Sic itur ad astra

Virgil

From Department of Molecular Medicine and Surgery
Karolinska Institutet, Stockholm, Sweden

BIOINFORMATIC METHODS IN RARE DISEASE GENOMICS

Måns Magnusson

Karolinska Institutet

Stockholm 2021

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2021

ISBN 978-91-8016-276-0
Cover illustration: Chromosomes II by Maya Brandi

Bioinformatic Methods in Rare Disease Genomics

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Måns Magnusson

The thesis will be defended in public at Eva & Georg Klein, Biomedicum, Stockholm on Friday 22/10-2021 at 10 a.m.

Principal Supervisor:
Professor Anna Wedell
Karolinska Institutet
Department of Molecular Medicine and Surgery
Centre for Inherited Metabolic Diseases

Co-supervisor(s):
Dr Henrik Stranneheim
Karolinska Institutet
Department of Molecular Medicine and Surgery
Centre for Inherited Metabolic Diseases

Dr Daniel Nilsson
Karolinska Institutet
Department of Molecular Medicine and Surgery
Division of Clinical Genetics

Opponent:
Professor Clare Turnbull
Institute of Cancer Research, London
Department of Genetics and Epidemiology
Division of Translational Genetics

Examination Board:
Docent Hans Ehrencrona
Lund University
Department of Laboratory Medicine
Division of Clinical Genetics

Professor Kerstin Lindblad-Toh
Uppsala University
Department of Medical Biochemistry and Microbiology
Division of Clinical Genetics

Docent Joakim Klar
Uppsala University
Department of Immunology, Genetics and Pathology
Division of Medical Genetics and Genomics

To my wonderful children

Iker

Inez

Abbe

Mauritz

and the love of my life

Ann

what would this life be without all of you

POPULAR SCIENCE SUMMARY OF THE THESIS

Health care is in the middle of a revolution, even though it might not be obvious to everyone. During the last decade it has become affordable to access the whole genome of patients entering the clinic. To convert the biological molecule to human readable text is known as **DNA sequencing** or simply "sequencing". It is a complex process that have evolved rapidly since the first time the **DNA** molecule was discovered in the 1950s. The first time a complete human genome was sequenced was in 2001 when the giant Human Genome Project finished. This is one of humanity's mayor efforts and has been compared to the moon landing that was completed in the same amount of time, about 50 years, after the first flying machine was constructed.

Genes in the genome act as blueprints for **proteins** that are involved in most processes that take place on the molecular level in all living organisms. Disruptions of the DNA is known as **mutations** or **genomic variations** and can lead to devastating implications for the individual, such as severe diseases or different types of cancer. As our understanding of the genome increases, the mechanisms behind these conditions become clearer and the chances of curing or preventing them in time gets better.

The genome is vast, every cell of every living organism holds a full copy of its genome, in humans the genome consists of about **3 billion** molecular bases or positions. The discipline of processing and analyzing these large datasets is known as **bioinformatics**.
During this thesis work I have developed bioinformatic methods and tools that have been crucial in the process of implementing whole genome sequencing into health care in the Stockholm region. In this thesis I will go through the current state of the field then continue with explaining how the methods I have developed work and are implemented. Finally, I will show how the sum of this work have impacted health care, both on a large scale and for specific patients.

POPULÄRVETENSKAPLIG SAMMANFATTNING

Det pågår en revolution inom vården som inte är uppenbar för alla. Under det senaste årtiondet har det blivit möjligt att analysera hela den mänskliga arvsmassan för patienter i sjukvården. Processen att konvertera den biologiska molekylen till läsbar text kallas för **DNA-sekvensering** eller endast "sekvensering". Det är en komplicerad process som har utvecklats snabbt sedan DNA-molekylen upptäcktes för första gången under 1950-talet. Det mänskliga genomet kartlades fullständigt för första gången år 2001 då det gigantiska **Human Genome Project** avslutades. Det måste anses vara en av mänsklighetens största triumfer och har jämförts med månlandningen som uppnåddes på ungefär lika lång tid, runt 50 år, räknat från när den första flygmaskinen byggdes.

Det finns delar av genomet som fungerar som ritningar till **protein**, dessa delar kallas för **gener**. Proteinerna är i sin tur involverade i alla biologiska processer som pågår i alla levande organismer. Förändringar i DNA kallas för **mutationer** eller **genomiska variationer**, dessa förändringar kan leda till förödande konsekvenser hos individen så som svåra sjukdomar och olika typer av cancer. I och med att vår förståelse för hur genomet fungerar ökar, så medför detta att vi kan förstå hur mekanismerna bakom dessa sjukdomar fungerar. I förlängningen leder denna kunskap till ökade chanser att hitta behandlingar och större förståelse för hur vi kan förebygga sjukdomarna.

Genomet är väldigt stort och ändå innehåller varje cell i varje levande varelse en fullständig kopia av dess genom. I människor består genomet av **3 miljarder** molekylära baser, eller positioner. Vetenskapen som handlar om att processa och analysera dessa gigantiska data kallas för **Bioinformatik**. Under min tid som doktorand så har jag utvecklat bioinformatiska metoder och verktyg, dessa har varit avgörande för en lyckad implementering av helgenomsekvensering i Stockholms sjukvård. I den här avhandlingen så kommer jag göra en sammanfattning av forskningsfältet idag för att sedan förklara hur metoderna som jag utvecklat fungerar och hur dom är skapade. Avslutningsvis så kommer jag visa hur summan av detta arbete har varit med och påverkat sjukvården, både i stor skala och på individnivå.

ABSTRACT

The larger goal of medical genetics is to map genotype to phenotype and to understand how genomic variation affects human health. In the field of rare disease genomics, there is a mendelian assumption that states: one disease one variant. This is simplified and means that when we observe the phenotype of a rare disease patient, we suspect that there is one or two genetic variations in one gene that cause the disease. It might sound like a simple problem to solve at first, especially compared to other fields in genomics, such as cancer and common disease where multiple loci, unrelated, together are expected to cause the biological state. However, it can be a daunting task to find this variant among the handful of million variants that each human individual is carrying in the genome. This thesis is focused on the problem of finding the causative variants in patients with suspected rare inherited disorders even though some of the tools and methods are applicable in other areas as well.

Many challenges arise in the sequencing analysis as the amount of data grows, requiring development of novel methods and algorithms to enable handling and interpretation of the massive amounts of data. Hundreds of millions of short sequence reads are produced for a single individual in a whole genome sequencing experiment. These are mapped to a reference genome and the positions and regions that differ from the reference are identified or “called” as variants. The variants are annotated with as much relevant information as possible, so that prediction algorithms and humans can determine which variant or small number of variants among the millions identified that are pathogenic in a particular genomic or phenotypic context.

This thesis was created in parallel with the process of establishing a genomics platform in the Stockholm region, to provide the hospitals with state-of-the-art genome analysis. The tools and methods that were developed during these years were implemented and tested in a production setting immediately.

In this thesis work I will illustrate the field of Clinical Genomics from different perspectives, from the components of a rare disease analysis pipeline to the integration of whole genome sequencing in a clinical setting via a close-up case study.

LIST OF SCIENTIFIC PAPERS INCLUDED IN THE THESIS

I. **MultiQC: summarize analysis results for multiple tools and samples in a single report.**
Ewels P, Magnusson M, Lundin S, Käller M.
Bioinformatics. 2016 Oct 1;32(19):3047-8.

II. **Loqusdb: added value of an observations database of local genomic variation.**
Magnusson M, Eisfeldt J, Nilsson D, Rosenbaum A, Wirta V, Lindstrand A, Wedell A, Stranneheim H.
BMC Bioinformatics. 2020 Jul 1;21(1):273.

III. **Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients.**
Stranneheim H[*], Lagerstedt-Robinson K[*], Magnusson M, Kvarnung M, Nilsson D, Lesko N, Engvall M, Anderlid BM, Arnell H, Johansson CB, Barbaro M, Björck E, Bruhn H, Eisfeldt J, Freyer C, Grigelioniene G, Gustavsson P, Hammarsjö A, Hellström-Pigg M, Iwarsson E, Jemt A, Laaksonen M, Enoksson SL, Malmgren H, Naess K, Nordenskjöld M, Oscarson M, Pettersson M, Rasi C, Rosenbaum A, Sahlin E, Sardh E, Stödberg T, Tesi B, Tham E, Thonberg H, Töhönen V, von Döbeln U, Vassiliou D, Vonlanthen S, Wikström AC, Wincent J, Winqvist O, Wredenberg A, Ygberg S, Zetterström RH, Marits P, Soller MJ, Nordgren A, Wirta V, Lindstrand A[†], Wedell A[†].
Genome Med. 2021 Mar 17;13(1):40.

IV. ***SLC12A2* mutations cause NKCC1 deficiency with encephalopathy and impaired secretory epithelia.**
Stödberg T[*], Magnusson M[*], Lesko N, Wredenberg A, Martin Munoz D, Stranneheim H, Wedell A.
Neurol Genet. 2020 Jul 2;6(4):e478.

