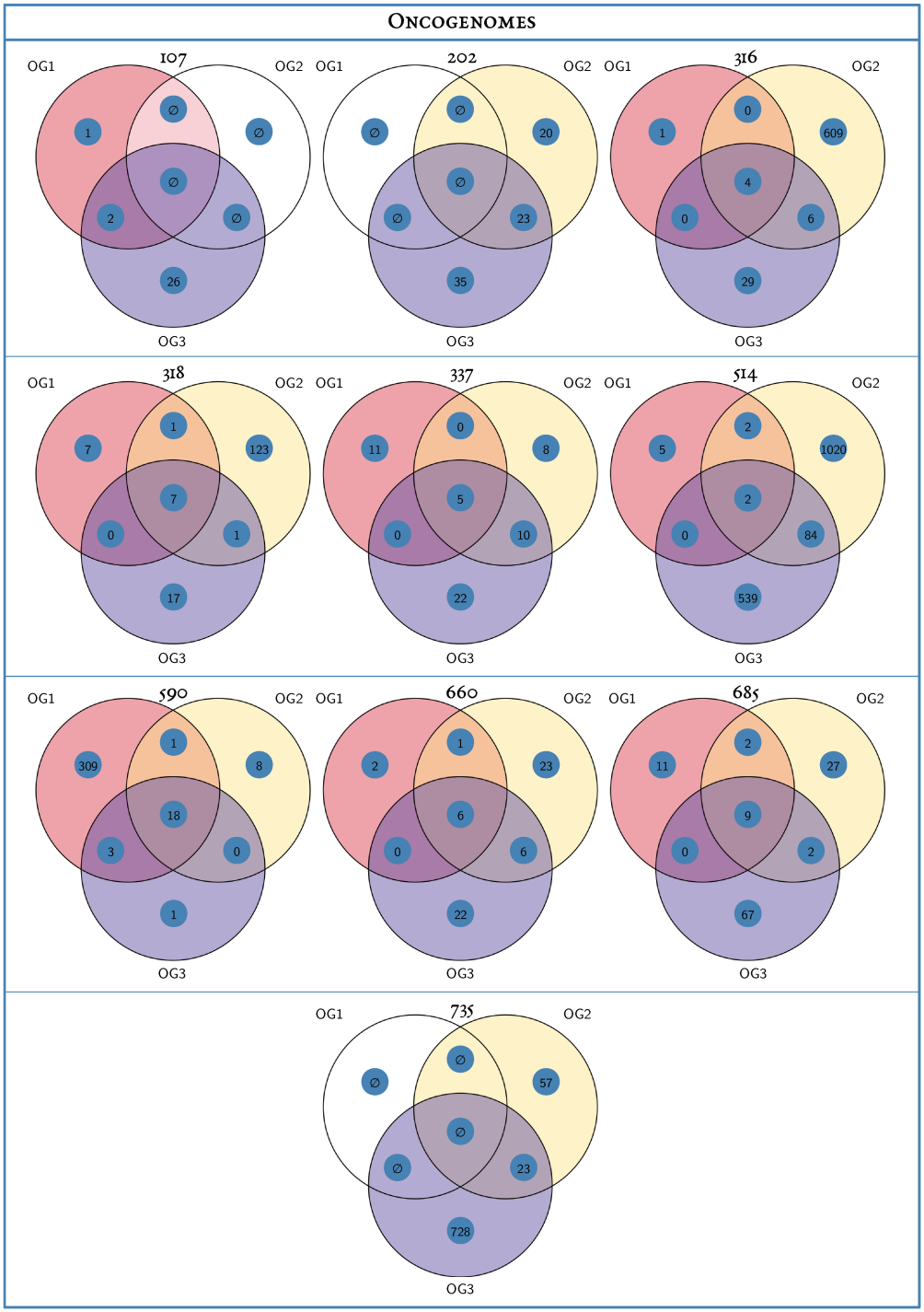


Figure 11.2: Overlaps of affected genes in 10 patients with post-allo-SCT ALL. White oncogenomes display incomplete datasets (missing controls or tumors).

The venn diagrams (see Figure 11.2) reveal that all studied patients carry mutations in genes that survive the treatment process through all of it's stages, ranging between 2 and 18 genes.

Median number of somatic variants in OG1, OG2 and OG3 were 13.5, 50 and 55.5, respectively. Individually the sizes varied substantially and stayed below 200, with one outlier in OG1, and two for OG2 and OG3. These outliers carried acquired variants in DNA polymerase or DNA repair/Fanconi anemia genes, possibly explaining the hypermutator phenotype. The oncogenomes of other patients did not carry mutations in these genes.

Although a stable core set of genes seems to manifest in these patients, some genes, such as NT5C2, a gene described to drive therapy resistance, were absent in all OG1, while 3/10 patients carried somatic mutations in OG2, disappearing in OG3[182].

We applied density-based clustering[183] to the log-scaled variant frequencies of the tumors to identify possible co-evolution of clones during tumor progression. This analysis reveals several clusters per patient, which undergo a similar tumor progression. Surprisingly, mutational clusters containing known cancer associated genes are as variable as clusters not containing these genes, hinting at different drivers for each tumor (data not shown, manuscript submitted).

The rigorous and flexible sample annotation schema, as well as the dynamic sample state representation (see Figure 7.2), allowed for a systematic analysis of 53 whole-exome sequencing datasets from 10 patients, resulting in 201 variant datasets. Because the study was conducted as part of a national multicenter trial, samples from different locations with different identifiers required a flexible data management and quality control to allow consistency and integrity checks.

The pre-defined query feature, as well as the dynamic queries and the external SNuPy interface enabled the integration of all somatic variants from all stages and the exploration of the patients mutational landscape. The ability of SNuPy to analyze hundreds of samples in parallel and dynamically query large number of variant datasets was fundamental for this study, that required explorative queries of more than 200 variant datasets and the integration of multiple somatic callers. Furthermore, statistical analysis (e.g. oncogenome sizes and overlaps) was carried out utilizing the SNuPy R package, the state of the art framework and programming language to target such statistical tasks.

Details of this work have been submitted for publication under the title "Pediatric ALL relapses after allo-SCT show high individuality, plasticity, selective pressure and druggable targets" to the Blood advances journal.

11.4 OTHER CASE STUDIES

The publications and projects described above describe some highlights of the usage of SNuPy which has also been used as part of other publications that are summarized in this section.

Next-generation-sequencing-based risk stratification and identification of new genes involved in structural and sequence variations in near haploid lymphoblastic leukemia. (Chen et al. (2013))
Whole-genome sequencing provides a feasible tool to distinguish high hyperdiploidy and near haploidy samples. Whole exome variant analysis using SNuPy was employed to gain insight into tumor promoting events, contributing to near haploidy in five patients. For this task the mutational landscape of relevant mutations was described, with a focus on the difference between a new haploid cell line and high hyperdiploid samples from patients. Utilizing SNuPy genomic experts were able to analyze the sample of complex chromosomal compositions.

Whole-genome paired-end analysis confirms remarkable genomic stability of atypical teratoid/rhabdoid tumors (Hoell et al. (2013))
This report analyzes two patients suffering from atypical teratoid/rhabdoid tumors (AT/RT) and reports on the somatic mutations, which have been identified by tumor-normal matched sequencing. SNuPy allowed clinicians to evaluate and analyze the exomes and somatic mutations and independently confirm previous reports of the role of SMARCB1 mutations in this type of brain malignancy.

Combined immunodeficiency with life-threatening EBV-associated lymphoproliferative disorder in patients lacking functional CD27 (Salzer et al. (2013))
This study attempted to establish if CD27 mutations are causative for a EVB-associated lymphoproliferative disorder. Previous analysis of Patient 1 identified a homozygous c.G158A mutation in CD27, that was shared in a homozygous state among siblings and found to be heterozygous in parents with consanguine background. SNuPy was used to analyze the whole-exome results of two more families and allowed the identification of the same mutation in two more patients. It allowed rapid verification of a hypothesis, previously derived from another experiment.

Constitutional Mismatch Repair-deficiency and Whole-exome Sequencing as the Means of the Rapid Detection of the Causative MSH6 Defect (Hoell et al. (2014))
Studying the feasibility to use whole-exome sequencing as a tool to diagnose constitutional mismatch repair-deficiency (CMMR-D), linked to increased risk of childhood hematologic malignancies, brain tumors, colorectal cancers and other rare malignancies. It is caused by mutations in MLH1, MSH2, MSH6 and PMS2, which are essential to the DNA replication control and repair process. Rapid diagnosis is necessary to tailor therapeutic factors, such as radiation exposure. SNuPy was used by clinicians to explore and interpret the variations of the patient, parents and siblings. Without additional bioinformatic support the trained clinicians were able

to identify two homozygous MSH6 mutation in the patient among more than 200,000 variants detected by whole-exome sequencing.

Inherited susceptibility to pre B-ALL caused by germline transmission of PAX5 c.547G>A (Auer et al. (2014))
This study confirmed a previous report PAX5 to confer an inherited susceptibility to pre-B-ALL. For this the pedigree of two children of the families of two brothers were sequenced. SNuPy was used to identify 518 mutations, common between the children and both fathers, that made up the initial list of candidates. The three affected children, and the two brothers shared a transmitted PAX5 mutation. Because of the presence of the same mutation in other healthy sibling, we conclude that the mutation itself may have reduced penetrance. Suggesting secondary factors to trigger the development on pre B-ALL.

Infection Exposure Is a Causal Factor in B-cell Precursor Acute Lymphoblastic Leukemia as a Result of Pax5-Inherited Susceptibility (Martin-Lorenzo et al. (2015))
This study presents in-vivo genetic evidence for the hypothesis that infections by common pathogens are causative for pre B-ALL. Pax5 positive mice developed pre B-ALL when removed from specific pathogen-free (SPF) environment. There was significant difference to Pax5 negative mice that were exposed to the same conditions. SNuPy supported the variant analysis and exploration of the mice samples, by providing access to protein consequence prediction for mouse transcripts and SIFT scores. This allowed to reduce the large number of whole exome variants of three mice (between 27,000 and 31,000 mutations) and identify additional secondary hits in known oncogenes Il7R, Jak3 and Stat5. A recurrent Jak3 mutation in two mice was found and further sanger sequencing reveals that 6 of 9 other mice carry non-synonymous Jak3 mutations as well.

Genomics and drug profiling of fatal TCF3-HLF-positive acute lymphoblastic leukemia identifies recurrent mutation patterns and therapeutic options (Fischer et al. (2015))
TCF3-HLF-fusion positive acute lymphoblastic leukemia (ALL), an ALL subtype was analyzed using a multitude of NGS and screening technologies, including whole-genome, whole-exome, whole-transcriptome, drug profiling using patient and xenograft models. This highly cooperative project was carried out by a national consortium of 25 institutions. In this consortium a specific annotation release (Ensembl v70) was used as a basis for integration and SNuPy provided the whole-exome variant analysis using this version. This was achieved using the flexible variant annotation system that the AQuA framework provides in SNuPy . Quality control measures in SNuPy was also fundamental to ensure that samples of the study were not contaminated.

Next-generation-sequencing of recurrent childhood high hyperdiploid acute lymphoblastic leukemia reveals mutations typically associated with high risk patients (Chen et al. (2015))
This study investigated the factors contributing to recurrent high hyperdiploid

ALL. In total 5 patients, with tumor, germline and relapse samples were investigated. Additionally, samples from 24 other ALL-associated patients were used as controls to find mutations and affected genes specific to the recurrent phenotype. SNuPy allowed to use previously analyzed samples to be integrated in this project, allowing more specific description of separating factors. This highlights the usefulness of a central database, making studies available to all investigators of an institution to integrate them into their filter strategies.

Fatal Lymphoproliferative Disease in Two Siblings Lacking Functional FAAP24 (Daschkey et al. (2016))
Epstein Barr virus (EBV) infections are often asymptomatic in children and are met with a vigorous immune response in health individuals. However, there are cases when immunodeficiencies lead to a response failure, causing a severe lymphoproliferative disorder. A family with consanguine background, from which two children died from progressive EBV-associated lymphoproliferative disease was studied, including two healthy siblings. SNuPy supported the comparison of the affected and unaffected children and the identification a recurrent homozygous mutation in Fanconi Anemia Core Complex Associated Protein 24 (FAAP24).

Specific antibody deficiency and autoinflammatory disease extend the clinical and immunological spectrum of heterozygous NFKB1 loss-of-function mutations in humans (Schipp et al. (2016))
This report investigates two patients with severe autoinflammatory disease. Both patients carry mutations in the nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (NFKB1), which resulted in significantly reduced expression of this immune and inflammatory response regulator. SNuPy was used to analyze the patients whole exome variant datasets and identify candidate mutations that broaden the spectrum of NFKB1-associated phenotypes.

A novel homozygous mutation in UNC13D presenting as Epstein-Barr-virus-associated lymphoproliferative disease at 9 years of age (Bienemann et al. (2016))
This study, as the study from Daschkey et al. (2016) described above, studied a patient with EBV-associated hematologic disease. The child was born to parents with consanguine background and was analyzed based on this hypothesis. SNuPy enabled the identification of a recessive UNC13D variant.

Human RAD52 - a novel player in DNA repair in cancer and immunodeficiency (Ghosh et al. (2017))
This case report presents an 18-year old man with immunodeficiencies, EBV-lymphoproliferative disease and chemosensitivity of fibroblasts. Fibroblast is a type of connective tissue with mesenchym origin and is commonly used as germline control to identify tumor specific mutations. The chemosensitivity of this tissue from this patient resembled that of Fanconi anemia (FA), an autosomal disorder affecting the DNA repair mechanisms during cell division and association to increased cancer risk. This circumstance and the family history of malignancies prompted the investigators to

search for factors of inherited diseases, by sequencing the father, sister and stored samples of the diseased mother, who died early of breast cancer carcinoma. Using SNuPy a total of 122 variants were detected, of which three could be excluded using deduction from clinical data. Only one mutation in RAD52 homolog, DNA repair protein (RAD52) was detected to be private to the patient and mother after comparison to database-wide available control samples. This massive reduction of possible candidates highlights the importance of institution wide control samples.