[*] Shared first authorship
[†] Shared senior authorship

ADDITIONAL SCIENTIFIC PAPERS

Listed in chronological order:

- **High diagnostic yield in skeletal ciliopathies using massively parallel genome sequencing, structural variant screening and RNA analyses.**
  Hammarsjö A, Pettersson M, Chitayat D, Handa A, Anderlid BM, Bartocci M, Basel D, Batkovskyte D, Beleza-Meireles A, Conner P, Eisfeldt J, Girisha KM, Chung BH, Horemuzova E, Hyodo H, Korņejeva L, Lagerstedt-Robinson K, Lin AE, Magnusson M, Moosa S, Nayak SS, Nilsson D, Ohashi H, Ohashi-Fukuda N, Stranneheim H, Taylan F, Traberg R, Voss U, Wirta V, Nordgren A, Nishimura G, Lindstrand A, Grigelioniene G.
  J Hum Genet. 2021 Apr 20.
- **Chanjo: Clincal grade sequence coverage analysis**.
  Andeer R, Magnusson M, Wedell A and Stranneheim H.
  F1000Research 2020, **9**:615
- **From cytogenetics to cytogenomics: whole-genome sequencing as a first-line test comprehensively captures the diverse spectrum of disease-causing genetic variation underlying intellectual disability.**
  Lindstrand A, Eisfeldt J, Pettersson M, Carvalho CMB, Kvarnung M, Grigelioniene G, Anderlid BM, Bjerin O, Gustavsson P, Hammarsjö A, Georgii-Hemming P, Iwarsson E, Johansson-Soller M, Lagerstedt-Robinson K, Lieden A, Magnusson M, Martin M, Malmgren H, Nordenskjöld M, Norling A, Sahlin E, Stranneheim H, Tham E, Wincent J, Ygberg S, Wedell A, Wirta V, Nordgren A, Lundin J, Nilsson D.
  Genome Med. 2019 Nov 7;11(1):68.
- **Rescue of primary ubiquinone deficiency due to a novel COQ7 defect using 2,4-dihydroxybensoic acid.**
  Freyer C, Stranneheim H, Naess K, Mourier A, Felser A, Maffezzini C, Lesko N, Bruhn H, Engvall M, Wibom R, Barbaro M, Hinze Y, Magnusson M, Andeer R, Zetterström RH, von Döbeln U, Wredenberg A, Wedell A.
  J Med Genet. 2015 Nov;52(11):779-83.
- **Rapid pulsed whole genome sequencing for comprehensive acute diagnostics of inborn errors of metabolism.**
  Stranneheim H, Engvall M, Naess K, Lesko N, Larsson P, Dahlberg M, Andeer R, Wredenberg A, Freyer C, Barbaro M, Bruhn H, Emahazion T, Magnusson M, Wibom R, Zetterström RH, Wirta V, von Döbeln U, Wedell A.
  BMC Genomics. 2014 Dec 11;15(1):1090.
- **An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge.**
  Brownstein CA, Beggs AH, Homer N, Merriman B, Yu TW, Flannery KC, DeChene ET, Towne MC, Savage SK, Price EN, Holm IA, Luquette LJ, Lyon E, Majzoub J, Neupert P, McCallie D Jr, Szolovits P, Willard HF, Mendelsohn NJ, Temme R, Finkel RS, Yum SW, Medne L, Sunyaev SR, Adzhubey I, Cassa CA, de Bakker PI, Duzkale H, Dworzyński P, Fairbrother W, Francioli L, Funke BH, Giovanni MA, Handsaker RE, Lage K, Lebo MS, Lek M, Leshchiner I, MacArthur DG, McLaughlin HM, Murray MF, Pers TH, Polak PP, Raychaudhuri S, Rehm HL, Soemedi R, Stitziel NO, Vestecka S, Supper J, Gugenmus C, Klocke B, Hahn A, Schubach M, Menzel M, Biskup S, Freisinger P, Deng M, Braun M, Perner S,

Smith RJ, Andorf JL, Huang J, Ryckman K, Sheffield VC, Stone EM, Bair T, Black-Ziegelbein EA, Braun TA, Darbro B, DeLuca AP, Kolbe DL, Scheetz TE, Shearer AE, Sompallae R, Wang K, Bassuk AG, Edens E, Mathews K, Moore SA, Shchelochkov OA, Trapane P, Bossler A, Campbell CA, Heusel JW, Kwitek A, Maga T, Panzer K, Wassink T, Van Daele D, Azaiez H, Booth K, Meyer N, Segal MM, Williams MS, Tromp G, White P, Corsmeier D, Fitzgerald-Butt S, Herman G, Lamb-Thrush D, McBride KL, Newsom D, Pierson CR, Rakowsky AT, Maver A, Lovrečić L, Palandačić A, Peterlin B, Torkamani A, Wedell A, Huss M, Alexeyenko A, Lindvall JM, Magnusson M, Nilsson D, Stranneheim H, Taylan F, Gilissen C, Hoischen A, van Bon B, Yntema H, Nelen M, Zhang W, Sager J, Zhang L, Blair K, Kural D, Cariaso M, Lennon GG, Javed A, Agrawal S, Ng PC, Sandhu KS, Krishna S, Veeramachaneni V, Isakov O, Halperin E, Friedman E, Shomron N, Glusman G, Roach JC, Caballero J, Cox HC, Mauldin D, Ament SA, Rowen L, Richards DR, San Lucas FA, Gonzalez-Garay ML, Caskey CT, Bai Y, Huang Y, Fang F, Zhang Y, Wang Z, Barrera J, Garcia-Lobo JM, González-Lamuño D, Llorca J, Rodriguez MC, Varela I, Reese MG, De La Vega FM, Kiruluta E, Cargill M, Hart RK, Sorenson JM, Lyon GJ, Stevenson DA, Bray BE, Moore BM, Eilbeck K, Yandell M, Zhao H, Hou L, Chen X, Yan X, Chen M, Li C, Yang C, Gunel M, Li P, Kong Y, Alexander AC, Albertyn ZI, Boycott KM, Bulman DE, Gordon PM, Innes AM, Knoppers BM, Majewski J, Marshall CR, Parboosingh JS, Sawyer SL, Samuels ME, Schwartzentruber J, Kohane IS, Margulies DM.
Genome Biol. 2014 Mar 25;15(3):R53.

- **Association and Mutation Analyses of the IRF6 Gene in Families With Nonsyndromic and Syndromic Cleft Lip and/or Cleft Palate.**
  Pegelow M, Koillinen H, Magnusson M, Fransson I, Unneberg P, Kere J, Karsten A, Peyrard-Janvid M.
  Cleft Palate Craniofac J. 2014 Jan;51(1):49-55.
- **Dominant mutations in GRHL3 cause Van der Woude Syndrome and disrupt oral periderm development.**
  Peyrard-Janvid M, Leslie EJ, Kousa YA, Smith TL, Dunnwald M, Magnusson M, Lentz BA, Unneberg P, Fransson I, Koillinen HK, Rautio J, Pegelow M, Karsten A, Basel-Vanagaite L, Gordon W, Andersen B, Svensson T, Murray JC, Cornell RA, Kere J, Schutte BC.
  Am J Hum Genet. 2014 Jan 2;94(1):23-32.

The first gulp from the glass of natural sciences will turn you into an atheist, but at the bottom of the glass God is waiting for you.

Werner Heisenberg

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
HGP	Human genome project
PTV	Protein truncating variant
LoF	Loss of function variant
RD	Rare disease
OMIM	Online mendelian inheritance in man
SNV	Single nucleotide variant
SV	Structural variant
INDEL	Insertion or deletion
CNV	Copy number variation
HPO	Human phenotype ontology
MPS	Massive parallel sequencing
WES	Whole exome sequencing
WGS	Whole genome sequencing
MQ	Mapping quality
BAM	Binary alignment map
VCF	Variant call format
MAF	Minor allele frequency
HWE	Hardy-Weinberg equilibrium
VEP	Variant effect predictor
SO	Sequence ontology
AR Hom	Autosomal recessive homozygous
AR Comp	Autosomar recessive compound
XR	X-linked recessive
AD	Autosomal dominant
AD dn	Autosomal dominant de novo
QC	Quality control
CADD	Combined annotation dependent depletion
VUS	Variant of unknown significance
GMCK-RD	Genomic medicine centre Karolinska – Rare disease
CLI	Command line interface

1 INTRODUCTION

Bioinformatics
The science of collecting and analyzing complex biological data such as DNA and RNA sequences

Genome
All genetic information of an organism, including all genes and the non-coding regions

Human Genome
≈ 3 billion bases of DNA distributed on 23 chromosomes. Includes a little less than 20.000 genes

To understand the complex world around us we represent entities with models that are understandable to humans and sometimes computers. A well-known example is how we learn in school to model an atom like a small solar system with a nucleus that is surrounded by circling electrons. As our understanding of the microcosmos increases we learn how far away this model is from reality, however it is still useful for understanding the larger picture. In molecular biology we represent biological macromolecules such as DNA, RNA and proteins with sequences of letters that correspond to nucleotides or amino acids. These molecules have been shown to hold massive amounts of information. The discipline that evolves around processing this vast amount of data and to increase the understanding of living systems is named **bioinformatics**. In this thesis, I will explore how bioinformatic methods impact clinical genomics and how improvements of these methods increase our understanding about life and disease.