Loss of Pax5 Exploits Sca1-BCR-ABL p190 Susceptibility to Confer the Metabolic Shift Essential for pB-ALL (Martín-Lorenzo et al. (2018))
Preleukemic cells are cells, which later develop into full leukemic cells, such as precursor B-cell acute lymphoblastic leukemia (pre B-ALL). These preleukemic clones can be found early in neonatal cord blood, e.g. harboring BCR-ABL^{p190} (BCR-ABL) lesions. However, these lesions often remain silent and fraction of normal B-cells in healthy adults harbor them without developing into pre B-ALL. This study set to investigate factors contributing to tumorigenesis of BCR-ABL positive cells. For this task sixteen tumors, 3 regular Sca1-BCR-ABL and 13 Sca1-BCR-ABL+Pax5$^{+/-}$, as well as the respective germline samples were sequenced using whole-exome sequencing. This revealed Pax5 and Jak3 to be recurrently mutated[g] in Sca1-BCR-ABL+Pax5$^{+/-}$ pre B-ALL, while Sca1-BCR-ABL were more heterogeneous. SNuPy was used to identify and describe the mutational landscape in these two different classes, which was enabled by allowing multiple samples to be queried in parallel.

g see also *Infection Exposure Is a Causal Factor in B-cell Precursor Acute Lymphoblastic Leukemia as a Result of Pax5-Inherited Susceptibility (Martin-Lorenzo et al. (2015))* described above

Part VII

CONCLUSION

12

CONCLUSION, DISCUSSION AND FUTURE WORK

Next generation sequencing technologies are widely seen as a driving force of fundamental changes in medical research, diagnosis and treatment in the coming years. What has started in the late '00 years is now becoming more and more accessible in terms of financial and timely investment. With the simultaneous increase in depth and breadth that genome wide sequencing offers data accessibility is becoming a more and more pressing issue.

These developments are not unique to the medical sector as the field of data related sciences is beginning to emerge in production and service-driven industries as well. But despite the success of data science applications, it is not always met with acceptance in the general population. The health-care aspect of medical NGS offers great benefits to increase acceptance for this field by offering better treatment and prevention through precision medicine utilizing data-driven processes.

However, data literacy requires tools that supports experts to access large datasets, both digitally and mentally and allow them to interpret their findings. SNuPy is a tool that empowers genomic scientists to do exactly this and lets users with limited computer literacy become data literate.

Bio-medical research is a rapidly evolving field and NGS is one of its current driving forces. There are many data sources available, many of them are specifically designed to focus on a single aspect of variants or diseases. Therefore, it is fundamental for any variant discovery platform to be extendable and integrate new data sources in a flexible fashion. Although most genomic information is gathered and published for the human genome, model organisms are important subjects of genomic studies as well. Consequently SNuPy enabled, users to not only work with human genomic data, but also with those of other organisms.

The Clinical Genome Resource project[a] aims to built a curated knowledge base for clinical genomics. Standardized variant interpretation criteria have been recognized by this project as a fundamental part of reproducible and coherent variant classification by experts. Thus variant discovery platforms need to be able to support users in complying with such standardized approaches, which SNuPy does.

As I have demonstrated SNuPy was used with great success for different research projects on varying topics (see chapter 11). The underlying AQuA framework concept allows sustainable and flexible future developments that can be adapted to many research tasks that analyze genome-wide variant datasets. It was successfully used to integrate the necessary data sources (see section 9.1) that allow standardized variant interpretation and explorative analysis (see section 10.3).

a https://www.clinicalgenome.org

As we have seen when looking at other solutions that provide a subset of the necessary variant discovery features, there is no clear state-of-the-art concept on how to store variants and the variant annotation data. The VCF standard defines a data exchange format and allows position based retrieval of annotations, but falls short for example when one needs to query a dataset by gene name. GEMINI, a popular command line based filtering tool uses a approach using file based relational database, utilizing indexing on arbitrary attributes. SNuPy utilizes a remote relational database to store variants and the associated raw variant data, utilizing the increased computational power of a centralized setup. The database size has grown significantly over the years, harboring more than 5.000 variant datasets with more than 300 million genotypes. Despite the large volume, SNuPy manages to query datasets in a time frame that allows dynamic and explorative work (section 10.2), thus providing another successful example of storing and managing variant datasets in a relational database management system.

With use of SNuPy in research and the overall increasing number of publication in this field, the next step is to utilize the knowledge that can be derived from the results. One way an integrated variant discovery platform can support variant interpretation efforts is to identify recurring aspects of query results. For example, analyzing the results of the queries of a project to identify patterns in the returned result that are not apparent from a single query, but follow a pattern (e.g. biological function) that show up more frequently as part of the result than expected. Using expert knowledge to drive this process will allow to learn important aspects of disease causing variants. Moreover, variant filtering and interpretation standards may benefit from such observation-based approaches, in order to find new criteria, refine existing ones or weight individual aspects more precisely.

I will be working on the extension of SNuPy by utilizing the existing AQuA framework to integrate new data sources, such as those published by the clinical genome project and disease specific databases. While SNuPy in its current state allows quality control of individual samples, I further plan on extending the quality controls capabilities to a database wide level, specifically to ensure that data is correct across all project and no NGS samples have been swapped. One promising way to address this are local sensitive hashing algorithms that allow to calculate distances based on hashed values, which can be pre-computed and stored when adding variant datasets.

Furthermore, the extension of SNuPy towards other NGS experiments such as RNASeq, whole-genome structural data, and microbiome data will be an important stepping stone towards comprehensive NGS sample analysis. This is expected to give insight into complex and regulatory disease progression mechanisms as well as to allow more targeted drug profiling and in the context of precision medicine approaches.

In conclusion: SNuPy has proven its capabilities in numerous practical case studies to be a user-friendly, comprehensive and versatile tool that empowers genomic

scientists to manage, query and interpret genomic variant datasets without additional bioinformatic support . The DNA data deluge [32] was and will remain a challenge for the next years, but we are sure to have the tools and know-how to navigate it.

Part VIII

APPENDIX

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

Entity	Entities represent individuals that are under investigation and for which biological specimen were retrieved.
Entity Group	Entity groups represent datasets which can be analyzed independently.
Project	Projects can be compiled by a user and consist of any number of entity Groups
Sample	Extract from a VcfFile, which may contain many samples.
Specimen	A biological specimen taken from an Entity.
VcfFile	Result of a NGS analysis pipeline

ACRONYMS

AQuA	Annotation QUery and Aggregation framework
CNV	Copy Number Variant
FDA	U. S. Food and Drug Administration
GUI	Graphical User Interface
HGSV	Human Genome Variation Society
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
HTTPS	Secure Hypertext Transfer Protocol
InDel	Insertion/Deletion
LOVD	Leiden Open Variation Database
MeSH	Medical Subject Heading
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
PM	precision medicine
SNuPy	Single NUcleotide PolYmorphism platform
SNV	Single Nucleotide Variant
SQL	Structured Query Language
SV	Structural Variant
SVI WG	Sequence Variant Interpretation Working Group
VCF	Variant Call Format (developed by the 1000 genomes project)

[1] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333–351, jun 2016. [Online]. Available: http://www.nature.com/articles/nrg.2016.49

[2] T. U. Consortium, "The UK10K project identifies rare variants in health and disease," *Nature*, vol. 526, no. 7571, pp. 82–90, oct 2015. [Online]. Available: http://www.nature.com/articles/nature14962

[3] T. . G. P. Consortium, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, nov 2012. [Online]. Available: http://www.nature.com/articles/nature11632

[4] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, and D. G. MacArthur, "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, no. 7616, pp. 285–291, aug 2016. [Online]. Available: http://www.nature.com/articles/nature19057

[5] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, oct 2013. [Online]. Available: http://www.nature.com/articles/ng.2764

[6] I. F. A. C. Fokkema, P. E. M. Taschner, G. C. P. Schaafsma, J. Celli, J. F. J. Laros, and J. T. den Dunnen, "LOVD v.2.0: the next generation in gene variant databases," *Human Mutation*, vol. 32, no. 5, pp. 557–563, may 2011. [Online]. Available: http://doi.wiley.com/10.1002/humu.21438

[7] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott, "ClinVar: public archive of relationships among sequence variation and human phenotype," *Nucleic Acids Research*, vol. 42, no. D1, pp. D980–D985, jan 2014. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1113

[8] Y. Yang, D. M. Muzny, J. G. Reid, M. N. Bainbridge, A. Willis, P. A. Ward, A. Braxton, J. Beuten, F. Xia, Z. Niu, M. Hardison, R. Person, M. R. Bekheirnia, M. S. Leduc, A. Kirby, P. Pham, J. Scull, M. Wang, Y. Ding, S. E. Plon, J. R. Lupski, A. L. Beaudet, R. A. Gibbs, and C. M. Eng, "Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders," *New England Journal of Medicine*, vol. 369, no. 16, pp. 1502–1511, oct 2013. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMoa1306555

[9] F. S. Collins and H. Varmus, "A New Initiative on Precision Medicine," *New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, feb 2015. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMp1500523

[10] B. C. Worst, C. M. van Tilburg, G. P. Balasubramanian, P. Fiesel, R. Witt, A. Freitag, M. Boudalil, C. Previti, S. Wolf, S. Schmidt, S. Chotewutmontri, M. Bewerunge-Hudler, M. Schick, M. Schlesner, B. Hutter, L. Taylor, T. Borst, C. Sutter, C. R. Bartram, T. Milde, E. Pfaff, A. E. Kulozik, A. von Stackelberg, R. Meisel, A. Borkhardt, D. Reinhardt, J.-H. Klusmann, G. Fleischhack, S. Tippelt, U. Dirksen, H. Jürgens, C. M. Kramm, A. O. von Bueren, F. Westermann, M. Fischer, B. Burkhardt, W. Wößmann, M. Nathrath, S. S. Bielack, M. C. Frühwald, S. Fulda, T. Klingebiel, E. Koscielniak, M. Schwab, R. Tremmel, P. H. Driever, J. H. Schulte, B. Brors, A. von Deimling, P. Lichter, A. Eggert, D. Capper, S. M. Pfister, D. T. Jones, and O. Witt, "Next-generation personalised medicine for high-risk paediatric cancer patients – The INFORM pilot study," *European Journal of Cancer*, vol. 65, pp. 91–101, sep 2016. [Online]. Available: http://dx.doi.org/10.1016/j.ejca.2016.06.009https://linkinghub.elsevier.com/retrieve/pii/S0959804916322122

[11] S. Schleidgen, C. Klingler, T. Bertram, W. H. Rogowski, and G. Marckmann, "What is personalized medicine: sharpening a vague term based on a systematic literature review," *BMC Medical Ethics*, vol. 14, no. 1, p. 55, dec 2013. [Online]. Available: http://www.biomedcentral.com/1472-6939/14/55

[12] J. S. Ross, E. A. Slodkowska, W. F. Symmans, L. Pusztai, P. M. Ravdin, and G. N. Hortobagyi, "The HER-2 Receptor and Breast Cancer: Ten Years of Targeted Anti-HER-2 Therapy and Personalized Medicine," *The Oncologist*, vol. 14, no. 4, pp. 320–368, apr 2009. [Online]. Available: http://theoncologist.alphamedpress.org/cgi/doi/10.1634/theoncologist.2008-0230

[13] L. G. Biesecker and R. C. Green, "Diagnostic Clinical Genome and Exome Sequencing," *New England Journal of Medicine*, vol. 370, no. 25, pp. 2418–2425, jun 2014. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMra1312543

[14] J. Gagan and E. M. Van Allen, "Next-generation sequencing to guide cancer therapy," *Genome Medicine*, vol. 7, no. 1, p. 80, dec 2015. [Online]. Available: http://genomemedicine.com/content/7/1/80

[15] J. I. Hoell, M. Gombert, S. Ginzel, S. Loth, P. Landgraf, V. Käfer, M. Streiter, A. Prokop, M. Weiss, R. Thiele, and A. Borkhardt, "Constitutional Mismatch Repair-deficiency and Whole-exome Sequencing as the Means of the Rapid Detection of the Causative MSH6 Defect," *Klinische Pädiatrie*, vol. 226, no. 06/07, pp. 357–361, nov 2014. [Online]. Available: http://www.thieme-connect.de/DOI/DOI?10.1055/s-0034-1389905

[16] R. J. Mody, J. R. Prensner, J. Everett, D. W. Parsons, and A. M. Chinnaiyan, "Precision medicine in pediatric oncology: Lessons learned and next steps," *Pediatric Blood & Cancer*, vol. 64, no. 3, p. e26288, mar 2017. [Online]. Available: http://doi.wiley.com/10.1002/pbc.26288

[17] Y. Sekiya, Y. Xu, H. Muramatsu, Y. Okuno, A. Narita, K. Suzuki, X. Wang, N. Kawashima, H. Sakaguchi, N. Yoshida, A. Hama, Y. Takahashi, K. Kato, and S. Kojima, "Clinical utility of next-generation sequencing-based minimal residual disease in paediatric B-cell acute lymphoblastic leukaemia," *British Journal of Haematology*, vol. 176, no. 2, pp. 248–257, jan 2017. [Online]. Available: http://doi.wiley.com/10.1111/bjh.14420

[18] M. Kotrova, V. H. J. van der Velden, J. J. M. van Dongen, R. Formankova, P. Sedlacek, M. Brüggemann, J. Zuna, J. Stary, J. Trka, and E. Fronkova, "Next-generation sequencing

indicates false-positive MRD results and better predicts prognosis after SCT in patients with childhood ALL," *Bone Marrow Transplantation*, vol. 52, no. 7, pp. 962–968, jul 2017. [Online]. Available: http://www.nature.com/doifinder/10.1038/bmt.2017.16

[19] E. B. Heikamp and C.-H. Pui, "Next-Generation Evaluation and Treatment of Pediatric Acute Lymphoblastic Leukemia," *The Journal of Pediatrics*, sep 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022347618309442

[20] C. Chen, C. Bartenhagen, M. Gombert, V. Okpanyi, V. Binder, S. Röttgers, J. Bradtke, A. Teigler-Schlegel, J. Harbott, S. Ginzel, R. Thiele, U. Fischer, M. Dugas, J. Hu, and A. Borkhardt, "Next-generation-sequencing-based risk stratification and identification of new genes involved in structural and sequence variations in near haploid lymphoblastic leukemia." *Genes Chromosomes Cancer*, vol. 52, no. 6, pp. 564–579, jun 2013. [Online]. Available: http://doi.wiley.com/10.1002/gcc.22054

[21] Z. Liu, B. Delavan, R. Roberts, and W. Tong, "Lessons Learned from Two Decades of Anticancer Drugs," *Trends in Pharmacological Sciences*, vol. 38, no. 10, pp. 852–872, oct 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165614717301268