To our current knowledge, deoxyribonucleic acid (DNA) is the only carrier of information that is passed between generations of life, meaning that everything that is needed to build, develop and maintain an organism is stored in this molecule. The total amount of DNA for an organism is called its **genome** and exists as a complete copy in the nucleus of each and every one of the billions of cells that make up an organism. When the double-helix nature of DNA was first discovered in 1953 it led to that Francis Crick, together with James Watson and Maurice Wilkins were awarded the Nobel prize in 1962 *"for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material".* Rosalind Franklin should probably have been credited as well, however she died of cancer 1958 and due to the rules of the Nobel prize she could not be awarded. Only about 50 years later the Human Genome Project (HGP) presented a complete map of the human genome including about 3 billion positions distributed over 23 chromosome pairs. Today, thousands of whole human genomes are sequenced every day. The scientific journey of mapping the human genome is one of humanity's greatest efforts and arguably one of its greatest achievements.

1.1 CENTRAL DOGMA OF MOLECULAR BIOLOGY

This thesis evolves around the biological entities DNA, ribonucleic acid (RNA) and proteins and how changes in these molecules affect humans. DNA is the material that transfers inherited information between generations of all living cells. **DNA** and **RNA** are constructed by only four different, but similar, building blocks called **nucleotides** or **bases**. Each base of DNA and RNA can only be connected to one or two other bases, in that way they form a linear structure that can be read forwards or backwards, we say that DNA and RNA are **linear polymers**. **Proteins** are macromolecules constructed from chains of amino acids and participate in all cellular activity. Simplified, DNA holds the blueprint of proteins and RNA is the messenger molecule that transfers the blueprint from DNA to ribosomes where proteins are constructed.
To understand the content of this thesis it is essential that the reader agrees on *the central dogma of molecular biology, describing the unidirectional flow of genetic information*. This idea was first published by Francis Crick in 1958 where he stated:

" The Central Dogma. This states that once "information" has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein."

Figure 1 – The Central Dogma of Molecular Biology

Arrows illustrate how information flow from DNA to RNA to protein

1.2 THE HUMAN GENOME

There are approximately 3 billion base pairs in the human genome organised into 23 chromosomes. *Autosomes* are the chromosomes for which humans have two copies, or two alleles, for each position. The *sex chromosomes*, X and Y, differ in such a way that females have two versions of X and inherit one copy from each parent and no version of the Y chromosome. Males on the other hand have one version of X that is always inherited from the mother and one version of Y that is inherited from father to son. The genome can roughly be divided into two sets of sequences where one set consists of the protein coding regions, or genes, and the other set comprises non-coding regions, or intergenic regions. Everyone differs from the reference genome in on average 3 million positions: these nucleotide changes are called variants. Out of these variants, about 5000 are private, which means that they are unique to the individual (Anon 2015). On average, there is one variant identified for every eight base pairs in the human exomes collected so far (Exome Aggregation Consortium et al. 2016). The variability, defined as the number of genetic variants, is unevenly distributed over the chromosomes where chromosomes X and Y are least variable. This is most likely due to purifying selection since males only have one copy each of X and Y, which makes them more intolerant to deleterious mutations (Sayres, Lohmueller, and Nielsen 2014; Schaffner 2004). A variant is called **pathogenic**, or disease causing, if it gives rise to a disease and **deleterious** if it reduces organismal fitness (Kircher et al. 2014). A good example to aid in understanding the difference between these two concepts is the BRCA2 gene. BRCA2 is a gene where a heterozygous protein truncating variant (PTV) will cause severe disease late in life, while a homozygous PTV leads to death. The heterozygous variant is pathogenic, but not deleterious.

Several large-scale sequencing initiatives have been completed during the last decade enabling deep insights into the structure and functionality of the human genome (Ameur et al. 2017, Anon 2014, Anon 2015:10, Anon n.d.; Gurdasani et al. 2015; Nagasaki et al. 2015; Telenti et al. 2016). Surprisingly, these large datasets have revealed that any individual carries a fairly large number of **loss-of-function** (LoF) variants, resulting in gene-knockout, indicating functional redundancy for some genes (Narasimhan et al. 2016). It has been estimated that an average human genome carries about 100 LoF variants that will knock out the functionality of about 20 genes (MacArthur et al. 2012; Narasimhan et al. 2016). Following this idea, the information from the sequencing projects described above has enabled the definition of a set of essential genes, that is genes where LoF variants are never or rarely observed since they probably lead to embryonic death or significant drop of fitness.

Loss-of-function and pathogenicity
A loss-of-function (LoF) variant affects the gene in such a way that there is no usable product after translation. A pathogenic variant is a disease-causing variant. However, a LoF variant is not necessarily a disease-causing variant since there are many non-essential genes in the genome. On the other hand, a pathogenic variant does not necessarily have to be a LoF variant, it can disrupt biological functionality in many ways.

1.3 RARE DISEASE

A disease is defined as rare if it affects less than 1 in 2000 individuals (www.eurordis.org) and it is estimated that about 7000 rare diseases (RD) exist (Amberger et al. 2015). Most of these are inherited single gene diseases that get transmitted in a recessive or dominant fashion. Although numbers like these are hard to deduce (Hartley et al. 2018) , what we do know is that, in total, rare diseases affect hundreds of millions of individuals. In a recent update (2019-04-24) of the Online Mendelian Inheritance in Man (OMIM) there were 3696 known single gene disorders. The field is currently in a discovery phase where hundreds of new gene-phenotype relationships and thousands of new variant-phenotype relationships are reported every year (Wenger et al. 2017), even though the pace for (reported) novel findings is decreasing (https://omim.org/statistics/update). Rare diseases were thought to follow the Mendelian laws of inheritance since the hypothesis was that they were always monogenic. Therefore, the terms rare disease, monogenic disease and mendelian disease are used to describe almost the same phenomenon. Today, it is well known that rare diseases can have both *locus heterogeneity* and *allelic heterogeneity.* Locus heterogeneity is the case where variation in multiple loci can cause the same phenotypic effect, like in Bardet-Biedl syndrome where variants in over 20 genes can give rise to the same phenotype (Ece Solmaz et al. 2015). Allelic heterogeneity is when variants at the same locus can give rise to different phenotypes. However, these diseases are still monogenic since only one gene is affected in everyone. Cases of digenic or oligogenic inheritance are also known, but are very rare. The types of mutations that cause rare diseases span the whole spectrum of genomic variation from **single nucleotide variations** (SNV), small **insertions and deletions** (INDEL) to larger **structural variations** (SV), altered **copy number variations** (CNV) and duplications and deletions of whole chromosomes.

Variant types
Due to historical reasons and limitation of sequencing technologies variants are divided into:
SNV – Single base alteration
INDEL – insertion/deletion < 150bp
CNV – segments that are copied or deleted
SV – Large structural aberrations

1.3.1 Rare Disease Genomics in the Clinic

Working in a clinical setting, it is common to make a distinction between a *clinical* investigation and *research*. An individual is considered as a clinical patient, when entering into the health care system and the process starts by clinical investigation/phenotyping. Ideally the symptoms are described in a controlled vocabulary, such as the Human Phenotype Ontology (HPO) (Köhler et al. 2021), with the goal to identify the pathogenic variant and derive a molecular diagnosis. Based on the phenotype, patients will be screened against gene panels including known genes related to phenotypically similar diseases. It is important to understand that in this step the clinician wants to avoid handling information that is not relevant to the disease in question. If no diagnosis has been made, the investigation may transcend into research setting and with informed consent all available genetic information can be considered. The value of integrating modern sequencing technologies into the clinic cannot be overstated. Together with other medical tests, multidisciplinary teams have high opportunities to accurately diagnose patients that in the best case can lead to individualized treatment of disease (Bainbridge et al. 2011). This is particularly important for patients where acute symptoms present early in life and a quick diagnosis can lead to treatment preventing serious handicaps or death (Stranneheim et al. 2014).

1.4 CLINICAL BIOINFORMATIC ANALYSIS

1.4.1 Sequencing

Whole exome
All protein coding parts of the genome. In other words, all exons of all genes. In humans this is about 1% of the genome.

Current sequencing technologies are performed by decomposing the genome into small fragments, called **inserts**, that get partially read by a sequencing instrument, this is also known as *shotgun sequencing*. The parts of the insert that get sequenced are called **reads**. When the fragment is read from both ends, which is often the case, they make up a *read pair*, this is known as **paired end sequencing**. When running a **whole exome** or **whole genome** sequencing experiment today the best practice is to aim for an insert size of 350 base pairs and a read length of 150 bases. The length of the sequenced insert is limited by technology and will probably increase with time. The entire human genome is not completely accessible, even for whole genome sequencing. Approximately 84% of the genome can be sequenced with confidence, including 91.5% of the protein coding parts and 95.2% of the known pathogenic variant positions (Telenti et al. 2016). The reason why some genomic regions are notoriously hard to sequence is mainly due to low complexity regions, such as centromeres, telomeres and other repetitive regions. If the size of the low complexity region is significantly larger than the insert size it will be impossible to deduce the exact origin of the inserts covering the problematic region. Increased read lengths will open for larger insert sizes which in turn will give more completeness of the sequenced genome (Li and Freudenberg 2014).

Sequencing depth
Number of times, on average, that each base in the genome was read during sequencing

When planning to sequence a genome, there is a tradeoff between cost and **sequencing depth**. A higher sequencing depth allows for increased sensitivity and precision with a high variant call set quality but is more expensive to perform. Especially in a clinical setting it is desirable to find the balance between cost versus sensitivity and precision. When aiming for 30X coverage, 95% of the high confidence region described above will be covered to 10X (Telenti et al. 2016), which is usually the minimum number of reads that is needed to make a confident variant call.

Linkage analysis
The process of identifying regions that are more likely to harbor pathogenic variants. Requires multiple affected individuals and preferably a large pedigree tree.

In the early days of clinical genetics there were not many alternatives when searching for mutations explaining a patient phenotype. If the investigator was lucky enough to have multiple affected individuals, **linkage analysis** could be performed (Slatkin 2008) to determine genetic markers that segregated with the disease. Based on linkage analysis, the most likely candidate genes in these regions would be sequenced, one at a time. If only a single subject was available, one had to make a qualified guess as to which genes might harbor the disease-causing variants and **Sanger sequence** those genes sequentially. With the advent of **Massive Parallel Sequencing** (MPS) in 2005 (Shendure et al. 2005), it became possible to sequence a panel of genes and look at a targeted section of the genome at once. The investigator still had to have an idea of where in the genome to search for causative variants based on the symptomatic picture of the patient. The entrance of **whole exome sequencing** (WES) (Ng et al. 2009) was a revolution to the field of inherited disorders. Now the

Sanger sequencing
The predecessor of MPS where one fragment at a time (typically a gene) was sequenced

whole exome, which means all coding parts of the genome, was sequenced in one run on a sequencing machine. This marks the first time that investigators could have a hypothesis free approach to the analysis, shifting the analysis from phenotype driven to genotype driven. When looking at a case with a novel phenotype it was now possible to perform a genotype to phenotype analysis and first determine which mutations from an unbiased perspective look most pathogenic and then functionally validate if there could be a connection from the genotype to the phenotype. WES is still widely used today, but there are some well-known caveats:

- Different types of bias are added to the analysis during the capture and amplification process (Warr et al. 2015)
- Non-coding regions like deep intronic and regulatory variants will not be captured and can therefore not be analysed
- SVs are challenging due to the fact that only small portions of the genome are sequenced and with variable read depth (Tattini, D'Aurizio, and Magi 2015)

The advantage of WES is deep sequencing of the coding part, which includes most disease-causing variants, to a low cost. With the introduction of the Illumina Hiseq X sequencing system in the beginning of 2014, **whole genome sequencing** (WGS) became affordable enough for all centers who could pay upfront for the machines to motivate this as an alternative when trying to diagnose patients with rare disease. The latest instrument from Illumina, the NovaSeq 6000 hit the market in 2017 and pushed the price of sequencing a whole human genome below 1000$ for the first time. Benefits of WGS include even and complete coverage of the whole genome that allows for better SV calling and access to potential disease-causing non-coding variants.

1.4.2 Mapping

Read mapping is a widely used solution to the challenge of having hundreds of millions of unordered reads from a sequenced genome, or part of a genome. This is called mapping, since the reads are aligned back to a **reference genome** - which can be done in an efficient and reliable manner. Each read or read pair if paired end sequencing, is aligned to the reference genome at the position maximizing an alignment score function, effectively highest when the highest number of bases are the same on the read as on the reference. Each mapping will get a Mapping Quality (MQ) score. If many bases differ between the read and the reference or if the read could be mapped to multiple positions, the MQ score will be lower. The advantage of this method is the speed and relatively good mapping accuracy. An Illumina WGS run with an average read depth of 30X, will give rise to about (number of base pairs in a human genome x time each position is read / length of a read):

Reference genome
A fixed sequence of nucleotides that represents the most likely version of an organism's genome.

$$3 \times 10^9 \times 30 \div 150 \approx 6 \times 10^8$$

or 600 million reads, equivalent to 300 million read pairs. With modern algorithms, implemented in alignment tools such as bwa (Li and Durbin 2009), this number of reads can be mapped back to a reference in just a few hours on a modern personal computer. The disadvantage is that the reference genome only represents one version of a human genome with the result that most individual genomes will differ a lot from the reference genome. This leads to many poorly mapped and unmapped reads, which are still biologically relevant. To achieve a representation that is closer to the biological truth one can assemble all the reads together in an all-versus-all read comparison (Sohn and Nam 2018) to recreate the original sequences. Unfortunately, this is not feasible today since these algorithms take

days and require heavy computational power for just a single sample. The final result after mapping is a file in the Binary Alignment Map (BAM) (Li et al. 2009) format that describes the best alignment location(s) in the used reference genome and the MQ of each mapped read.

1.4.3 Variant Calling

When reads have been aligned to the reference genome it is time to find out where and how the aligned representation of the genome differs from the reference genome. Each segment of the genome that differs from the reference is referred to as a **variant**, otherwise we call it **homozygous reference**. The variants vary in size: from SNVs that are one nucleotide changes; to INDELS where several bases are inserted or deleted; to larger structural changes in the genome, SVs. All called variants from the alignment are collected in a Variant Call Format (VCF) file. It contains the information we have at this point for each variant in the genome, for instance, the variant call quality, which is a measurement of the certainty of the call. The concept of a variant can be confusing since all it tells us is that a part of the genome differs from the reference genome, though we do not know how common the sequence is in a specific population. There are many genomic positions and differing segments that are not in the human reference genome, although these can be common in a population. These are referred to as **normal variation** and are not likely to cause rare disease. SNVs and small INDELS are easy to detect with modern applications while SVs are challenging (English et al. 2015), especially with the short-read technology that is mostly used today. It has even been suggested that the quality of called SVs is so low that it is not feasible to use in clinical practice (Telenti et al. 2016). The variant calls of SVs were inconsistent when a sample was replicated 200 times and the called SVs was compared between these. We addressed this problem in **paper II**.

Variant
Position(s) where one or both alleles differ from the reference genome.

Homozygous reference
Position where both alleles are the same as the reference

1.4.4 Annotation

Before interpretation can begin, we need to collect as much relevant information as possible on each variant. This information can then be used in the clinical investigation of the variants and to design algorithms to prioritize variants before manual inspection. The first step is called annotation, where the variants are labelled with information from a range of sources. At this stage in the analysis pipeline, all we know about the variants is what genotype call the included individuals have. A variant call can be:

- *Heterozygous* when the two alleles are different from each other and at least one of the alleles differs from the reference sequence
- *Homozygous* alternative when both alleles have the variant allele
- *Hemizygous* when the single allele on a sex chromosome has the variant allele or the allele is deleted from the homologous chromosome

This zygosity information is recorded in the VCF-files' genotype field. To facilitate the investigation, we need to gather more information about the effects of the variant, for example, how common it is in different populations, how conserved the position is in different species, the predicted functional effect of a variant etc. The aim of the annotation step is to collect as much information as possible on each variant and ultimately try to decide if the variant can have deleterious effects to the patient. Some of the annotations

described here do not at all indicate deleteriousness but are still valuable by allowing investigators to disregard a significant part of the variant set. A disease-causing candidate will almost exclusively be a rare variant with protein altering effects. There are mainly two reasons for this:

I. Protein altering variants are more plausible disease candidates
II. It is still hard to predict the effects of non coding variation

1.4.4.1 Population frequency

Prevalence
How widespread a disease is in the population.

As the frequency of the disease allele is directly related to the **prevalence** of the disease, the rarity of a variant is fundamental in the rare disease context. This implies that common alleles, in any population, cannot be causing rare disease and should therefore be disregarded. The reasoning is that a deleterious variant will be depleted by natural selection due to negative effects on survival and reproductive rates (Tennessen et al. 2012).

When is a variant too common in the population for being disease causing? Thresholds have varied over the years, when the first larger data sets of human variation became available they included thousands of individuals and conservative thresholds were suggested with values between 1% and 0.1% for recessive and dominant diseases, respectively (Bamshad et al. 2011), where a recessive variant is allowed to be more common since it requires two variants to cause disease. In general, a rare variant is defined as a variant with a minor allele frequency (MAF) < 1% in the population. With a resource such as ExAC (Exome Aggregation Consortium et al. 2016), that includes > 60,000 individuals, and gnomAD (Karczewski et al. 2019) (> 100,000 ind.) the possibilities in rare disease open up dramatically as variant frequencies are getting closer to the true population frequencies. Since the prevalence of rare diseases are low, it has been suggested to significantly lower these rare variant thresholds to around 0.01% for any rare disease (Kobayashi et al. 2017). The **Hardy-Weinberg Equilibrium** (HWE) is a principle that states: *"genotype frequencies in a population remain constant between generations in the absence of disturbance by outside factors"* (Edwards 2008), this means that the proportion of heterozygotes and homozygotes should be stable between generations. According to this principle, the relationship between two alleles can be described as $p^2 + 2pq + q^2 = 1$ where p is the frequency of the common and q is the frequency of the rare allele at a given locus. For an allele causing a recessive disease (q), a frequency of 1% in the population thus corresponds to a disease prevalence (q^2) of 1/10 000. When the frequency threshold is low enough most variants could be filtered, simplifying the analysis. For previously known diseases, it is feasible to base the threshold on the prevalence of the disease in question as arbitrary cutoffs are not efficient and often too lenient (Whiffin et al. 2017). One exception to using the low thresholds is the case of **founder mutations**. These are variants that have been enriched in isolated population that stems from a few numbers of individuals, or founders. In these cases where HWE does not apply and founder mutations can be disease causing, despite having a higher frequency in certain populations. However, these mutations are often well known within the healthcare system of each particular population.