[22] U. Fischer, M. Forster, A. Rinaldi, T. Risch, S. Sungalee, H.-J. Warnatz, B. Bornhauser, M. Gombert, C. Kratsch, A. M. Stütz, M. Sultan, J. Tchinda, C. L. Worth, V. Amstislavskiy, N. Badarinarayan, A. Baruchel, T. Bartram, G. Basso, C. Canpolat, G. Cario, H. Cavé, D. Dakaj, M. Delorenzi, M. P. Dobay, C. Eckert, E. Ellinghaus, S. Eugster, V. Frismantas, S. Ginzel, O. A. Haas, O. Heidenreich, G. Hemmrich-Stanisak, K. Hezaveh, J. I. Höll, S. Hornhardt, P. Husemann, P. Kachroo, C. P. Kratz, G. Te Kronnie, B. Marovca, F. Niggli, A. C. McHardy, A. V. Moorman, R. Panzer-Grümayer, B. S. Petersen, B. Raeder, M. Ralser, P. Rosenstiel, D. Schäfer, M. Schrappe, S. Schreiber, M. Schütte, B. Stade, R. Thiele, N. von der Weid, A. Vora, M. Zaliova, L. Zhang, T. Zichner, M. Zimmermann, H. Lehrach, A. Borkhardt, J.-P. Bourquin, A. Franke, J. O. Korbel, M. Stanulla, M.-L. Yaspo, G. te Kronnie, B. Marovca, F. Niggli, A. C. McHardy, A. V. Moorman, R. Panzer-Grümayer, B. S. Petersen, B. Raeder, M. Ralser, P. Rosenstiel, D. Schäfer, M. Schrappe, S. Schreiber, M. Schütte, B. Stade, R. Thiele, N. von der Weid, A. Vora, M. Zaliova, L. Zhang, T. Zichner, M. Zimmermann, H. Lehrach, A. Borkhardt, J.-P. Bourquin, A. Franke, J. O. Korbel, M. Stanulla, and M.-L. Yaspo, "Genomics and drug profiling of fatal TCF3-HLF-positive acute lymphoblastic leukemia identifies recurrent mutation patterns and therapeutic options." *Nature Genetics*, vol. 47, no. 9, pp. 1020–1029, sep 2015. [Online]. Available: http://www.nature.com/articles/ng.3362

[23] M. Joosten, S. Ginzel, C. Blex, D. Schmidt, M. Gombert, C. Chen, R. M. R. Linka, O. Gräbner, A. Hain, B. Hirsch, A. Sommerfeld, A. Seegebarth, U. Gruber, C. Maneck, L. Zhang, K. Stenin, H. Dieks, M. Sefkow, C. Münk, C. D. C. Baldus, R. Thiele, A. Borkhardt, M. Hummel, H. Köster, U. Fischer, M. Dreger, and V. Seitz, "A novel approach to detect resistance mechanisms reveals FGR as a factor mediating HDAC inhibitor SAHA resistance in B-cell lymphoma," *Molecular Oncology*, vol. 10, no. 8, pp. 1232–1244, oct 2016. [Online]. Available: http://doi.wiley.com/10.1016/j.molonc.2016.06.001

[24] A. Martín-Lorenzo, F. Auer, L. N. Chan, I. García-Ramírez, I. González-Herrero, G. Rodríguez-Hernández, C. Bartenhagen, M. Dugas, M. Gombert, S. Ginzel, O. Blanco, A. Orfao, D. Alonso-López, J. D. L. Rivas, M. B. García-Cenador, F. J. García-Criado, M. Müschen, I. Sánchez-García, A. Borkhardt, C. Vicente-Dueñas, and J. Hauer, "Loss of Pax5 Exploits Sca1-BCR-ABL p190 Susceptibility to Confer the Metabolic Shift Essential for pB-ALL," *Cancer Research*, vol. 78, no. 10, pp. 2669–2679, may 2018. [Online]. Available: http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-17-3262

[25] A. Martin-Lorenzo, J. Hauer, C. Vicente-Duenas, F. Auer, I. Gonzalez-Herrero, I. Garcia-Ramirez, S. Ginzel, R. Thiele, S. N. Constantinescu, C. Bartenhagen, M. Dugas, M. Gombert, D. Schafer, O. Blanco, A. Mayado, A. Orfao, D. Alonso-Lopez, J. D. L. Rivas, C. Cobaleda, M. B. Garcia-Cenador, F. J. Garcia-Criado, I. Sanchez-Garcia, and A. Borkhardt, "Infection Exposure Is a Causal Factor in B-cell Precursor Acute Lymphoblastic Leukemia as a Result of Pax5-Inherited Susceptibility," *Cancer Discovery*, vol. 5, no. 12, pp. 1328–1343, dec 2015. [Online]. Available: http://cancerdiscovery.aacrjournals.org/cgi/doi/10.1158/2159-8290.CD-15-0892

[26] V. Prasad, "Perspective: The precision-oncology illusion," *Nature*, vol. 537, no. 7619, pp. S63–S63, sep 2016. [Online]. Available: http://www.nature.com/articles/537S63a

[27] G. Lightbody, V. Haberland, F. Browne, L. Taggart, H. Zheng, E. Parkes, and J. K. Blayney, "Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application," *Briefings in Bioinformatics*, pp. 1–17, jul 2018. [Online]. Available: https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby051/5062275

[28] D. Senft, M. D. Leiserson, E. Ruppin, and Z. A. Ronai, "Precision Oncology: The Road Ahead," *Trends in Molecular Medicine*, vol. 23, no. 10, pp. 874–898, oct 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1471491417301430

[29] S. H. Shin, A. M. Bode, and Z. Dong, "Precision medicine: the foundation of future cancer therapeutics," vol. 1, p. 12, 2017. [Online]. Available: www.nature.com/npjprecisiononcology

[30] K. Wetterstrand, "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)." [Online]. Available: https://www.genome.gov/sequencingcostsdata/

[31] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big Data: Astronomical or Genomical?" *PLOS Biology*, vol. 13, no. 7, p. e1002195, jul 2015. [Online]. Available: http://dx.plos.org/10.1371/journal.pbio.1002195

[32] M. C. Schatz and B. Langmead, "The DNA data deluge," *IEEE Spectrum*, vol. 50, no. 7, pp. 28–33, jul 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6545119/

[33] A. Belkadi, A. Bolze, Y. Itan, A. Cobat, Q. B. Vincent, A. Antipenko, L. Shang, B. Boisson, J.-L. Casanova, and L. Abel, "Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants," *Proceedings of the National Academy of Sciences*, vol. 112, no. 17, pp. 5473–5478, apr 2015. [Online]. Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1418631112

[34] S. Coutant, C. Cabot, A. Lefebvre, M. Léonard, E. Prieur-Gaston, D. Campion, T. Lecroq, and H. Dauchel, "EVA: Exome Variation Analyzer, an efficient and versatile tool for filtering strategies in medical genomics," *BMC Bioinformatics*, vol. 13, no. SUPPL 1, p. S9, 2012. [Online]. Available: http://www.biomedcentral.com/1471-2105/13/S14/S9

[35] C. Gilissen, A. Hoischen, H. G. Brunner, and J. A. Veltman, "Disease gene identification strategies for exome sequencing," *European Journal of Human Genetics*, vol. 20, no. 5, pp. 490–497, may 2012. [Online]. Available: http://www.nature.com/articles/ejhg2011258

[36] C. A. Brownstein and Others, "An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge." *Genome Biol.*, vol. 15, no. 3, p. R53, 2014.

[37] L. M. Amendola, G. P. Jarvik, M. C. Leo, H. M. McLaughlin, Y. Akkari, M. D. Amaral, J. S. Berg, S. Biswas, K. M. Bowling, L. K. Conlin, G. M. Cooper, M. O. Dorschner, M. C. Dulik, A. A. Ghazani, R. Ghosh, R. C. Green, R. Hart, C. Horton, J. J. Johnston, M. S. Lebo, A. Milosavljevic, J. Ou, C. M. Pak, R. Y. Patel, S. Punj, C. S. Richards, J. Salama, N. T. Strande, Y. Yang, S. E. Plon, L. G. Biesecker, and H. L. Rehm, "Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium," *The American Journal of Human Genetics*, vol. 98, no. 6, pp. 1067–1076, jun 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002929716300593

[38] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, and H. L. Rehm, "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology," *Genetics in Medicine*, vol. 17, no. 5, pp. 405–423, may 2015. [Online]. Available: http://www.nature.com/articles/gim201530

[39] R. M. Durbin, D. L. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, F. S. Collins, F. M. De La Vega, P. Donnelly, M. Egholm, P. Flicek, S. B. Gabriel, R. A. Gibbs, B. M. Knoppers, E. S. Lander, H. Lehrach, E. R. Mardis, G. A. McVean, D. A. Nickerson, L. Peltonen, A. J. Schafer, S. T. Sherry, J. J. Wang, R. K. Wilson, R. A. Gibbs, D. Deiros, M. Metzker, D. Muzny, J. Reid, D. Wheeler, J. J. Wang, J. Li, M. Jian, G. Li, R. Li, H. Liang, G. Tian, B. Wang, J. J. Wang, W. Wang, H. Yang, X. Zhang, H. H. Zheng, E. S. Lander, D. L. Altshuler, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, S. B. Gabriel, D. B. Jaffe, E. Shefler, C. L. Sougnez, D. R. Bentley, N. Gormley, S. Humphray, Z. Kingsbury, P. Koko-Gonzales, J. Stone, K. J. McKernan, G. L. Costa, J. K. Ichikawa, C. C. Lee, R. Sudbrak, H. Lehrach, T. A. Borodina, A. Dahl, A. N. Davydov, P. Marquardt, F. Mertes, W. Nietfeld, P. Rosenstiel, S. Schreiber, A. V. Soldatov, B. Timmermann, M. Tolzmann, M. Egholm, J. Affourtit, D. Ashworth, S. Attiya, M. Bachorski, E. Buglione, A. Burke, A. Caprio, C. Celone, S. Clark, D. Conners, B. Desany, L. Gu, L. Guccione, K. Kao, A. Kebbel, J. Knowlton, M. Labrecque, L. McDade, C. Mealmaker, M. Minderman, A. Nawrocki, F. Niazi, K. Pareja, R. Ramenani, D. Riches, W. Song, C. Turcotte, S. Wang, E. R. Mardis, R. K. Wilson, D. Dooling, L. Fulton, R. Fulton, G. Weinstock, R. M. Durbin, J. Burton, D. M. Carter, C. Churcher, A. A. Coffey, A. Cox, A. Palotie, M. Quail, T. Skelly, J. Stalker, H. P. Swerdlow, D. Turner, A. De Witte, S. Giles, R. A. Gibbs, D. Wheeler, M. Bainbridge, D. Challis, A. Sabo, F. Yu, J. Yu, J. J. Wang, X. Fang, X. Guo, R. Li, Y. Y. Li, R. Luo, S. Tai, H. Wu, H. H. Zheng, X. Zheng, Y. Zhou, G. Li, J. J. Wang, H. Yang, G. T. Marth, E. P. Garrison, W. Huang, A. Indap, D. Kural, W.-P. Lee, W. Fung Leong, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. C. Lee, R. E. Mills, X. Shi, M. J. Daly, M. A. DePristo, D. L. Altshuler, A. D. Ball, E. Banks, T. Bloom, B. L. Browning, K. Cibulskis, T. J. Fennell, K. V. Garimella, S. R. Grossman, R. E. Handsaker, M. Hanna, C. Hartl, D. B. Jaffe, A. M. Kernytsky, J. M. Korn, H. Li, J. R. Maguire, S. A. McCarroll, A. McKenna, J. C. Nemesh, A. A. Philippakis, R. E. Poplin, A. Price, M. A. Rivas, P. C. Sabeti, S. F. Schaffner, E. Shefler, I. A. Shlyakhter, D. N. Cooper, E. V. Ball, M. Mort, A. D. Phillips, P. D. Stenson, J. Sebat, V. Makarov, K. K. Ye, S. C. Yoon, C. D. Bustamante, A. G. Clark, A. Boyko, J. Degenhardt, S. Gravel, R. N. Gutenkunst, M. Kaganovich, A. Keinan, P. Lacroute, X. Ma, A. Reynolds, L. Clarke, P. Flicek, F. Cunningham, J. Herrero, S. Keenen, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, R. E. Smith, V. Zalunin, X. Zheng-Bradley, J. O. Korbel, A. M. Stütz, S. Humphray, M. Bauer, R. Keira Cheetham, T. Cox, M. Eberle, T. James, S. Kahn, L. Murray, A. Chakravarti, K. K. Ye, F. M. De La Vega, Y. Fu, F. C. L. Hyland, J. M. Manning, S. F. McLaughlin, H. E. Peckham, O. Sakarya, Y. A. Sun, E. F. Tsung, M. A. Batzer, M. K. Konkel, J. A. Walker, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, R. Herwig, D. V. Parkhomchuk, S. T. Sherry, R. Agarwala,