1.4.4.2 Conservation

Homology
Genomic elements in different species that are inherited from a common ancestor

Neutral selection
Positions in the genome that are not under evolutionary constraint. Indicating that variation is neither beneficial nor deleterious to the organism.

Important positions in the human genome are evolutionarily conserved. The degree of conservation can be measured by comparing the variability of **homologous positions** in other species. When homologous loci in the genome have the same or very similar bases in other species, these loci are considered as conserved. This assumes that if a locus is under **neutral selection**, we should observe that mutations have been established as species diverge. If the locus is conserved it is probable that introduced variability would reduce rather than improve fitness. The conservation can be estimated by comparing multiple sequence alignments in a range of species. There are some caveats of using conservation as a pathogenicity predictor. The most important is that a highly conserved loci does not necessarily mean that a mutation there will be deleterious. There are numerous examples of positions with human mutations in ultra-conserved regions without any noticeable effect. It could be that the function has changed during evolution and is not as crucial in humans as it is in the other species. One also must make a trade-off in what organisms to include when performing a comparative sequence analysis. The further away two species are in the evolutionary tree, the less likely it will be to find homologous sequences between them. And the closer they are the less the level of conservation will tell us about deleteriousness. As an example, between human and chimpanzee ~98.8% of the nucleotides are conserved (Waterson et al. 2005). There are several conservation based methods that are used in pathogenicity prediction with fairly accurate results, some that strictly quantify conservation such as **gerp++** (Davydov et al. 2010) and others that predict how changes to amino acids affect protein function based on sequence homology, such as **SIFT** (Ng and Henikoff 2003) and **polyPhen** (Adzhubei, Jordan, and Sunyaev 2013).

1.4.4.3 Variant effects

The variant effect describes the impact of a variant on the corresponding reference sequence and its derivatives RNA and proteins. Since the **codons** for all amino acids are known it is straightforward to determine if an exonic variant is **synonymous**, which means no effect on amino acid, or **non-synonymous**, which will change the amino acid. Estimating how an amino acid change affects the function of the protein is harder. Tools such as the **Variant Effect Predictor** (VEP) (McLaren et al. 2016) and **snpEff** (Cingolani et al. 2012) can be used to annotate each variant with its effects. They use terms from a controlled vocabulary called the **Sequence Ontology** (SO) (Eilbeck et al. 2005) to describe the consequence of a variant. Predicting the variant effect is not trivial. One example could be when a variant seems to be synonymous based on the codon alteration while it could be a **splice affecting** variant that dramatically affect the protein sequence. Therefore these tools are variant effect *predictors,* they will sometimes produce different predictions for the same variant (McCarthy et al. 2014) and the use of a common set of terms facilitates comparisons between effect predictors.

Codon
Triplet of nucleotides that codes for one amino acid. The conversion table for all $4^3 = 64$ codons is known.

1.4.4.4 Inheritance

Autosome
All the 22 chromosomes that are not sex chromosomes. Humans have a diploid genome meaning that there are two copies of

The Mendelian rules of inheritance, which were discovered by Gregor Mendel in 1866 (Mendel 1866), are still relevant for monogenic disorders and represent a powerful way to reduce the number of potential disease causing candidates. Clinical phenotype, previous knowledge about the disease and family history indicates what inheritance patterns to look for when analyzing a patient. Inheritance patterns are divided into two groups, **dominant** and **recessive** which behave differently depending on if the relevant locus resides on any of the 22 **autosomes** or the **allosomes**.

Recessive inheritance

In recessive inheritance both alleles of a locus must be disrupted for the disease to occur. The *Autosomal Recessive Homozygous* (**AR hom**) inheritance pattern is followed when all affected individuals in a family are homozygous alternative for a locus. Since all autosomal positions have one version inherited from the father and one from the mother, this means that both parents need to carry the same variant. In the rare disease case, where variants are extremely rare, this event is highly unlikely in a mixed population. However, if a family is **consanguineous**, variants following this pattern are likely candidates for causing the disorder. If two disrupting variants are inherited in *trans*, on different alleles, the *Autosomal Recessive Compound* (**AR comp**) inheritance pattern is followed. This pattern is expected for a recessive disease in a non-consanguineous pedigree.

Allosome
The sex chromosomes which consist of one chromosome pair in humans. Females have XX and males XY

Consanguinity
Two persons are consanguineous if they share the same ancestor. Common in cultures where cousin marriage in encouraged.

A recessive disorder on the X-chromosome follows the *X-linked recessive* (**XR**) inheritance pattern. This will affect males and females in different ways, females can be carriers of the disease and men will always be affected if they inherit the disease-causing allele from the mother. Affected females will always inherit from an affected father and a carrier (heterozygote) or affected (homozygote) mother. In females that hold two copies of X there is a phenomenon known as X-inactivation where, from early development, only one of the copies gets expressed. Since females have twice as many X genes compared to males, this machinery works as a dosage compensation. As a result, females heterozygous for an X-linked disease allele can display variable symptoms depending on the degree of inactivation of the normal versus the mutant allele.

Dominant inheritance

For dominant diseases it is enough to have one mutated allele to be affected. This means that if the disease is inherited, one of the parents must be affected and the disease is following the *Autosomal Dominant* (**AD**) inheritance pattern. In a pedigree analysis this type of disease is highly likely if multiple individuals in several generations are affected. There is also the *Autosomal Dominant De Novo* (**AD dn**) inheritance pattern where a newly arising heterozygous mutation is causing the disease. AD dn is extremely hard to discover if only a single affected individual is studied since there will be millions of heterozygous candidates. If both parents are included almost all of the variants can be dismissed since the

De novo mutation
Genomic variant that is not inherited, instead they are presented as a mutation in the germ cells of the parents or in the fertilized egg.

number of **de novo** mutations will be around 60 in the whole genome with 1-2 affecting the protein coding parts (Acuna-Hidalgo, Veltman, and Hoischen 2016).
Dominant diseases on the sex chromosomes exist, but they are rare. *X linked dominant* (XD) diseases are inherited from an affected father or affected mother, or more often arise *de novo*. *Y linked dominant* (Y) will obviously only affect males and are always inherited from an affected father or de novo, one example is Spermatogenic Failure (MIM:41500) (Tiepolo and Zuffardi 1976).

Reduced penetrance

Biology does not follow any man-made rules and there are always exceptions to constructions like these patterns of inheritance described above. There are several diseases where dominant variants are not always directly disease causing, which means that some individuals carry a "dominant" variant even though they are not affected by the disease. This phenomenon is called *reduced penetrance* and the biology behind it is not always entirely clear. One example where reduced penetrance is explained is inherited Retinoblastoma which is a dominant disease caused by mutations in RB Susceptibility gene (*RB1*). The disease is recessive on the cellular level meaning that both copies need to be lost in the cancer cell. Affected individuals inherit a pathogenic mutation from one of the parents, in a dominant fashion, and get a second hit in the form of a somatic mutation in the cells which give rise to cancer (Knudson 1971). In the cases where reduced penetrance is not understood the explanation could be that some unknown factors or variants need to coexist for the disease to occur, such as non-coding regulatory variants that are not necessarily located in the vicinity of the gene.
Although inheritance patterns are essential when analyzing variants in rare disease the power in reducing variants in the analysis is not well investigated. A British study, which screened 1133 undiagnosed children, found that if all members of parent-child trios with unaffected parents were sequenced, the number of disease candidates could be reduced tenfold (Wright et al. 2015). When one or both parents were affected the number of variants could be reduced 3-times or 1.5-times, respectively. However, it is not clear how these results are affected by sequencing extended families or consanguineous families. It is also urgent to routinely consider more complex situations such as the case where a larger structural variant is in compound with an SNV, or when one of the potential causatives resides in the regulatory region of the gene. We have designed a tool called **Genmod** (Magnusson [2014] 2018) to address these challenges and study the power in reduction of variants for pedigrees with varying number of family members.

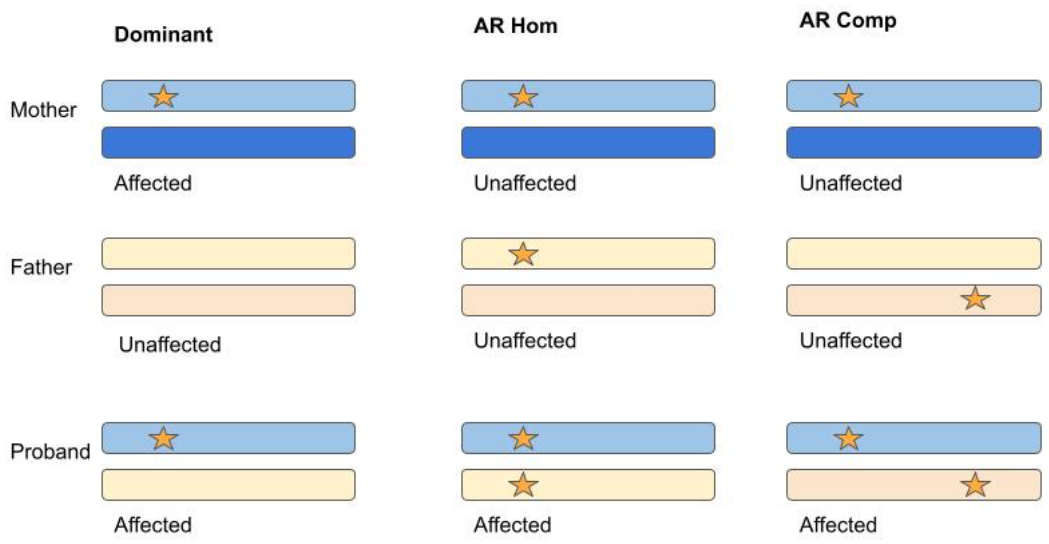

Figure 2 - Inheritance patterns

Illustration of how pathogenic variants are inherited for the different inheritance patterns

1.4.5 Quality Control

The amount of data produced from a MPS experiment is massive and in order to transfer this amount of data to something useful, it must be processed through several bioinformatic tools with the more common steps described above. In each step, there will be errors increasing the number of false positive and false negative variants, so it is of great importance to understand the quality of the data. Many of the bioinformatic tools used to process the data output quality control (QC) reports and there are dedicated QC programs such as **fastQC** (Andrews et al. 2012), to analyse results from sequencing, and **qualimap** (Okonechnikov, Conesa, and García-Alcalde 2016), analyses alignments. With several tools producing metrics on different formats, some on the sample level and others for batches it is a complicated task to manage and report all this information. We addressed this problem in **paper I**.

1.5 VARIANT PREDICTION

As mentioned earlier, sequencing a human genome results in tens of thousands (WES) to millions (WGS) of variants. After annotating all of them with information as described above, it is time to classify all variants as potentially pathogenic or **benign** to reduce the list of candidates that goes to manual interpretation. A perfect classifier would, based on the genotype and phenotype present a list of one or few disease-causing variants, making it easy to solve the case. Unfortunately, we are far away from this scenario today.

Benign variant
Variation that has no impact on health

There is a range of tools that are used to classify variants, some are based on conservation like **gerp++** (Davydov et al. 2010) and **SIFT** (Ng and Henikoff 2003) whereas others try to measure the impact on protein function or protein structure like **PolyPhen2** (Adzhubei et al. 2013). Modern classifiers are based on machine learning algorithms and us many annotations to learn how known pathogenic variants behave compared to known benign such as **Combined Annotation Dependent Depletion** (CADD) (Kircher et al. 2014). Hard filters have been used extensively in the past, today this is seen as a crude way to handle the variants for several reasons:

- It is likely that variants that fall out of a filter threshold could be interesting (eg. false negatives)
- Choosing a threshold is often a balance between **sensitivity** and **precision**. A low threshold will increase the true positive rate, but lead to a large candidate list with many potentially false positive variants. A stringent threshold yields a manageable number of variants, but could mean that the causative variant is filtered out.

The problem with using a classifier is mainly the ascertainment bias that is unavoidable. There are a few numbers of known disease-causing variants compared to the millions of benign. This is not an ideal situation to apply a machine learning algorithm to. Moreover, this approach suffers from a circular proof problem, since there is only a limited set of known pathogenic variants, found by previous methods that searched for a certain type of variants. Each algorithm separates a part of these variants as a training dataset and the remaining is left to study how the algorithm behaves, the validation set. When classifiers are trained at finding variants with the same properties as the ones already found they will have a hard time finding new types of variants (Grimm et al. 2015).

1.6 VARIANT INTERPRETATION

With as much essential information about the variants gathered as possible, the investigator is presented a list of prioritized variants with many types of relevant information. It is now time to vet the variants based on all the parameters described above. This is done in a pedigree context where parameters such as how a variant segregates in the family can cause a variant that looks pathogenic to be highly interesting in one case or dismissed in another. It is of high value to share manually derived classifications and there are many initiatives taken to solve this problem where the most successful and widely used is **ClinVar** (Landrum et al. 2018). ClinVar is a curated public resource where researchers submit their classification and support for evidence to make them publicly available. Variant classifications are rated based on number of submitters, if the classifications agree, if they have been reviewed by an expert panel etc. To rate the clinical significance of a variant there are classification guidelines presented jointly by the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) (Richards et al. 2015) known as the ACMG-guidelines. Based on some defined criteria, a variant is categorized as *pathogenic, likely pathogenic, likely benign, benign* or *variant of unknown significance* (VUS). This formalized way of classifying is sometimes not enough, and investigators will look at more criteria e.g., gene information, what pathways that are affected, biology, literature search, etc. This is by far the most time-consuming step in the whole analysis process described above and tools that help structuring the interpretation process are of great importance. There are a few solutions available via web browsers. Some are open source such as **iobio** (Miller et al. 2014) and **Variation Viewer** (Harrison et al. 2016) while other commonly used commercial solutions are the Agilent Alissa Clinical Informatics Platform and Ingenuity Variant Analysis by QIAGEN bioinformatics. However, there is a need for improved solutions that enable larger case-control and cohort studies (Eilbeck, Quinlan, and Yandell 2017). We have developed and implemented a solution called **Scout** (Andeer et al. [2014] 2021) that has been growing organically over the last 5 years. Scout is a web-based solution where multiple labs can access thousands of cases and analyse them individually or perform aggregated analyses where for example phenotype terms can be considered. This work has been critical for clinical implementation of MPS and is described more in **paper III.** Scout is built to handle issues such as data security and privacy while allowing users to share information with other trusted users.

1.7 FUNCTIONAL VALIDATION

Even after making a novel finding in a rare disease case, the road can be long to strengthen the case that the variant is disease causing. In many cases the complete aetiology of the disease is not known, which means that the condition needs to be delineated by arguing how a particular genotype gives rise to the phenotype or, in other words, to establish a genotype-phenotype connection. In these cases, the only way to *show* how a genotype impacts the phenotype is to use functional studies. Functional studies can be done in a range of ways on different levels of complexity, from studying how RNA expression is affected by genomic variants to constructing disease models using patient cells or model organisms. For some disease groups, such as inborn errors of metabolism or primary immunodeficiencies, clinical diagnostics rely heavily on detailed biochemical or immunological investigations. In these cases, the combined analysis of genomic and functional data provides a strong advantage when homing in on the underlying genetic defect, as described in **paper III**.

1.8 SHARING DATA

To functionally validate a genetic finding is often difficult, expensive and sometimes not even possible for several reasons. On the other hand, if there are other patients that share the phenotypes and have a similar genotype the case that the variant in question causes disease becomes much stronger. As of today, there are hundreds of thousands rare disease genomes and exomes sequenced every year. Most of this data are siloed and never leave the lab. To share this information and group cases in cohorts would increase the power in the analysis. The big hurdle of sharing is that the integrity of a patient can be compromised. Today there are hard regulations on what is allowed to share and how to share (https://gdpr-info.eu/). The legal systems have a hard time to catch up with the pace of advances in technology and allowing its implementation into healthcare and often choose to be restrictive when decisions must be made. Sharing can be done on different levels, ranging from single variants to whole datasets including genotype and phenotype. Solutions that exists today include the **beacon network** (Fiume et al. 2019) where specific variants can be searched via "beacons" to find other cases with same causatives and ClinVar (Landrum et al. 2016) where variant classifications are shared and curated in a public resource. There are also commercial resources like the **Human Gene Mutation Database** (HGMD) (Stenson et al. 2017) where users need to pay for a license to get access to the curated database. **The Matchmaker Exchange** (MME) (Philippakis et al. 2015) is an open-source initiative to link labs with case information on a global scale; MME is built and maintained by 7 organizations that contribute with about 70,000 rare disease cases (March 2019). It is possible to share both genotype and/or phenotype which makes it a powerful tool in the search for similar patients or related genetic findings. Our in-house decision support tool Scout is a node in the MME-network (**paper III**). Scout also enables data sharing via Beacon and facilitates reporting of findings to ClinVar.

2 RESEARCH AIMS

The overall aim of this thesis was to develop tools that enable and improve large scale whole genome sequencing analysis in a clinical setting. Furthermore, the aim was to apply these improvements in a context where rare diseases are being analysed routinely using WGS.

The specific aims are:

- To improve components of pipelines for analysing genomic data (**Papers I-II**)
- To study how whole genome sequencing impacts the diagnostic rate for rare genetic disorders in a clinical setting (**Paper III**)
- To identify the cause of and delineate a Mendelian condition using our in-house developed bioinformatic tools (**Paper IV**)

3 MATERIALS AND METHODS

3.1 PATIENTS AND CLINICAL DATA

In **paper I,** we present MultiQC which is a tool to gather reports for quality control. No patient samples were included in this paper.

In **paper II,** we present the tool LoqusDB and show how information from local populations improve rare disease analysis. Whole genomes from 1000 individuals were collected from a public dataset named the **SweGen** cohort, presented in (Ameur et al. 2017). The SweGen cohort is a collection of individuals from the Swedish twin registry that was chosen as a cross-section of the Swedish population ranging from south to north.