H. M. Khouri, A. O. Morgulis, J. E. Paschall, L. D. Phan, K. E. Rotmistrovsky, R. D. Sanders, M. F. Shumway, C. Xiao, G. A. McVean, A. Auton, Z. Iqbal, G. Lunter, J. L. Marchini, L. Moutsianas, S. Myers, A. Tumian, B. Desany, J. Knight, R. Winer, D. W. Craig, S. M. Beckstrom-Sternberg, A. Christoforides, A. A. Kurdoglu, J. V. Pearson, S. A. Sinari, W. D. Tembe, D. Haussler, A. S. Hinrichs, S. J. Katzman, A. Kern, R. M. Kuhn, M. Przeworski, R. D. Hernandez, B. Howie, J. L. Kelley, S. Cord Melton, G. R. Abecasis, Y. Y. Li, P. Anderson, T. Blackwell, W. Chen, W. O. Cookson, J. Ding, H. Min Kang, M. Lathrop, L. Liang, M. F. Moffatt, P. Scheet, C. Sidore, M. M. Snyder, X. Zhan, S. Zöllner, P. Awadalla, F. Casals, Y. Idaghdour, J. Keebler, E. A. Stone, M. Zilversmit, L. Jorde, J. Xing, E. E. Eichler, G. Aksay, C. Alkan, I. Hajirasouliha, F. Hormozdiari, J. M. Kidd, S. Cenk Sahinalp, P. H. Sudmant, E. R. Mardis, K. Chen, A. Chinwalla, L. Ding, D. C. Koboldt, M. D. McLellan, D. Dooling, G. Weinstock, J. W. Wallis, M. C. Wendl, Q. Zhang, R. M. Durbin, C. A. Albers, Q. Ayub, S. Balasubramaniam, J. C. Barrett, D. M. Carter, Y. Chen, D. F. Conrad, P. Danecek, E. T. Dermitzakis, M. Hu, N. Huang, M. E. Hurles, H. Jin, L. Jostins, T. M. Keane, S. Quang Le, S. Lindsay, Q. Long, D. G. MacArthur, S. B. Montgomery, L. Parts, J. Stalker, C. Tyler-Smith, K. Walter, Y. Zhang, M. B. Gerstein, M. M. Snyder, A. Abyzov, S. Balasubramanian, R. Bjornson, J. Du, F. Grubert, L. Habegger, R. Haraksingh, J. Jee, E. Khurana, H. Y. K. Lam, J. Leng, X. Jasmine Mu, A. E. Urban, Z. Zhang, Y. Y. Li, R. Luo, G. T. Marth, E. P. Garrison, D. Kural, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. C. Lee, R. E. Mills, X. Shi, S. A. McCarroll, E. Banks, M. A. DePristo, R. E. Handsaker, C. Hartl, J. M. Korn, H. Li, J. C. Nemesh, J. Sebat, V. Makarov, K. K. Ye, S. C. Yoon, J. Degenhardt, M. Kaganovich, L. Clarke, R. E. Smith, X. Zheng-Bradley, J. O. Korbel, S. Humphray, R. Keira Cheetham, M. Eberle, S. Kahn, L. Murray, K. K. Ye, F. M. De La Vega, Y. Fu, H. E. Peckham, Y. A. Sun, M. A. Batzer, M. K. Konkel, J. A. Walker, C. Xiao, Z. Iqbal, B. Desany, T. Blackwell, M. M. Snyder, J. Xing, E. E. Eichler, G. Aksay, C. Alkan, I. Hajirasouliha, F. Hormozdiari, J. M. Kidd, K. Chen, A. Chinwalla, L. Ding, M. D. McLellan, J. W. Wallis, M. E. Hurles, D. F. Conrad, K. Walter, Y. Zhang, M. B. Gerstein, M. M. Snyder, A. Abyzov, J. Du, F. Grubert, R. Haraksingh, J. Jee, E. Khurana, H. Y. K. Lam, J. Leng, X. Jasmine Mu, A. E. Urban, Z. Zhang, R. A. Gibbs, M. Bainbridge, D. Challis, C. Coafra, H. Dinh, C. Kovar, S. Lee, D. Muzny, L. Nazareth, J. Reid, A. Sabo, F. Yu, J. Yu, G. T. Marth, E. P. Garrison, A. Indap, W. Fung Leong, A. R. Quinlan, C. Stewart, A. N. Ward, J. Wu, K. Cibulskis, T. J. Fennell, S. B. Gabriel, K. V. Garimella, C. Hartl, E. Shefler, C. L. Sougnez, J. Wilkinson, A. G. Clark, S. Gravel, F. Grubert, L. Clarke, P. Flicek, R. E. Smith, X. Zheng-Bradley, S. T. Sherry, H. M. Khouri, J. E. Paschall, M. F. Shumway, C. Xiao, G. A. McVean, S. J. Katzman, G. R. Abecasis, T. Blackwell, E. R. Mardis, D. Dooling, L. Fulton, R. Fulton, D. C. Koboldt, R. M. Durbin, S. Balasubramaniam, A. A. Coffey, T. M. Keane, D. G. MacArthur, A. Palotie, C. Scott, J. Stalker, C. Tyler-Smith, M. B. Gerstein, S. Balasubramanian, A. Chakravarti, B. M. Knoppers, G. R. Abecasis, C. D. Bustamante, N. Gharani, R. A. Gibbs, L. Jorde, J. S. Kaye, A. Kent, T. Li, A. L. McGuire, G. A. McVean, P. N. Ossorio, C. N. Rotimi, Y. Su, L. H. Toji, C. Tyler-Smith, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, A. Abdallah, C. R. Juenger, N. C. Clemm, F. S. Collins, A. Duncanson, E. D. Green, M. S. Guyer, J. L. Peterson, A. J. Schafer, G. R. Abecasis, D. L. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, oct 2010. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature09534http://www.ncbi.nlm.nih.gov/snp

[40] J. L. Freeman, "Copy number variation: New insights in genome diversity," *Genome Research*, vol. 16, no. 8, pp. 949–961, jun 2006. [Online]. Available: http://www.genome.org/cgi/doi/10.1101/gr.3677206

[41] NICHOLAS WADE, "Genome of DNA Pioneer Is Deciphered," New York, may 2007. [Online]. Available: https://www.nytimes.com/2007/05/31/science/31cnd-gene.html

[42] "E pluribus unum," *Nature Methods*, vol. 7, no. 5, pp. 331–331, may 2010. [Online]. Available: http://www.nature.com/articles/nmeth0510-331

[43] S. E. Levy and R. M. Myers, "Advancements in Next-Generation Sequencing," *Annual Review of Genomics and Human Genetics*, vol. 17, no. 1, pp. 95–115, aug 2016. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev-genom-083115-022413

[44] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, "Ten years of next-generation sequencing technology," *Trends in Genetics*, vol. 30, no. 9, pp. 418–426, sep 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0168952514001127

[45] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, and T. J. Hubbard, "GENCODE: The reference human genome annotation for The ENCODE Project," *Genome Research*, vol. 22, no. 9, pp. 1760–1774, sep 2012. [Online]. Available: http://genome.cshlp.org/cgi/doi/10.1101/gr.135350.111

[46] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt, "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation," *Nucleic Acids Research*, vol. 44, no. D1, pp. D733–D745, jan 2016. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1189

[47] J. T. den Dunnen, R. Dalgleish, D. R. Maglott, R. K. Hart, M. S. Greenblatt, J. McGowan-Jordan, A.-F. Roux, T. Smith, S. E. Antonarakis, and P. E. Taschner, "HGVS Recommendations for the Description of Sequence Variants: 2016 Update," *Human Mutation*, vol. 37, no. 6, pp. 564–569, jun 2016. [Online]. Available: http://doi.wiley.com/10.1002/humu.22981

[48] A. Bateman, "The Pfam protein families database," *Nucleic Acids Research*, vol. 32, no. 90001, pp. 138D–141, jan 2004. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh121

[49] I. Letunic, T. Doerks, and P. Bork, "SMART 7: recent updates to the protein domain annotation resource," *Nucleic Acids Research*, vol. 40, no. D1, pp. D302–D305, jan 2012. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr931

[50] C. J. A. Sigrist, L. Cerutti, E. de Castro, P. S. Langendijk-Genevaux, V. Bulliard, A. Bairoch, and N. Hulo, "PROSITE, a protein domain database for functional characterization and annotation," *Nucleic Acids Research*, vol. 38, no. suppl_1, pp. D161–D166, jan 2010. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp885

[51] S. T. Sherry, "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, jan 2001. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/29.1.308

[52] P. D. Stenson, M. Mort, E. V. Ball, K. Howells, A. D. Phillips, N. S. Thomas, and D. N. Cooper, "The Human Gene Mutation Database: 2008 update," *Genome Medicine*, vol. 1, no. 1, p. 13, jan 2009. [Online]. Available: http://genomemedicine.biomedcentral.com/articles/10.1186/gm13

[53] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P. A. Futreal, M. R. Stratton, and R. Wooster, "The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website," *British Journal of Cancer*, vol. 91, no. 2, pp. 355–358, jul 2004. [Online]. Available: http://www.nature.com/articles/6601894

[54] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, "OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders," *Nucleic Acids Research*, vol. 43, no. D1, pp. D789–D798, jan 2015. [Online]. Available: http://academic.oup.com/nar/article/43/D1/D789/2439148/OMIMorg-Online-Mendelian-Inheritance-in-Man-OMIM

[55] A. Siepel, K. S. Pollard, and D. Haussler, "New Methods for Detecting Lineage-Specific Selection," in *Annual International Conference on Research in Computational Molecular Biology*. Springer, Berlin, Heidelberg, 2006, pp. 190–205. [Online]. Available: http://link.springer.com/10.1007/11732990{_}17

[56] G. M. Cooper, "Distribution and intensity of constraint in mammalian genomic sequence," *Genome Research*, vol. 15, no. 7, pp. 901–913, jun 2005. [Online]. Available: http://www.genome.org/cgi/doi/10.1101/gr.3577405

[57] A. Siepel, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Research*, vol. 15, no. 8, pp. 1034–1050, aug 2005. [Online]. Available: http://www.genome.org/cgi/doi/10.1101/gr.3715005

[58] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248–249, apr 2010. [Online]. Available: http://www.nature.com/articles/nmeth0410-248

[59] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nature Protocols*, vol. 4, no. 7, pp. 1073–1081, jul 2009. [Online]. Available: http://www.nature.com/articles/nprot.2009.86

[60] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nature Genetics*, vol. 46, no. 3, pp. 310–315, mar 2014. [Online]. Available: http://www.nature.com/articles/ng.2892

[61] I. Meyts, B. Bosch, A. Bolze, B. Boisson, Y. Itan, A. Belkadi, V. Pedergnana, L. Moens, C. Picard, A. Cobat, X. Bossuyt, L. Abel, and J.-L. Casanova, "Exome and genome sequencing for inborn errors of immunity," *Journal of Allergy and Clinical Immunology*, vol. 138, no. 4, pp. 957–969, oct 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0091674916308818

[62] M. M. Li, M. Datto, E. J. Duncavage, S. Kulkarni, N. I. Lindeman, S. Roy, A. M. Tsimberidou, C. L. Vnencak-Jones, D. J. Wolff, A. Younes, and M. N. Nikiforova, "Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer," *The Journal of Molecular Diagnostics*, vol. 19, no. 1, pp. 4–23, jan 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1525157816302239

[63] C. A. Mather, S. D. Mooney, S. J. Salipante, S. Scroggins, D. Wu, C. C. Pritchard, and B. H. Shirts, "CADD score has limited clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel," *Genetics in Medicine*, vol. 18, no. 12, pp. 1269–1275, dec 2016. [Online]. Available: http://www.nature.com/articles/gim201644

[64] L. L. Andersen, E. Terczyńska-Dyla, N. Mørk, C. Scavenius, J. J. Enghild, K. Höning, V. Hornung, M. Christiansen, T. H. Mogensen, and R. Hartmann, "Frequently used bioinformatics tools overestimate the damaging effect of allelic variants," *Genes & Immunity*, no. December 2017, pp. 1–13, dec 2017. [Online]. Available: http://www.nature.com/articles/s41435-017-0002-z

[65] L. A. Miosge, M. A. Field, Y. Sontani, V. Cho, S. Johnson, A. Palkova, B. Balakishnan, R. Liang, Y. Zhang, S. Lyon, B. Beutler, B. Whittle, E. M. Bertram, A. Enders, C. C. Goodnow, and T. D. Andrews, "Comparison of predicted and actual consequences of missense mutations," *Proceedings of the National Academy of Sciences*, vol. 112, no. 37, pp. E5189–E5198, sep 2015. [Online]. Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1511585112

[66] T. A. Peterson, E. Doughty, and M. G. Kann, "Towards Precision Medicine: Advances in Computational Approaches for the Analysis of Human Variants," *Journal of Molecular Biology*, vol. 425, no. 21, pp. 4047–4063, nov 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23962656

[67] A. Niroula and M. Vihinen, "Variation Interpretation Predictors: Principles, Types, Performance, and Choice," *Human Mutation*, vol. 37, no. 6, pp. 579–597, jun 2016. [Online]. Available: http://doi.wiley.com/10.1002/humu.22987

[68] X. Liu, X. Jian, and E. Boerwinkle, "dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations," *Human Mutation*, vol. 34, no. 9, pp. E2393–E2402, sep 2013. [Online]. Available: http://doi.wiley.com/10.1002/humu.22376

[69] P. Cingolani, A. Platts, L. L. L. Wang, M. Coon, T. Nguyen, L. L. L. Wang, S. J. Land, X. Lu, and D. M. Ruden, "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff," *Fly*, vol. 6, no. 2, pp. 80–92, apr 2012. [Online]. Available: http://www.tandfonline.com/doi/abs/10.4161/fly.19695

[70] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Research*, vol. 38, no. 16, pp. e164–e164, sep 2010. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq603

[71] A. V. Kotlar, C. E. Trevino, M. E. Zwick, D. J. Cutler, and T. S. Wingo, "Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale," *Genome Biology*, vol. 19, no. 1, p. 14, dec 2018. [Online]. Available: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1387-3

[72] J. R. Grant, A. S. Arantes, X. Liao, and P. Stothard, "In-depth annotation of SNPs arising from resequencing projects using NGS-SNP," *Bioinformatics*, vol. 27, no. 16, pp. 2300–2301, aug 2011. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr372

[73] S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson, and J. Shendure, "Targeted capture and massively parallel sequencing of 12 human exomes," *Nature*, vol. 461, no. 7261, pp. 272–276, sep 2009. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature08250

[74] I. Medina, A. De Maria, M. Bleda, F. Salavert, R. Alonso, C. Y. Gonzalez, and J. Dopazo, "VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing," *Nucleic Acids Research*, vol. 40, no. W1, pp. W54–W58, jul 2012. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks572

[75] L. Habegger, S. Balasubramanian, D. Z. Chen, E. Khurana, A. Sboner, A. Harmanci, J. Rozowsky, D. Clarke, M. Snyder, and M. Gerstein, "VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment," *Bioinformatics*, vol. 28, no. 17, pp. 2267–2269, sep 2012. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts368

[76] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham, "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor," *Bioinformatics*, vol. 26, no. 16, pp. 2069–2070, aug 2010. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq330

[77] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham, "The Ensembl Variant Effect Predictor," *Genome Biology*, vol. 17, no. 1, pp. 1–14, 2016. [Online]. Available: http://dx.doi.org/10.1186/s13059-016-0974-4

[78] M. Sincan, Others, D. R. Simeonov, D. Adams, T. C. Markello, T. M. Pierson, C. Toro, W. A. Gahl, and C. F. Boerkoel, "VAR-MD: a tool to analyze whole exome-genome variants in small human pedigrees with mendelian inheritance." *Human Mutation*, vol. 33, no. 4, pp. 593–598, 2012.

[79] L. Zhang, J. Zhang, J. Yang, D. Ying, Y. lung Lau, and W. Yang, "PriVar: a toolkit for prioritizing SNVs and indels from next-generation sequencing data," *Bioinformatics*, vol. 29, no. 1, pp. 124–125, jan 2013. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts627

[80] U. Paila, B. A. Chapman, R. Kirchner, and A. R. Quinlan, "GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations," *PLoS Computational Biology*, vol. 9, no. 7, p. e1003153, jul 2013. [Online]. Available: http://dx.plos.org/10.1371/journal.pcbi.1003153

[81] D. Ge, E. K. Ruzzo, K. V. Shianna, M. He, K. Pelak, E. L. Heinzen, A. C. Need, E. T. Cirulli, J. M. Maia, S. P. Dickson, M. Zhu, A. Singh, A. S. Allen, D. B. Goldstein, and Others, "SVA: software for annotating and visualizing sequenced human genomes." *Bioinformatics*, vol. 27, no. 14, pp. 1998–2000, 2011.

[82] J. K. Teer, Others, E. D. Green, J. C. Mullikin, and L. G. Biesecker, "VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer." *Bioinformatics*, vol. 28, no. 4, pp. 599–600, feb 2012. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr711

[83] M. Akgün, H. Demirci, and B. Berger, "VCF-Explorer: filtering and analysing whole genome VCF files," *Bioinformatics*, vol. 33, no. 21, pp. 3468–3470, nov 2017. [Online]. Available: https://academic.oup.com/bioinformatics/article/33/21/3468/3931856

[84] A. Pietrelli, L. Valenti, and J. Wren, "myVCF: a desktop application for high-throughput mutations data management," *Bioinformatics*, vol. 33, no. 22, pp. 3676–3678, nov 2017. [Online]. Available: https://academic.oup.com/bioinformatics/article/33/22/3676/4004873

[85] S. N. Hart, P. Duffy, D. J. Quest, A. Hossain, M. A. Meiners, and J.-P. P. Kocher, "VCF-Miner: GUI-based application for mining variants and annotations stored in VCF files." *Brief. Bioinformatics*, vol. 17, no. 2, pp. 346–351, 2016.

[86] B. Maranhao, P. Biswas, J. Duncan, K. Branham, G. Silva, M. Naeem, S. Khan, S. Riazuddin, J. Hejtmancik, J. Heckenlively, S. Riazuddin, P. Lee, and R. Ayyagari, "exomeSuite: Whole exome sequence variant filtering tool for rapid identification of putative disease causing SNVs/indels," *Genomics*, vol. 103, no. 2-3, pp. 169–176, feb 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888754314000123https://linkinghub.elsevier.com/retrieve/pii/S0888754314000123

[87] J.-P. Desvignes, M. Bartoli, V. Delague, M. Krahn, M. Miltgen, C. Béroud, and D. Salgado, "VarAFT: a variant annotation and filtration system for human next generation sequencing data," *Nucleic Acids Research*, vol. 46, no. W1, pp. W545–W553, jul 2018. [Online]. Available: https://academic.oup.com/nar/article/46/W1/W545/5025894

[88] A. Alemán, F. Garcia-Garcia, F. Salavert, I. Medina, and J. Dopazo, "A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies," *Nucleic Acids Research*, vol. 42, no. W1, pp. W88–W93, jul 2014. [Online]. Available: http://academic.oup.com/nar/article/42/W1/W88/2437375/A-webbased-interactive-framework-to-assist-in-the

[89] M. J. Li, J. Deng, P. Wang, W. Yang, S. L. Ho, P. C. Sham, J. Wang, and M. Li, "wKGGSeq: A Comprehensive Strategy-Based and Disease-Targeted Online Framework to Facilitate Exome Sequencing Studies of Inherited Disorders," *Human Mutation*, vol. 36, no. 5, pp. 496–503, may 2015. [Online]. Available: http://doi.wiley.com/10.1002/humu.22766

[90] M.-X. Li, H.-S. Gui, J. S. H. Kwan, S.-Y. Bao, and P. C. Sham, "A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases," *Nucleic Acids Research*, vol. 40, no. 7, pp. e53–e53, apr 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22241780http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3326332https://academic.oup.com/nar/article/40/7/e53/1200399

[91] G. Vandeweyer, L. Van Laer, B. Loeys, T. Van den Bulcke, and R. F. Kooy, "VariantDB: a flexible annotation and filtering portal for next generation sequencing data," *Genome Medicine*, vol. 6, no. 10, p. 74, oct 2014. [Online]. Available: http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-014-0074-6

[92] I. Granata, M. Sangiovanni, F. Maiorano, M. Miele, and M. R. Guarracino, "Var2GO: a web-based tool for gene variants selection," *BMC Bioinformatics*, vol. 17, no. S12, p. 376, oct 2016. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1197-0

[93] J. Delafontaine, A. Masselot, R. Liechti, D. Kuznetsov, I. Xenarios, and S. Pradervand, "Varapp: A reactive web-application for variants filtering," *bioRxiv*, p. 060806, jun 2016. [Online]. Available: https://www.biorxiv.org/content/early/2016/06/27/060806

[94] R. G. C. C. L. Cardenas, N. D. Linhares, R. L. Ferreira, and S. D. J. Pena, "Mendel,MD: A user-friendly open-source web tool for analyzing WES and WGS in the diagnosis of patients with Mendelian disorders," *PLOS Computational Biology*, vol. 13, no. 6, p. e1005520, jun 2017. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1005520

[95] A. Sifrim, J. K. J. Van Houdt, L. C. Tranchevent, B. Nowakowska, R. Sakai, G. A. Pavlopoulos, K. Devriendt, J. R. Vermeesch, Y. Moreau, J. Aerts, and Others, "Annotate-it: a Swiss-knife approach to annotation, analysis and interpretation of single nucleotide variation in human disease." *Genome Medicine*, vol. 4, no. 9, p. 73, 2012. [Online]. Available: http://genomemedicine.com/content/4/9/73

[96] M. D. Preston, M. Manske, N. Horner, S. Assefa, S. Campino, S. Auburn, I. Zongo, J.-B. Ouedraogo, F. Nosten, T. Anderson, and T. G. Clark, "VarB: a variation browsing and analysis tool for variants derived from next-generation sequencing data," *Bioinformatics*, vol. 28, no. 22, pp. 2983–2985, nov 2012. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts557

[97] F. A. San Lucas, G. Wang, P. Scheet, and B. Peng, "Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools," *Bioinformatics*, vol. 28, no. 3, pp. 421–422, feb 2012. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr667

[98] H. Vuong, R. M. Stephens, and N. Volfovsky, "AVIA: an interactive web-server for annotation, visualization and impact analysis of genomic variations," *BMC Proceedings*, vol. 6, no. Suppl 6, p. P37, oct 2012. [Online]. Available: http://bmcproc.biomedcentral.com/articles/10.1186/1753-6561-6-S6-P37

[99] M. D'Antonio, P. D'Onorio De Meo, D. Paoletti, B. Elmi, M. Pallocca, N. Sanna, E. Picardi, G. Pesole, and T. Castrignanò, "WEP: a high-performance analysis pipeline for whole-exome data," *BMC Bioinformatics*, vol. 14, no. Suppl 7, p. S11, apr 2013. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S7-S11

[100] Y.-J. Na, Y. Cho, and J. H. Kim, "AnsNGS: An Annotation System to Sequence Variations of Next Generation Sequencing Data for Disease-Related Phenotypes," *Healthcare Informatics Research*, vol. 19, no. 1, p. 50, mar 2013. [Online]. Available: https://synapse.koreamed.org/DOIx.php?id=10.4258/hir.2013.19.1.50

[101] J. Yao, K. X. Zhang, M. Kramer, M. Pellegrini, and W. R. McCombie, "FamAnn: an automated variant annotation pipeline to facilitate target discovery for family-based sequencing studies," *Bioinformatics*, vol. 30, no. 8, pp. 1175–1176, apr 2014. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt749

[102] S. Salatino and V. Ramraj, "BrowseVCF: a web-based application and workflow to quickly prioritize disease-causative variants in VCF files," *Briefings in Bioinformatics*, vol. 21, no. 8, p. bbw054, jul 2016. [Online]. Available: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbw054

[103] H. Müller, R. Jimenez-Heredia, A. Krolo, T. Hirschmugl, J. Dmytrus, K. Boztug, and C. Bock, "VCF.Filter: interactive prioritization of disease-linked genetic variants from sequencing data," *Nucleic Acids Research*, vol. 45, no. W1, pp. W567–W572, jul 2017. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx425

[104] K. Chamberlain, Scott Boettiger, Carl Hart, Ted Ram, "rcrossref: Client for Various 'CrossRef' 'APIs'. R package version 0.7.0." 2017. [Online]. Available: https://cran.r-project.org/package=rcrossref

[105] M. Li, J. Li, M. J. Li, Z. Pan, J. S. Hsu, D. J. Liu, X. Zhan, J. Wang, Y. Song, and P. C. Sham, "Robust and rapid algorithms facilitate large-scale whole genome sequencing downstream analysis in an integrative framework," *Nucleic Acids Research*, vol. 45, no. 9, p. gkx019, jan 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28115622http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5435951https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx019

[106] J. Lopez, J. Coll, M. Haimel, S. Kandasamy, J. Tarraga, P. Furio-Tari, W. Bari, M. Bleda, A. Rueda, S. Gräf, A. Rendon, J. Dopazo, and I. Medina, "HGVA: the Human Genome Variation Archive," *Nucleic Acids Research*, vol. 45, no. W1, pp. W189–W194, jul 2017. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx445

[107] A. Frankish, B. Uszczynska, G. R. Ritchie, J. M. Gonzalez, D. Pervouchine, R. Petryszak, J. M. Mudge, N. Fonseca, A. Brazma, R. Guigo, and J. Harrow, "Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction," *BMC Genomics*, vol. 16, no. Suppl 8, p. S2, 2015. [Online]. Available: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-16-S8-S2

[108] World Health Organization, "WHO definitions of genetics and genomics," p. 1, 2018. [Online]. Available: http://www.who.int/genomics/geneticsVSgenomics/en/

[109] J. S. Black, M. Salto-Tellez, K. I. Mills, and M. A. Catherwood, "The impact of next generation sequencing technologies on haematological research – A review," *Pathogenesis*, vol. 2, no. 1-2, pp. 9–16, 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S221466361500005X

[110] E. Shumilov, J. Flach, A. Kohlmann, Y. Banz, N. Bonadies, M. Fiedler, T. Pabst, and U. Bacher, "Current status and trends in the diagnostics of AML and MDS," 2018.

[111] L. A. Garraway, J. Verweij, and K. V. Ballman, "Precision oncology: an overview." *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 31, no. 15, pp. 1803–5, may 2013. [Online]. Available: http://ascopubs.org/doi/10.1200/JCO.2013.49.4799

[112] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease," *Nature Genetics*, vol. 33, no. 3s, pp. 228–237, mar 2003. [Online]. Available: http://www.nature.com/doifinder/10.1038/ng1090

[113] K. Eilbeck, A. Quinlan, and M. Yandell, "Settling the score: variant prioritization and Mendelian disease," *Nature Reviews Genetics*, vol. 18, no. 10, pp. 599–612, aug 2017. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrg.2017.52

[114] D. G. MacArthur and C. Tyler-Smith, "Loss-of-function variants in the genomes of healthy humans," *Human Molecular Genetics*, vol. 19, no. R2, pp. R125–R130, oct 2010. [Online]. Available: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddq365

[115] R. Chen, L. Shi, J. Hakenberg, B. Naughton, P. Sklar, J. Zhang, H. Zhou, L. Tian, O. Prakash, M. Lemire, P. Sleiman, W.-y. Cheng, W. Chen, H. Shah, Y. Shen, M. Fromer, L. Omberg, M. A. Deardorff, E. Zackai, J. R. Bobe, E. Levin, T. J. Hudson, L. Groop, J. Wang, H. Hakonarson, A. Wojcicki, G. A. Diaz, L. Edelmann, E. E. Schadt, and S. H. Friend, "Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases," *Nature Biotechnology*, vol. 34, no. 5, pp. 531–538, may 2016. [Online]. Available: http://www.nature.com/articles/nbt.3514

[116] S. Serratì, S. De Summa, B. Pilato, D. Petriella, R. Lacalamita, S. Tommasi, and R. Pinto, "Next-generation sequencing: advances and applications in cancer diagnosis," *OncoTargets and Therapy*, vol. Volume 9, pp. 7355–7365, dec 2016. [Online]. Available: https://www.dovepress.com/next-generation-sequencing-advances-and-applications-in-cancer-diagnos-peer-reviewed-article-OTT

[117] R. Bao, L. Huang, J. Andrade, W. Tan, W. A. Kibbe, H. Jiang, and G. Feng, "Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing," *Cancer Informatics*, vol. 13s2, p. CIN.S13779, jan 2014. [Online]. Available: http://journals.sagepub.com/doi/10.4137/CIN.S13779

[118] H. L. Rehm, S. J. Bale, P. Bayrak-Toydemir, J. S. Berg, K. K. Brown, J. L. Deignan, M. J. Friez, B. H. Funke, M. R. Hegde, and E. Lyon, "ACMG clinical laboratory standards for next-generation sequencing," *Genetics in Medicine*, vol. 15, no. 9, pp. 733–747, sep 2013. [Online]. Available: http://www.nature.com/articles/gim201392

[119] D. J. McCarthy, P. Humburg, A. Kanapin, M. A. Rivas, K. Gaulton, J.-B. B. Cazier, and P. Donnelly, "Choice of transcripts and software has a large effect on variant annotation." *Genome Medicine*, vol. 6, no. 3, p. 26, 2014.