All patients in **paper III** were collected by three clinics that together make up the Genomic Medicine Centre Karolinska – Rare Disease (GMCK-RD) at the Karolinska University Hospital, Stockholm and sequenced at the Clinical Genomics facility in Stockholm. We name this set of patients the **ClinGen** cohort. The ClinGen cohort consists of whole genome sequencing data from 4437 individuals from 3219 rare disease cases. For some of the cases additional family members were included in the analysis, explaining the difference in number of cases/individuals.

In **paper IV,** we delineate a previously unknown Mendelian condition and show what genomic mutation caused the disease. All members of the family analysed in this study were part of the ClinGen cohort.

All individuals, or legal guardians, who participated in the studies (**paper II-IV**) provided informed consent. The studies were approved by the Regional Ethical Review Board in Stockholm.

3.2 DNA PREPARATION AND SEQUENCING

DNA from all samples in the SweGen cohort and ~95% in ClinGen cohort was extracted from blood, the remaining samples were taken from other tissues such as muscle or fetal tissue. Sample DNA was fragmented using Covaris E220 to an insert size of approximately 350 base pairs. The fragmented DNA was converted to sequencing libraries using PCR free sample preparation for paired end sequencing with a read length 150 bases, following the instructions from the manufacturer. Where the amount of DNA was enough, the Illumina TruSeq DNA PCR free sample kit was applied and for the few cases in the ClinGen cohort where the amount of DNA was low the Lucigen NxSeq AmpFree Low DNA protocol was used. Sequencing of the samples in the SweGen cohort was executed on the Illumina HiSeq X instrument. This was also true for all samples in the ClinGen cohort processed before December 2018 (n=2866). The rest of the samples in the ClinGen cohort (n=1571) were sequenced on the Illumina NovaSeq 6000 instrument.

3.3 DATA ANALYSIS

Even though the bioinformatic analysis of whole genome sequencing data follows the same pattern there are often slight differences between pipelines and therefore how the data is being processed. All individuals in the SweGen cohort were sequenced using a pipeline named **Piper** (https://github.com/johandahlberg/piper), while the ClinGen cohort samples

where all analysed with our in-house developed **Mutation Identification Pipeline** (MIP) (Stranneheim et al. 2014). In both pipelines, MultiQC was run to monitor that every step passes the quality control and to discover batch effects.

3.3.1 Alignment and variant calling

After demultiplexing of the raw information from the sequencing machines the unsorted reads from each sample ends up in two FASTQ files. All reads in both cohorts were mapped using **BWA** (Li and Durbin 2009) to reference genome GRCh37 utilizing the bwa-mem command. Single nucleotide variants and short insertions and deletions were called using GATK following the best practice workflow. Manta was used to call SVs in both cohorts, however in the ClinGen cohort the structural variant calls were complemented using CNVnator and TIDDIT as well, the results from these callers were combined using SVDB. In the ClinGen cohort repeat expansions were called using the software ExpansionHunter (Dolzhenko et al. 2019). The tool VT (Tan, Abecasis, and Kang 2015) was used to decompose multi allelic variants and to normalize INDELs.

3.3.2 Variant annotation

The predicted impact of a mutation, or the variant effect, as well as common resources including gnomAD, SIFT etc. were annotated for SNV/INDELs in both cohorts using VEP and the SVs were annotated using SVDB. SNVs in the ClinGen cohort were also annotated using Vcfanno, local variant observations using Loqusdb and genetic models in the same cohort were annotated using Genmod. Moreover, all variants in the ClinGen cohort were scored and ranked using a weighted sum model adapted to rare disease, this model was implemented and annotated using Genmod. Since there are discrepancies in how SNV/INDELS and SVs are annotated there where two different scoring models for the two different types of variants, these score models are available on GitHub (https://github.com/Clinical-Genomics/reference- files/tree/master/rare-disease/rank_model). The method of scoring variants and rank them based on the score is an essential part of how genetic analysis is being performed at Clinical Genomics. This allows the clinicians to examine the variants, starting with the ones that have a high potential of being disease causing and work their way down the list with confidence that all genomic variants from the original analysis are available for investigation, nothing has been filtered away except for “normal variation”, meaning variants with a population frequency above 40%.

3.3.3 Gene panels

One of the most effective ways to reduce the number of candidate variants is to search the genomic regions where mutations are known to cause disease, these regions are almost exclusively protein coding genes. The genes associated to a certain disease or disease group are gathered into gene panels, the gene panels are later used as a hard filter restricting the analysis so that the investigator can focus on the region most likely to harbour disease causing mutations. This is also a means to avoid incidental findings by narrowing the analysis to only the genes of choice.
The fact that certain genes are associated to specific phenotypes is utilized by phenotype ontologies such as HPO, where phenotype terms are associated to a set of genes. When presented with a phenotypic picture that does not resemble any known condition. It is possible to generate an individualized gene panel based on the phenotype terms used to describe a patient. Tools such as **phenomizer** (Köhler et al. 2009) perform a statistical test based on phenotype terms and suggest similar mendelian conditions. Phenomizer also

presents a gene panel that associates genes with the phenotype. The procedure described above, as well as phenomizer, is utilized in Scout (**paper III)**.
The specialist clinics at GMCK-RD curates gene panels that are specific for their respective area of expertise, these panels are shared within the collaboration and used extensively during the analysis of rare disease cases presented in **paper III**. Originally developed in the genomics England initiative there is a public resource called **PanelApp** (Martin et al. 2019) where researchers and clinicians from all over the world collaborate on curating gene panels. In **paper II,** the Intellectual Disability (ID) panel from PanelApp was used to illustrate the usage of LoqusDB in a clinical setting. Gene panels from PanelApp were also used to some extent in **paper III**. When performing more explorative analyses there is a need to include all known disease genes. A gene panel with *all* genes known to be associated with monogenic diseases was created by including all genes from OMIM where a disease relationship was either "established" or "provisional". This resulted in a panel with 3756 genes (latest update 27/8-21 includes 4199 genes) called OMIM-morbid and was used in **paper II**, **paper III** and **paper IV**.

3.3.4 Quality control

Monitoring the quality of the analysis steps and ensuring the integrity of samples are crucial. Most of the steps of the bioinformatic pipelines used in this work produce quality metrics and for every analysis in **paper II**, **paper III** and **paper IV** MultiQC, presented in **paper I**, was run after completed analysis to validate that the quality is high enough. To ensure that the reported gender of a sample corresponds to what is revealed in the DNA, a sex check as well as a coverage report was produced for all samples in **paper III** using Chanjo (Andeer et al. 2020). Chanjo is a command line base software that was developed at Clinical Genomics for persistent storage of coverage data, it is mainly used to summarize coverage information and produce coverage reports on gene panels.
Moreover, to ensure that familial relationship is correct the tools Peddy and Plink were run on all trios and extended families included in **paper III**. To minimize the risk of sample mix-up a small sample of DNA was taken before library prep and sent to a third-party provider that does genotyping using MassARRAY technology. These results were then compared to the in-house sequence data produced at Clinical Genomics.

3.3.5 Variant analysis

To determine if a variant is potentially disease causing in a particular case, variants were vetted using the in-house developed decision support tool Scout. Scout was developed to be used in an environment where multiple individuals from several different clinics collaborate on thousands of cases. Scout combines pre-annotated information from the analysis pipeline with links to sources on the internet as well as information that clinicians themselves add to the cases and variants. Cases go through a 3-step procedure during analysis, this procedure applies to all samples in **paper III** (including the family in **paper IV**)

- In the first step the number of variants to investigate is constrained to be included in only the most relevant genes, these are genes previously known to be associated to phenotypes that overlap with the patient phenotype. This reduces the risk of incidental findings.
- In the second step the data gets shared between the other nodes within GMCK-RD where teams of different clinical expertise further analyze the data.
- In the third and final step the case goes from being "clinical" to "research" and access to the whole genome is opened for analysis and potential discovery of new unknown disorders.

For the last step there are several methods available to reach out to the global community by utilizing modules in Scout that allows for data sharing. The services that are currently available for matching similar phenotypes or genotypes are Beacon and Matchmaker exchange

3.3.6 Data Sharing

There are efforts being made for sharing results back to the community by publishing the findings and analyses in a public fashion. With regular intervals the clinics gather their latest result and report them to ClinVar, making it possible for the global community to take part of the results. Scout has been integral for sharing data within GMCK-RD, over time modules have been added that enable the sharing of case- and variant information via MME, the Beacon network and ClinVar.

4 RESULTS

Several software tools have been implemented during the course of this thesis work, many that have been mentioned either as articles themselves, **paper I** (MultiQC) and **paper II** (LoqusDB), while others are mentioned within articles, these include Genmod, Scout, Stranger, Chanjo and MutAcc (**paper III** and **IV**). In addition to the tools that are explicitly mentioned in the articles there is a plethora of applications developed to enable the flow of data in a sequencing facility that processes and analyzes massive amounts of data. It is important to stress that the implementation of WGS on in the clinical setting is a great challenge and these tools are components of a larger complex system that together have solved this challenge.