[120] S. Nabhani, S. Ginzel, H. Miskin, S. Revel-Vilk, D. Harlev, B. Fleckenstein, A. Honscheid, P. T. Oommen, M. Kuhlen, R. Thiele, H.-J. Laws, A. Borkhardt, P. Stepensky, and U. Fischer, "Deregulation of Fas ligand expression as a novel cause of autoimmune lymphoproliferative syndrome-like disease," *Haematologica*, vol. 100, no. 9, pp. 1189–1198, sep 2015. [Online]. Available: http://www.haematologica.org/cgi/doi/10.3324/haematol.2014.114967

[121] M. A. Sukhai, K. J. Craddock, M. Thomas, A. R. Hansen, T. Zhang, L. Siu, P. Bedard, T. L. Stockley, and S. Kamel-Reid, "A classification system for clinical relevance of somatic variants identified in molecular profiling of cancer," *Genetics in Medicine*, vol. 18, no. 2, pp. 128–136, feb 2016. [Online]. Available: http://www.nature.com/articles/gim201547

[122] D. Salgado, M. I. Bellgard, J.-P. Desvignes, and C. Béroud, "How to Identify Pathogenic Mutations among All Those Variations: Variant Annotation and Filtration in the Genome Sequencing Era," *Human Mutation*, vol. 37, no. 12, pp. 1272–1282, dec 2016. [Online]. Available: http://doi.wiley.com/10.1002/humu.23110

[123] S. Chen, X. Hu, and Y. Shen, "Sequence Variant Interpretation 2.0: Perspective on New Guidelines for Sequence Variant Classification," *Clinical Chemistry*, vol. 61, no. 11, pp. 1317–1319, nov 2015. [Online]. Available: http://www.clinchem.org/cgi/doi/10.1373/clinchem.2015.240812

[124] D. C. Hoskinson, A. M. Dubuc, and H. Mason-Suares, "The current state of clinical interpretation of sequence variants," *Current Opinion in Genetics & Development*, vol. 42, pp. 33–39, feb 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959437X17300072

[125] S. Ellard, E. L. Baple, M. Owens, D. M. Eccles, C. Turnbull, S. Abbs, R. Scott, Z. C. Deans, T. Lester, J. Campbell, W. G. Newman, and D. J. Mcmullan, "ACGS Best Practice Guidelines for Variant Classification 2018," Tech. Rep., 2018. [Online]. Available: https://www.clinicalgenome.org/working-groups/sequence-variant-interpretation/

[126] L. G. Biesecker and S. M. Harrison, "The ACMG/AMP reputable source criteria for the interpretation of sequence variants," *Genetics in Medicine*, vol. 20, no. 12, pp. 1687–1688, dec 2018. [Online]. Available: http://www.nature.com/articles/gim201842

[127] C. S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, and H. L. Rehm, "Response to Biesecker and Harrison," *Genetics in Medicine*, vol. 20, no. 12, pp. 1689–1690, dec 2018. [Online]. Available: http://www.nature.com/articles/gim201843

[128] A. N. Abou Tayoun, T. Pesaran, M. T. DiStefano, A. Oza, H. L. Rehm, L. G. Biesecker, and S. M. Harrison, "Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion," *Human Mutation*, vol. 39, no. 11, pp. 1517–1524, nov 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/05/09/313718http://doi.wiley.com/10.1002/humu.23626

[129] R. Ghosh, S. M. Harrison, H. L. Rehm, S. E. Plon, and L. G. Biesecker, "Updated recommendation for the benign stand-alone ACMG/AMP criterion," *Human Mutation*, vol. 39, no. 11, pp. 1525–1530, nov 2018. [Online]. Available: http://doi.wiley.com/10.1002/humu.23642

[130] E. A. Rivera-Muñoz, L. V. Milko, S. M. Harrison, D. R. Azzariti, C. L. Kurtz, K. Lee, J. L. Mester, M. A. Weaver, E. Currey, W. Craigen, C. Eng, B. Funke, M. Hegde, R. E. Hershberger, R. Mao, R. D. Steiner, L. M. Vincent, C. L. Martin, S. E. Plon, E. Ramos, H. L. Rehm, M. Watson, and J. S. Berg, "ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation," *Human Mutation*, vol. 39, no. 11, pp. 1614–1622, nov 2018. [Online]. Available: http://doi.wiley.com/10.1002/humu.23645

[131] K. Nykamp, M. Anderson, M. Powers, J. Garcia, B. Herrera, Y.-Y. Ho, Y. Kobayashi, N. Patil, J. Thusberg, M. Westbrook, and S. Topper, "Sherloc: a comprehensive refinement of the ACMG–AMP variant classification criteria," *Genetics in Medicine*, vol. 19, no. 10, pp. 1105–1117, oct 2017. [Online]. Available: http://www.nature.com/doifinder/10.1038/gim.2017.37

[132] E. M. Van Allen, N. Wagle, P. Stojanov, D. L. Perrin, K. Cibulskis, S. Marlow, J. Jane-Valbuena, D. C. Friedrich, G. Kryukov, S. L. Carter, A. McKenna, A. Sivachenko, M. Rosenberg, A. Kiezun, D. Voet, M. Lawrence, L. T. Lichtenstein, J. G. Gentry, F. W. Huang, J. Fostel, D. Farlow, D. Barbie, L. Gandhi, E. S. Lander, S. W. Gray, S. Joffe, P. Janne, J. Garber, L. MacConaill, N. Lindeman, B. Rollins, P. Kantoff, S. A. Fisher, S. Gabriel, G. Getz, and L. A. Garraway, "Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine," *Nature Medicine*, vol. 20, no. 6, pp. 682–688, jun 2014. [Online]. Available: http://www.nature.com/articles/nm.3559

[133] A. Olivé, *Conceptual Modeling of Information Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. [Online]. Available: http://link.springer.com/10.1007/978-3-540-39390-0

[134] K. Sahatqija, J. Ajdari, X. Zenuni, B. Raufi, and F. Ismaili, "Comparison between relational and NOSQL databases," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, may 2018, pp. 0216–0221. [Online]. Available: https://ieeexplore.ieee.org/document/8400041/

[135] M. A. Mohamed, O. G. Altrafi, and M. O. Ismail, "Relational vs. NoSQL Databases: A Survey," *International Journal of Computer and Information Technology*, vol. 3, no. 3, pp. 2279–0764, 2014.

[136] N. Zeng, G. Q. Zhang, X. Li, and L. Cui, "Evaluation of Relational and NoSQL Approaches for Cohort Identification from Heterogeneous Data Sources in the National Sleep Research Resource," *Journal of Health & Medical Informatics*, vol. 08, no. 05, 2017. [Online]. Available: https://www.omicsonline.org/open-access/72148-2157-7420-8-295(3).pdf

[137] W. L. Schulz, B. G. Nelson, D. K. Felker, T. J. Durant, and R. Torres, "Evaluation of relational and NoSQL database architectures to manage genomic annotations," *Journal of Biomedical Informatics*, vol. 64, pp. 288–295, dec 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046416301526

[138] Y. Li and S. Manoharan, "A performance comparison of SQL and NoSQL databases," in *2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*. IEEE, aug 2013, pp. 15–19. [Online]. Available: http://ieeexplore.ieee.org/document/6625441/

[139] Z. Parker, S. Poe, and S. V. Vrbsky, "Comparing NoSQL MongoDB to an SQL DB," in *Proceedings of the 51st ACM Southeast Conference on - ACMSE '13*. New York, New York, USA: ACM Press, 2013, p. 1. [Online]. Available: http://delivery.acm.org/10.1145/2510000/2500047/a5-parker.pdf?ip=194.95.66.1{&}id=2500047{&}acc=ACTIVESERVICE{&}key=2BA2C432AB83DA15.D6DC39BE0BC9DD13.4D4702B0C3E38B35.4D4702B0C3E38B35{&}{_}{_}acm{_}{_}=1534350847{_}7293eb0785f3dc85507fdaa1d56f59a0

[140] J. R. Lourenço, B. Cabral, P. Carreiro, M. Vieira, and J. Bernardino, "Choosing the right NoSQL database for the job: a quality attribute evaluation," *Journal of Big Data*, vol. 2, no. 1, p. 18, dec 2015. [Online]. Available: http://www.journalofbigdata.com/content/2/1/18

[141] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears, "Benchmarking cloud serving systems with YCSB," in *Proceedings of the 1st ACM symposium on Cloud computing - SoCC '10*. New York, New York, USA: ACM Press, 2010, p. 143. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1807128.1807152

[142] Y.-T. Liao, J. Zhou, C.-H. Lu, S.-C. Chen, C.-H. Hsu, W. Chen, M.-F. Jiang, and Y.-C. Chung, "Data adapter for querying and transformation between SQL and NoSQL database," *Future Generation Computer Systems*, vol. 65, pp. 111–121, dec 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X16300085

[143] F. Mitha, H. Herodotou, N. Borisov, C. Jiang, J. Yoder, and K. Owzar, "SNPpy - Database Management for SNP Data from Genome Wide Association Studies," *PLoS ONE*, vol. 6, no. 10, p. e24982, oct 2011. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0024982

[144] R. N. Lichtenwalter, K. Zorina-Lichtenwalter, and L. Diatchenko, "Genotypic Data in Relational Databases: Efficient Storage and Rapid Retrieval," in *Advances in Databases and Information Systems*. Springer, 2017, pp. 408–421. [Online]. Available: http://link.springer.com/10.1007/978-3-319-66917-5{_}27

[145] S. De Capitani di Vimercati, S. Paraboschi, and P. Samarati, "Access control: principles and solutions," *Software: Practice and Experience*, vol. 33, no. 5, pp. 397–421, apr 2003. [Online]. Available: http://doi.wiley.com/10.1002/spe.513

[146] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype–phenotype interactions," *Nature Reviews Genetics*, vol. 16, no. 2, pp. 85–97, feb 2015. [Online]. Available: http://www.nature.com/articles/nrg3868

[147] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. v. Mering, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, no. Database, pp. D561–D568, jan 2011. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq973

[148] J. A. Cuff, "The Ensembl Computing Architecture," *Genome Research*, vol. 14, no. 5, pp. 971–975, may 2004. [Online]. Available: http://www.genome.org/cgi/doi/10.1101/gr.1866304

[149] P. Danecek and Others, "The variant call format and VCFtools." *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011.

[150] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature Genetics*, vol. 43, no. 5, pp. 491–498, may 2011. [Online]. Available: http://www.nature.com/articles/ng.806

[151] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson, "VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing," *Genome Research*, vol. 22, no. 3, pp. 568–576, mar 2012. [Online]. Available: http://genome.cshlp.org/cgi/doi/10.1101/gr.129684.111

[152] A. Magi, L. Tattini, I. Cifola, R. D'Aurizio, M. Benelli, E. Mangano, C. Battaglia, E. Bonora, A. Kurg, M. Seri, P. Magini, B. Giusti, G. Romeo, T. Pippucci, G. D. Bellis, R. Abbate, and G. F. Gensini, "EXCAVATOR: detecting copy number variants from whole-exome sequencing data," *Genome Biology*, vol. 14, no. 10, p. R120, dec 2013. [Online]. Available: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-10-r120

[153] H. Li, "Tabix: fast retrieval of sequence features from generic TAB-delimited files," *Bioinformatics*, vol. 27, no. 5, pp. 718–719, mar 2011. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq671

[154] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo, "From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline," in *Current Protocols in Bioinformatics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., oct 2013, vol. 43, no. 1, pp. 11.10.1–11.10.33. [Online]. Available: http://doi.wiley.com/10.1002/0471250953.bi1110s43

[155] K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz, "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples," *Nature Biotechnology*, vol. 31, no. 3, pp. 213–219, mar 2013. [Online]. Available: http://www.nature.com/articles/nbt.2514

[156] K. C. Cotto, A. H. Wagner, Y.-Y. Feng, S. Kiwala, A. C. Coffman, G. Spies, A. Wollam, N. C. Spies, O. L. Griffith, and M. Griffith, "DGIdb 3.0: a redesign and expansion of the drug–gene interaction database," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1068–D1073, jan 2018. [Online]. Available: https://academic.oup.com/nar/article/46/D1/D1068/4634012

[157] S. Janssen, "Spike." [Online]. Available: https://github.com/sjanssen2/spike

[158] J. Koster and S. Rahmann, "Snakemake–a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, oct 2012. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts480

[159] M. A. Kelly, C. Caleshu, A. Morales, J. Buchan, Z. Wolf, S. M. Harrison, S. Cook, M. W. Dillon, J. Garcia, E. Haverfield, J. D. H. Jongbloed, D. Macaya, A. Manrai, K. Orland, G. Richard, K. Spoonamore, M. Thomas, K. Thomson, L. M. Vincent, R. Walsh, H. Watkins, N. Whiffin, J. Ingles, J. P. van Tintelen, C. Semsarian, J. S. Ware, R. Hershberger, and B. Funke, "Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel," *Genetics in Medicine*, vol. 20, no. 3, pp. 351–359, mar 2018. [Online]. Available: http://www.nature.com/doifinder/10.1038/gim.2017.218

[160] G. W. Tam, R. Redon, N. P. Carter, and S. G. Grant, "The Role of DNA Copy Number Variation in Schizophrenia," *Biological Psychiatry*, vol. 66, no. 11, pp. 1005–1012, dec 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0006322309009007

[161] E. J. Hollox and B.-P. Hoh, "Human gene copy number variation and infectious disease," *Human Genetics*, vol. 133, no. 10, pp. 1217–1233, oct 2014. [Online]. Available: http://link.springer.com/10.1007/s00439-014-1457-x

[162] A. Shlien and D. Malkin, "Copy number variations and cancer," *Genome Medicine*, vol. 1, no. 6, p. 62, jun 2009. [Online]. Available: http://genomemedicine.biomedcentral.com/articles/10.1186/gm62

[163] N. Brouwers, C. Van Cauwenberghe, S. Engelborghs, J.-C. Lambert, K. Bettens, N. Le Bastard, F. Pasquier, A. G. Montoya, K. Peeters, M. Mattheijssens, R. Vandenberghe, P. P. De

Deyn, M. Cruts, P. Amouyel, K. Sleegers, and C. Van Broeckhoven, "Alzheimer risk associated with a copy number variation in the complement receptor 1 increasing C3b/C4b binding sites," *Molecular Psychiatry*, vol. 17, no. 2, pp. 223–233, feb 2012. [Online]. Available: http://www.nature.com/articles/mp201124

[164] D. Levy, M. Ronemus, B. Yamrom, Y.-h. Lee, A. Leotta, J. Kendall, S. Marks, B. Lakshmi, D. Pai, K. Ye, A. Buja, A. Krieger, S. Yoon, J. Troge, L. Rodgers, I. Iossifov, and M. Wigler, "Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders," *Neuron*, vol. 70, no. 5, pp. 886–897, jun 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21658582

[165] B. L. Grayson, M. E. Smith, J. W. Thomas, L. Wang, P. Dexheimer, J. Jeffrey, P. R. Fain, P. Nanduri, G. S. Eisenbarth, and T. M. Aune, "Genome-Wide Analysis of Copy Number Variation in Type 1 Diabetes," *PLoS ONE*, vol. 5, no. 11, p. e15393, nov 2010. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0015393

[166] N. Pankratz, A. Dumitriu, K. N. Hetrick, M. Sun, J. C. Latourelle, J. B. Wilk, C. Halter, K. F. Doheny, J. F. Gusella, W. C. Nichols, R. H. Myers, T. Foroud, and A. L. DeStefano, "Copy Number Variation in Familial Parkinson Disease," *PLoS ONE*, vol. 6, no. 8, p. e20988, aug 2011. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0020988

[167] D. T. Teachey, A. E. Seif, and S. A. Grupp, "Advances in the management and understanding of autoimmune lymphoproliferative syndrome (ALPS)," *British Journal of Haematology*, vol. 148, no. 2, pp. 205–216, jan 2010. [Online]. Available: http://doi.wiley.com/10.1111/j.1365-2141.2009.07991.x

[168] J. B. Oliveira, J. J. Bleesing, U. Dianzani, T. A. Fleisher, E. S. Jaffe, M. J. Lenardo, F. Rieux-Laucat, R. M. Siegel, H. C. Su, D. T. Teachey, and V. K. Rao, "Revised diagnostic criteria and classification for the autoimmune lymphoproliferative syndrome (ALPS): report from the 2009 NIH International Workshop," *Blood*, vol. 116, no. 14, pp. e35–e40, oct 2010. [Online]. Available: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2010-04-280347

[169] K. Bride and D. Teachey, "Autoimmune lymphoproliferative syndrome: more than a FAScinating disease," *F1000Research*, vol. 6, p. 1928, nov 2017. [Online]. Available: https://f1000research.com/articles/6-1928/v1

[170] T. A. Fleisher and J. B. Oliveira, "Monogenic defects in lymphocyte apoptosis," *Current Opinion in Allergy and Clinical Immunology*, vol. 12, no. 6, pp. 609–615, dec 2012. [Online]. Available: https://insights.ovid.com/crossref?an=00130832-201212000-00007

[171] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Zhang, C. Zeng, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, L. D. Stein, F. Cunningham, A. Kanani, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, P. Donnelly, J. Marchini, G. A. T. McVean, S. R. Myers, L. R. Cardon, G. R. Abecasis, A. Morris, B. S. Weir, J. C. Mullikin, S. T. Sherry, M. Feolo, D. Altshuler, M. J. Daly, S. F. Schaffner, R. Qiu, A. Kent, G. M. Dunston, K. Kato, N. Niikawa, B. M. Knoppers, M. W. Foster, E. W. Clayton, V. O. Wang, J. Watkin, R. A. Gibbs, J. W. Belmont, E. Sodergren, G. M. Weinstock, R. K. Wilson, L. L. Fulton, J. Rogers, B. W. Birren, H. Han, H. Wang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Todani, T. Fujita, S. Tanaka, A. L. Holden, E. H. Lai, F. S. Collins, L. D. Brooks, J. E. McEwen, M. S. Guyer, E. Jordan, J. L. Peterson, J. Spiegel, L. M. Sung, L. F. Zacharia, K. Kennedy, M. G. Dunn, R. Seabrook, M. Shillito, B. Skene, J. G. Stewart, D. L. Valle, E. W. Clayton, L. B. Jorde, J. W. Belmont, A. Chakravarti, M. K. Cho, T. Duster, M. W. Foster, M. Jasperse, B. M. Knoppers, P.-Y. Kwok, J. Licinio, J. C. Long, P. A. Marshall, P. N. Ossorio, V. O. Wang, C. N. Rotimi, C. D. M. Royal, P. Spallone, S. F. Terry, E. S. Lander, E. H. Lai, D. A. Nickerson, G. R. Abecasis, D. Altshuler, D. R. Bentley, M. Boehnke, L. R. Cardon,

M. J. Daly, P. Deloukas, J. A. Douglas, S. B. Gabriel, R. R. Hudson, T. J. Hudson, L. Kruglyak, P.-Y. Kwok, Y. Nakamura, R. L. Nussbaum, C. D. M. Royal, S. F. Schaffner, S. T. Sherry, L. D. Stein, and T. Tanaka, "The International HapMap Project," *Nature*, vol. 426, no. 6968, pp. 789–796, dec 2003. [Online]. Available: http://www.nature.com/articles/nature02168

[172] S. Nabhani, C. Schipp, H. Miskin, C. Levin, S. Postovsky, T. Dujovny, A. Koren, D. Harlev, A.-M. A.-M. Bis, F. Auer, B. Keller, K. Warnatz, M. Gombert, S. Ginzel, A. Borkhardt, P. Stepensky, and U. Fischer, "STAT3 gain-of-function mutations associated with autoimmune lymphoproliferative syndrome like disease deregulate lymphocyte apoptosis and can be targeted by BH3 mimetic compounds," *Clinical Immunology*, vol. 181, pp. 32–42, aug 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1521661616304478

[173] S. E. Flanagan, A.-M. Patch, and S. Ellard, "Using SIFT and PolyPhen to Predict Loss-of-Function and Gain-of-Function Mutations," *Genetic Testing and Molecular Biomarkers*, vol. 14, no. 4, pp. 533–537, aug 2010. [Online]. Available: http://www.liebertpub.com/doi/10.1089/gtmb.2010.0036

[174] B. S. Mann, J. R. Johnson, M. H. Cohen, R. Justice, and R. Pazdur, "FDA Approval Summary: Vorinostat for Treatment of Advanced Primary Cutaneous T-Cell Lymphoma," *The Oncologist*, vol. 12, no. 10, pp. 1247–1252, oct 2007. [Online]. Available: http://theoncologist.alphamedpress.org/cgi/doi/10.1634/theoncologist.12-10-1247

[175] M. Iwamoto, E. J. Friedman, P. Sandhu, N. G. B. Agrawal, E. H. Rubin, and J. A. Wagner, "Clinical pharmacology profile of vorinostat, a histone deacetylase inhibitor," *Cancer Chemotherapy and Pharmacology*, vol. 72, no. 3, pp. 493–508, sep 2013. [Online]. Available: http://link.springer.com/10.1007/s00280-013-2220-z

[176] H. Köster, D. P. Little, P. Luan, R. Muller, S. M. Siddiqi, S. Marappan, and P. Yip, "Capture Compound Mass Spectrometry: A Technology for the Investigation of Small Molecule Protein Interactions," *ASSAY and Drug Development Technologies*, vol. 5, no. 3, pp. 381–390, jun 2007. [Online]. Available: http://www.liebertpub.com/doi/10.1089/adt.2006.039

[177] C.-H. Pui, J. J. Yang, S. P. Hunger, R. Pieters, M. Schrappe, A. Biondi, A. Vora, A. Baruchel, L. B. Silverman, K. Schmiegelow, G. Escherich, K. Horibe, Y. C. Benoit, S. Izraeli, A. E. J. Yeoh, D.-C. Liang, J. R. Downing, W. E. Evans, M. V. Relling, and C. G. Mullighan, "Childhood Acute Lymphoblastic Leukemia: Progress Through Collaboration," *Journal of Clinical Oncology*, vol. 33, no. 27, pp. 2938–2948, sep 2015. [Online]. Available: http://ascopubs.org/doi/10.1200/JCO.2014.59.1636

[178] C. Eckert, G. Henze, K. Seeger, N. Hagedorn, G. Mann, R. Panzer-Grümayer, C. Peters, T. Klingebiel, A. Borkhardt, M. Schrappe, A. Schrauder, G. Escherich, L. Sramkova, F. Niggli, J. Hitzler, and A. von Stackelberg, "Use of Allogeneic Hematopoietic Stem-Cell Transplantation Based on Minimal Residual Disease Response Improves Outcomes for Children With Relapsed Acute Lymphoblastic Leukemia in the Intermediate-Risk Group," *Journal of Clinical Oncology*, vol. 31, no. 21, pp. 2736–2742, jul 2013. [Online]. Available: http://ascopubs.org/doi/10.1200/JCO.2012.48.5680

[179] M. Kato, Y. Horikoshi, Y. Okamoto, Y. Takahashi, D. Hasegawa, K. Koh, J. Takita, M. Inoue, H. Kigasawa, A. Ogawa, Y. Sasahara, K. Kawa, H. Yabe, H. Sakamaki, R. Suzuki, and K. Kato, "Second allogeneic hematopoietic SCT for relapsed ALL in children," *Bone Marrow Transplantation*, vol. 47, no. 10, pp. 1307–1311, oct 2012. [Online]. Available: http://www.nature.com/articles/bmt201229

[180] S. N. Gröbner, B. C. Worst, J. Weischenfeldt, I. Buchhalter, K. Kleinheinz, V. A. Rudneva, P. D. Johann, G. P. Balasubramanian, M. Segura-Wang, S. Brabetz, S. Bender, B. Hutter, D. Sturm, E. Pfaff, D. Hübschmann, G. Zipprich, M. Heinold, J. Eils, C. Lawerenz, S. Erkek, S. Lambo, S. Waszak, C. Blattmann, A. Borkhardt, M. Kuhlen, A. Eggert, S. Fulda, M. Gessler, J. Wegert, R. Kappler, D. Baumhoer, S. Burdach, R. Kirschner-Schwabe, U. Kontny, A. E. Kulozik, D. Lohmann, S. Hettmer, C. Eckert, S. Bielack, M. Nathrath, C. Niemeyer, G. H. Richter, J. Schulte, R. Siebert, F. Westermann, J. J. Molenaar, G. Vassal, H. Witt, P. Lichter, U. Weber, R. Eils, A. Korshunov, O. Witt, S. Pfister, G. Reifenberger, J. Felsberg, C. von Kalle, M. Schmidt, C. Bartholomä, M. Taylor, S. Pfister, D. Jones, P. Lichter, N. Jäger, I. Buchhalter, J. Korbel, A. Stütz, T. Rausch, B. Radlwimmer, M.-L. Yaspo, H. Lehrach, H.-J. Warnatz, P. Landgraf, A. Borkhardt, B. Brors, M. Zapatka, R. Eils, R. Eils, J. Eils, C. Lawerenz, R. Siebert, S. Wagner, A. Haake, J. Richter, G. Richter, R. Eils, C. Lawerenz, J. Eils, J. Kerssemakers, C. Jaeger-Schmidt, I. Scholz, A. K. Bergmann, C. Borst, B. Burkhardt, A. Claviez, M. Dreyling, S. Eberth, H. Einsele, N. Frickhofen, S. Haas, M.-L. Hansmann, D. Karsch, M. Kneba, J. Lisfeld, L. Mantovani-Löffler, M. Rohde, G. Ott, C. Stadler, P. Staib, S. Stilgenbauer, L. Trümper, T. Zenz, M.-L. Hansmann, D. Kube, R. Küppers, M. Weniger, M. Hummel, W. Klapper, U. Kostezka, D. Lenze, P. Möller, A. Rosenwald, G. Ott, M. Szczepanowski, O. Ammerpohl, S. M. Aukema, V. Binder, A. Borkhardt, A. Haake, J. I. Hoell, E. Leich, P. Lichter, C. López, I. Nagel, J. Pischimariov, B. Radlwimmer, J. Richter, P. Rosenstiel, A. Rosenwald, M. Schilhabel, S. Schreiber, I. Vater, R. Wagener, R. Siebert, S. H. Bernhart, H. Binder, B. Brors, G. Doose, R. Eils, S. Hoffmann, L. Hopp, D. Hübschmann, K. Kleinheinz, H. Kretzmer, M. Kreuz, J. Korbel, D. Langenberger, M. Loeffler, M. Rosolowski, M. Schlesner, P. F. Stadler, S. Sungalee, B. Burkhardt, C. P. Kratz, O. Witt, C. M. van Tilburg, C. M. Kramm, G. Fleischhack, U. Dirksen, S. Rutkowski, M. Frühwald, K. von Hoff, S. Wolf, T. Klingebiel, E. Koscielniak, P. Landgraf, J. Koster, A. C. Resnick, J. Zhang, Y. Liu, X. Zhou, A. J. Waanders, D. A. Zwijnenburg, P. Raman, B. Brors, U. D. Weber, P. A. Northcott, K. W. Pajtler, M. Kool, R. M. Piro, J. O. Korbel, M. Schlesner, R. Eils, D. T. W. Jones, P. Lichter, L. Chavez, M. Zapatka, and S. M. Pfister, "The landscape of genomic alterations across childhood cancers," *Nature*, vol. 555, no. 7696, pp. 321–327, feb 2018. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature25480

[181] M. Bodini, C. Ronchini, L. Giacò, G. Giacò, A. Russo, G. E. M. Melloni, L. Luzi, D. Sardella, S. Volorio, S. K. Hasan, T. Ottone, S. Lavorgna, F. Lo-Coco, A. Candoni, R. Fanin, E. Toffoletti, I. Iacobucci, G. Martinelli, A. Cignetti, C. Tarella, L. Bernard, P. G. Pelicci, and L. Riva, "The hidden genomic landscape of acute myeloid leukemia: subclonal structure revealed by undetected mutations," 2015. [Online]. Available: www.bloodjournal.org

[182] J. I. Hoell, S. Ginzel, C. Eckert, M. Gombert, U. Fischer, M. Stanulla, M. Schrappe, U. zur Stadt, P. Bader, B. Strahm, J. Alten, A. Moericke, G. Escherich, A. Stackelberg, C. Peters, A. Borkhardt, and R. Meisel, "Mutational Landscape of Pediatric Acute Lymphoblastic Leukemia Relapsing after Allogeneic Stem Cell Transplantation," in *58th ASH Annual Meeting*, vol. 128, no. 22, 2016.

[183] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1996, pp. 226–231. [Online]. Available: https://dl.acm.org/citation.cfm?id=3001507

[184] J. I. Hoell, M. Gombert, C. Bartenhagen, S. Ginzel, P. Husemann, J. Felsberg, G. Reifenberger, A. Eggert, M. Dugas, S. Schönberger, A. Borkhardt, and U. Fischer, "Whole-genome paired-end analysis confirms remarkable genomic stability of atypical teratoid/rhabdoid tumors," *Genes, Chromosomes and Cancer*, vol. 52, no. 10, pp. 983–985, oct 2013. [Online]. Available: http://doi.wiley.com/10.1002/gcc.22092

[185] E. Salzer, S. Daschkey, S. Choo, M. Gombert, E. Santos-Valente, S. Ginzel, M. Schwendinger, O. A. Haas, G. Fritsch, W. F. Pickl, E. Forster-Waldl, A. Borkhardt, K. Boztug, K. Bienemann, and M. G. Seidel, "Combined immunodeficiency with life-threatening EBV-associated lymphoproliferative disorder in patients lacking functional CD27," *Haematologica*, vol. 98, no. 3, pp. 473–478, mar 2013. [Online]. Available: http://www.haematologica.org/cgi/doi/10.3324/haematol.2012.068791

[186] F. Auer, F. Rüschendorf, M. Gombert, P. Husemann, S. Ginzel, S. Izraeli, M. Harit, M. Weintraub, O. Y. Weinstein, I. Lerer, P. Stepensky, A. Borkhardt, and J. Hauer, "Inherited susceptibility to pre B-ALL caused by germline transmission of PAX5 c.547G>A," *Leukemia*, vol. 28, no. 5, pp. 1136–1138, may 2014. [Online]. Available: http://www.nature.com/articles/leu2013363

[187] C. Chen, C. Bartenhagen, M. Gombert, V. Okpanyi, V. Binder, S. Röttgers, J. Bradtke, A. Teigler-Schlegel, J. Harbott, S. Ginzel, R. Thiele, P. Husemann, P. F. P. Krell, A. Borkhardt, M. Dugas, J. Hu, and U. Fischer, "Next-generation-sequencing of recurrent childhood high hyperdiploid acute lymphoblastic leukemia reveals mutations typically associated with high risk patients," *Leukemia Research*, vol. 39, no. 9, pp. 990–1001, sep 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0145212615303325

[188] S. Daschkey, K. Bienemann, V. Schuster, H. H. W. Kreth, R. M. R. Linka, A. Hönscheid, G. Fritz, C. Johannes, B. Fleckenstein, B. Kempkes, M. Gombert, S. Ginzel, and A. Borkhardt, "Fatal Lymphoproliferative Disease in Two Siblings Lacking Functional FAAP24," *Journal of Clinical Immunology*, vol. 36, no. 7, pp. 684–692, oct 2016. [Online]. Available: http://link.springer.com/10.1007/s10875-016-0317-y

[189] C. Schipp, S. Nabhani, K. Bienemann, N. Simanovsky, S. Kfir-Erenfeld, N. Assayag-Asherie, P. T. Oommen, S. Revel-Vilk, A. Honscheid, M. Gombert, S. Ginzel, D. Schafer, H.-J. Laws, E. Yefenof, B. Fleckenstein, A. Borkhardt, P. Stepensky, and U. Fischer, "Specific antibody deficiency and autoinflammatory disease extend the clinical and immunological spectrum of heterozygous NFKB1 loss-of-function mutations in humans," *Haematologica*, vol. 101, no. 10, pp. e392–e396, oct 2016. [Online]. Available: http://www.haematologica.org/cgi/doi/10.3324/haematol.2016.145136

[190] K. Bienemann, S. Daschkey, J. Sörensen, D. Schwabe, T. Klingebiel, A. Hönscheid, M. Gombert, S. Ginzel, and A. Borkhardt, "A novel homozygous mutation in UNC13D presenting as Epstein–Barr-virus-associated lymphoproliferative disease at 9 years of age," *Leukemia & Lymphoma*, vol. 57, no. 12, pp. 2949–2951, dec 2016. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10428194.2016.1177724

[191] S. Ghosh, A. Hönscheid, G. Dückers, S. Ginzel, H. Gohlke, M. Gombert, B. Kempkes, W. Klapper, M. Kuhlen, H.-J. Laws, R. M. Linka, R. Meisel, C. Mielke, T. Niehues, D. Schindler, D. Schneider, F. R. Schuster, C. Speckmann, and A. Borkhardt, "Human RAD52 – a novel player in DNA repair in cancer and immunodeficiency," *Haematologica*, vol. 102, no. 2, pp. e69–e72, feb 2017. [Online]. Available: http://www.haematologica.org/lookup/doi/10.3324/haematol.2016.155838

Part IX

SUPPLEMENT

REPORT EXAMPLES

1 ACMG/AMP REPORT TEMPLATE

The ACMG/AMP report template can be used as a standardized document for variant interpretation after users have identified variants of interest from their query results. The tables of the evidence framework can be edited using common word processing programs (such as LibreOffice or Word). The report template can be share among multiple users and after all details are filled out, can also be re-uploaded to SNuPy to document the findings.

The first page shows an overview of the entities of the investigated entity group, giving users the necessary information about available samples and states. The second page shows the evidence framework as an overview, to help users navigate the complex evidence framework. The next pages contain the same evidence framework table, but augmented with information from SNuPy . Each page only displays information for a single variant. The last pages show the detailed descriptions used for the evidence framework, again helping users to navigate the interpretation framework.

Genetic Analysis ALPS 6

Sebastian Ginzel | sginze2s@inf.h-brs.de | 11 April 2019

Name: ALPS 6
Parents: ALPS 6 Father & ALPS 6 Mother
Siblings:
GENDER: unknown
DISEASE: Autoimmune Lymphoproliferative Syndrome [C20.683.515.124]
CLASS: primary immune de ciency

ACMG/AMP Criteria overview (Richards et al. 2015)

Category|Evidence	Strong benign	Supporting benign	Supporting pathogenic	Moderate pathogenic	Strong pathogenic	Very strong pathogenic
Population data	MAF is too high for disorder BA1/BS1 OR observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affected statistically increased over controls PS4	
Compulational predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4. Missense in gene where only truncating cause disease BP1. Silent variant with non predicted splice impact BP7. In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5. Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4 Observed		Cosegregation with disease in multiple affected family members PP1 (possibly increasing towards 'Very strong pathogenic')			
Denovo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in trans with a dominant variant BP2. Observed in cis with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source without shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

IL12RB1(ENST00000600835.2:c.634C>T)

Category|Evidence	Strong benign	Supporting benign	Supporting pathogenic	Moderate pathogenic	Strong pathogenic	Very strong pathogenic
Population data	BA1: EXaC: ["A(1.652e-05)"], BS1: less frequent than expected, BS2: PATIENT: ALPS 6 (1.0) FATHER: ALPS 6 Father (0.5893) MOTHER: ALPS 6 Father (0.5893) SIBLINGS:			PM2: EXaC: ["A(1.652e-05)"]	PS4: Case: 1, Control: 1, Total: 956	
Compulational predictive data		BP4: POLYPHEN2_HDVI_PRED: . POLYPHEN2_HVAR_PRED: . SIFT_PRED: T GERP_RS > 2?: -8.14 CADD > 15?: true BP1: not missense BP7: NA BP3: NA	PP3: POLYPHEN2_HDVI_PRED: . POLYPHEN2_HVAR_PRED: . SIFT_PRED: T GERP_RS > 2?: -8.14 CADD > 15?: true	PM4: IL12RB1(stop_gained) PM5: Is pathogenic, but not missense	PS1: Is pathogenic, but not missense	IL12RB1(stop_gained)
Functional data	BS3: https://www.genecards.org/cgi-bin/carddisp.pl?gene=IL12RB1#function		PP2: not missense	PM1: [IL12RB1] Immunodeficiency_30 (distance: -3)	PS3: https://www.ncbi.nlm.nih.gov/pubmed/	
Segregation data	BS4: PATIENT: ALPS 6 (1.0) FATHER: ALPS 6 Father (0.5893) MOTHER: ALPS 6 Father (0.5893) SIBLINGS:		PP1: PATIENT: ALPS 6 (1.0) FATHER: ALPS 6 Father (0.5893) MOTHER: ALPS 6 Father (0.5893) SIBLINGS:			
Denovo data				PM6: Missing in dbSNP - assumed DE NOVO	PS2: PATIENT: ALPS 6 (1.0) FATHER: ALPS 6 Father (0.5893) MOTHER: ALPS 6 Father (0.5893) SIBLINGS:	
Allelic data		BP2: NA		PM3: NA		
Other database						
Other data		BP5: Case with disease: 1, Cases w/o disease: 0, Total Entities: 956	PP4: ["Autoimmune Lymphoproliferative Syndrome [C20.683.515.124]"]			

ACMG/AMP Code Description (Richards et al. 2015)

ACMG/AMP Code	Description
Pathogenic	
Very Strong	
PVS1	null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease
Strong	
PS1	Same amino acid change as a previously established pathogenic variant regardless of nucleotide change
PS2	De novo (both maternity and paternity confirmed) in a patient with the disease and no family history
PS3	Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product
PS4	The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls
Moderate	
PM1	Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation
PM2	Absent from controls (or at extremely low frequency if recessive) (table 6) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium
PM3	For recessive disorders, detected in trans with a pathogenic variant
PM4	Protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants
PM5	Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before
PM6	Assumed de novo, but without confirmation of paternity and maternity
Supporting	
PP1	Cosegregation with disease in multiple affected family members in a gene definitively known to cause the disease
PP2	Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease

PP3	Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.)
PP4	Patient’s phenotype or family history is highly specific for a disease with a single genetic etiology
PP5	Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation
Benign	
Stand-alone	
BA1	Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium
Strong	
BS1	Allele frequency is greater than expected for disorder (see table 6)
BS2	Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected at an early age
BS3	Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing
BS4	Lack of segregation in affected members of a family
Supporting	
BP1	Missense variant in a gene for which primarily truncating variants are known to cause disease
BP2	Observed in trans with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in cis with a pathogenic variant in any inheritance pattern
BP3	In-frame deletions/insertions in a repetitive region without a known function
BP4	Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.)
BP5	Variant found in a case with an alternate molecular basis for disease
BP6	Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation
BP7	A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved

2 DGIDB DRUG INTERACTION REPORT TEMPLATE

The DGIDB drug interaction report is used to identify druggable targets of a variant list in a standardized fashion. The gene-drug interaction data is based on the work by Cotto et al. [156].

The example report shows data from the Ashkenazim Trio[b] that is used as a gold standard for variant datasets.

To better illustrate a real reporting scenario, the son of the family trio is exemplary treated as a malignant sample with diagnosis for Burkitt Lymphoma. The first page shows a summary of the entity group, giving investigators the necessary information about the sample backgrounds. The second and following pages show a table with the coordinates of the mutation, the affected genes and the gene-drug interaction information, including the sources for the interaction. In order to make it easier for users to identify variants potential functional variants CADD scores are added to the table as well.

b `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/OsloUniversityHospital_Exome_GATK_jointVC_11242015/HG002-HG003-HG004.jointVC.filter.vcf`

Genetic Analysis AshkenazimTrio-child

Senor Developer | none@example.com | 16 April 2019

Name: AshkenazimTrio-child
Parents: AshkenazimTrio-father & AshkenazimTrio-mother
Siblings:
GENDER: unknown
CLASS: malignant
AGE_GROUP: child
DISEASE: Burkitt Lymphoma [C20.683.515.761.480.150.165]

Data is based on Drug Gene Interaction Database (https://doi.org/10.1093/nar/gkx1143)

Variant	Gene	Mutation	CADD	DGIDB-drugs	DGIDB-sources
6:32191658-32191670TAGCAGCAGCAGC>T	NOTCH4	ENST00000375023.3:c.36_47delGCTGCTGCTGCT)	10.23	MK0752(inhibitor); PF-03084014(inhibitor); REGN421(antibody); RO4929097(inhibitor)	MyCancerGenome
19:17650229-17650229C>A	FAM129C	ENST00000335393.4:c.878C>A)	9.466		

3 QUERY SUMMARY REPORT TEMPLATE

This report template displays a summary of the query result and can be used as a generic report and template for users to mark variants of interest.

The example report shows data from the Ashkenazim Trio[c] that is used as a gold standard for variant datasets.

To better illustrate a real reporting scenario, the son of the family trio is exemplary treated as a malignant sample with diagnosis for Burkitt Lymphoma. The first page shows a summary of the entity group, giving investigators the necessary information about the sample backgrounds. Following pages of the report show tables from three sections: Transcripts, Inheritance and Phenotypes. The transcript category displays the mutation coordinates, gene and transcript location as well as information from population and variant databases. The Inheritance table shows the variant frequencies in the diseased entity, its parents and possibly the siblings. This allows users to judge the penetrance of a variant. The last phenotype category lists all OMIM and Clinvar phenotypes for the genes that a variant is associated with, additional CADD scores are shown to support users in prioritizing variants of interest.

On the last page, SNuPy documents which filter criteria have been used to filter the variants, helping users to reproduce the query.

---

c `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/OsloUniversityHospital_Exome_GATK_jointVC_11242015/HG002-HG003-HG004.jointVC.filter.vcf`

Genetic Analysis AshkenazimTrio-child

Senor Developer | none@example.com | 16 April 2019

Name: AshkenazimTrio-child

Parents: AshkenazimTrio-father & AshkenazimTrio-mother

Siblings:

GENDER: unknown

CLASS: malignant

AGE_GROUP: child

DISEASE: Burkitt Lymphoma [C20.683.515.761.480.150.165]

PROFILE

Transcripts

Variant	Symbol	Transcript	Exac	Dbsnp
6:32191658-32191670TAGCAGCAGCAGC>T	NOTCH4(inframe_deletion)	ENST00000375023.3:c.36_47delGCTGCTGCTGCT		rs150280230
19:17650229-17650229C>A	FAM129C(missense_variant)	ENST00000335393.4:c.878C>A	0.1154 (A)	rs114207587

Inheritance

Variant	Ashkenazi mtrio-child (child sp2-test)	Ashkenazi mtrio-child (child sp2-testa)	Ashkenazi mtrio-child (child-sp1)	Father	Mother	Siblings	Autosomal dominant	Autosomal recessive	Compound heterozygous	Denovo
6:32191658-32191670TAGCAGCAGCAGC>T	NaN	NaN	1.0	1.0	NaN	NaN	N	N	ENST00000375023	N
19:17650229-17650229C>A	NaN	NaN	0.49	0.45	NaN	NaN	N	N	ENST00000335393	N

Phenotypes

Variant	Cadd	Omim	Clinvar
6:32191658-32191670TAGCAGCAGCAGC>T	10.23		NOTCH4: not_speci ed
19:17650229-17650229C>A	9.466		

FILTER CRITERIA

Query	Value	Filters
Read depth	13	Read depth of variant.
Consequence	["frameshift_variant", "incomplete_terminal_codon_variant", "inframe_deletion", "inframe_insertion", "initiator_codon_variant", "mature_miRNA_variant", "missense_variant", "splice_acceptor_variant", "splice_donor_variant", "start_lost", "5_prime_UTR_premature_start_codon_gain_variant", "stop_gained", "stop_lost", "stop_retained_variant", "TF_binding_site_variant", "TFBS_ablation"]	Most severe consequence
Population frequency	0.15	1000 Genomes Phase 1 AND ExAc Adj. Freq.
Compound heterozygous in parents	false	Compound heterozygous - requires entity relation and parents (based on VEP)