All these tools are implemented in the programming language Python and adapted to the later versions above **3.6**. Most of the tools use a CLI for communicating with the user. In some cases, especially for the internal works, the user is often another application that is part of an automated process, in these cases the interface is a REST API. The code is open source and is hosted on GitHub with the intention to encourage collaboration on the code and the tools themselves. As an example, the decision support tool Scout has a lively community where users, both in the form of clinicians and developers, highlight issues, make requests and improve the code base. The power of collaboration should not be understated, this is the most important part of Scout's success. MultiQC was created with collaboration in mind from the beginning, design choices and usage have all been adapted to simplify for others. The purpose of MultiQC is to gather data from the multiple steps in bioinformatic analyses and wrap all the information into one report. Since the number of tools to use is ever growing it would not be possible for one or a few maintainers to adapt MultiQC to the needs of all. Fortunately, it seems like it is straightforward how to extend MultiQC considering the growth pace of the number of modules.

4.1 PAPER I

When running bioinformatic analyses many tools are executed in multiple steps, each of these steps alter the data in some way and they often produce a QC report so that the user can monitor the process. This leaves the user with several reports making it difficult to oversee if the analysis was successful. To address this problem, we developed MultiQC (**paper I**) - a tool to gather quality metrics and visualize them in one report. MultiQC is an established tool used by labs all over the world with a living community that continuously adds modules with support for more bioinformatic tools.

MultiQC is a command line tool that collects information from multiple bioinformatic pipeline components and summarizes the information in a beautiful report. These reports are an outstanding way of discovering outliers and batch effects when running analyses on multiple samples. Up until today MultiQC has more than 100 modules developed by researchers from all over the world, indicating how easy it is to extend the tool. MultiQC is nowadays a standard component of almost every bioinformatic pipeline, as of August 2021 MultiQC is being run about 100k times every day.

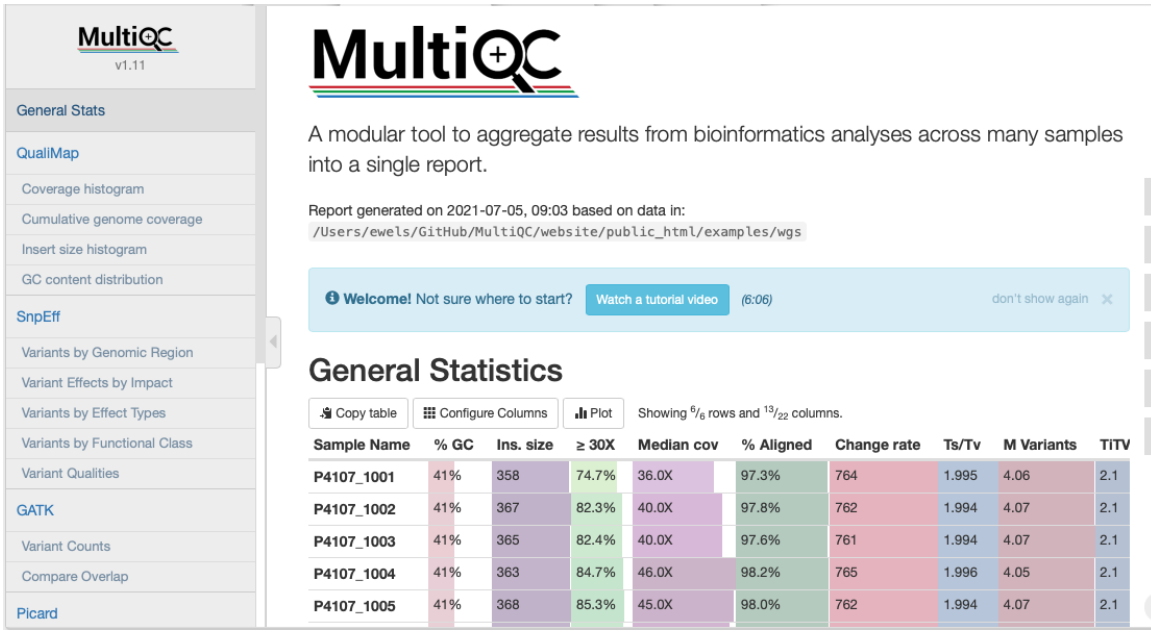

Figure 3 - A MultiQC example report

4.2 PAPER II

Population frequencies are essential to rare disease analysis - we developed a tool called **Loqusdb** to keep track of local variant observations. This will aid in the analysis by highlighting problematic sites due to the local setting and provide detailed information about the local population that might be lacking in public databases.
In **paper II** we study the added value of keeping a database of local population variation for different types of genetic variation, we also present LoqusDB to maintain such a database. To study the value of a local variant database we randomly removed 98 individuals from the SweGen cohort and used the remaining 888 to build a database that represents genetic population variation in Sweden. We call this database **SweGenDB**. 14 individuals were removed from the experiment due to problems during preprocessing. Finally, we annotated all genetic variants in the 98 individuals with the observations from the SweGenDB as well as the population frequencies from the largest public dataset available, gnomAD. Results were compared by counting how many variants we could dismiss when using a threshold of population frequency above 1%. To illustrate the differences in number of filtered variants we compared three different combinations of databases:

1) Only gnomAD
2) Only SweGenDB
3) Any of the two databases

Due to their different nature, we separated SNVs and SVs and performed two parallel analyses.

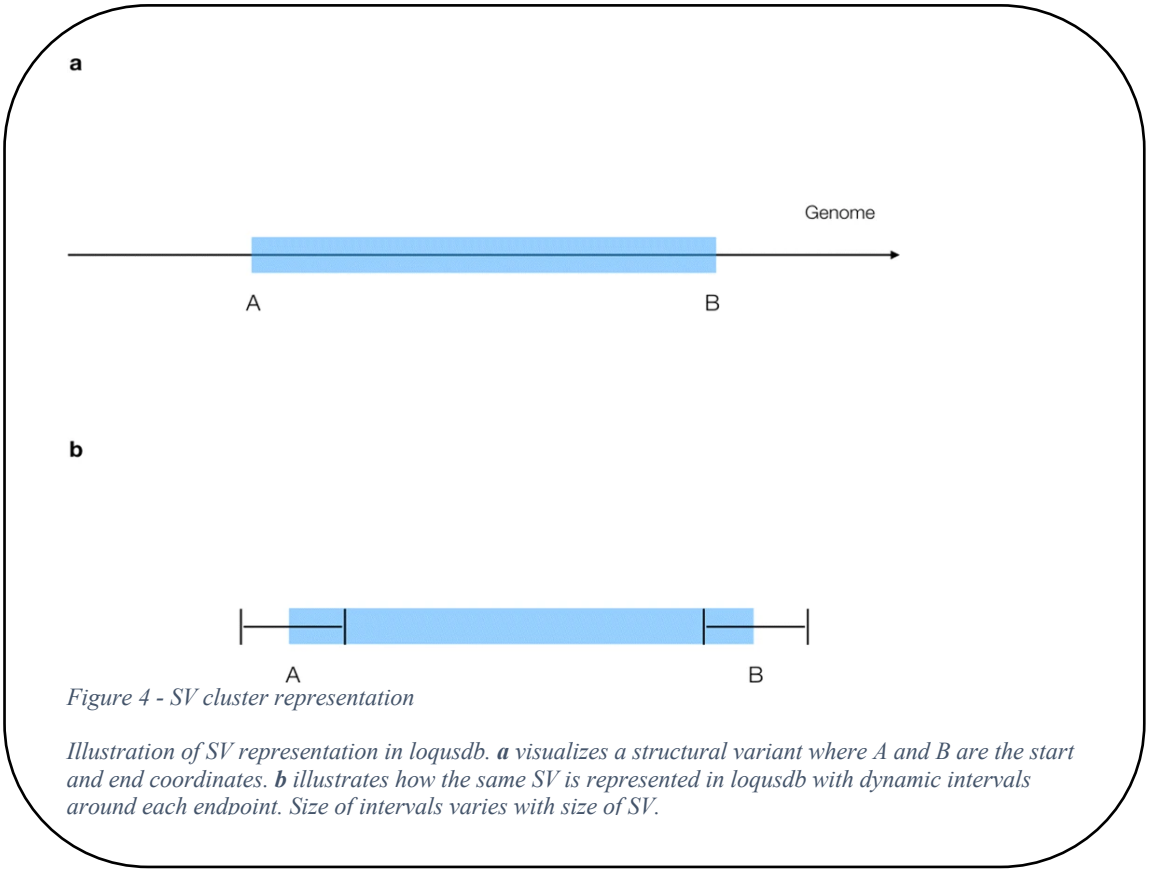

Figure 4 - SV cluster representation

*Illustration of SV representation in loqusdb. **a** visualizes a structural variant where A and B are the start and end coordinates. **b** illustrates how the same SV is represented in loqusdb with dynamic intervals around each endpoint. Size of intervals varies with size of SV.*

Databases are constructed by loading all variants from an analysis using the CLI included in LoqusDB. SNVs/INDELs are straight forward, if the coordinates and the nucleotide change is identical to a previous finding the observation count is increased otherwise a new object is created. However, SVs are more complicated since current short read sequencing technologies are not optimal to accurately identify large genomic aberrations, resulting in inaccurate and variable representations of SVs. LoqusDB handles this problem by allowing “floating” representations of structural variants, meaning that the start and end coordinates are allowed to differ within a defined interval. The size of the interval is proportional to the size of the SV in such a way that small SVs have small intervals and larger SVs allow for larger interval sizes. In this way, LoqusDB increases the chance of representing the same biological event even though the sequencing instrument and variant caller failed to do so.

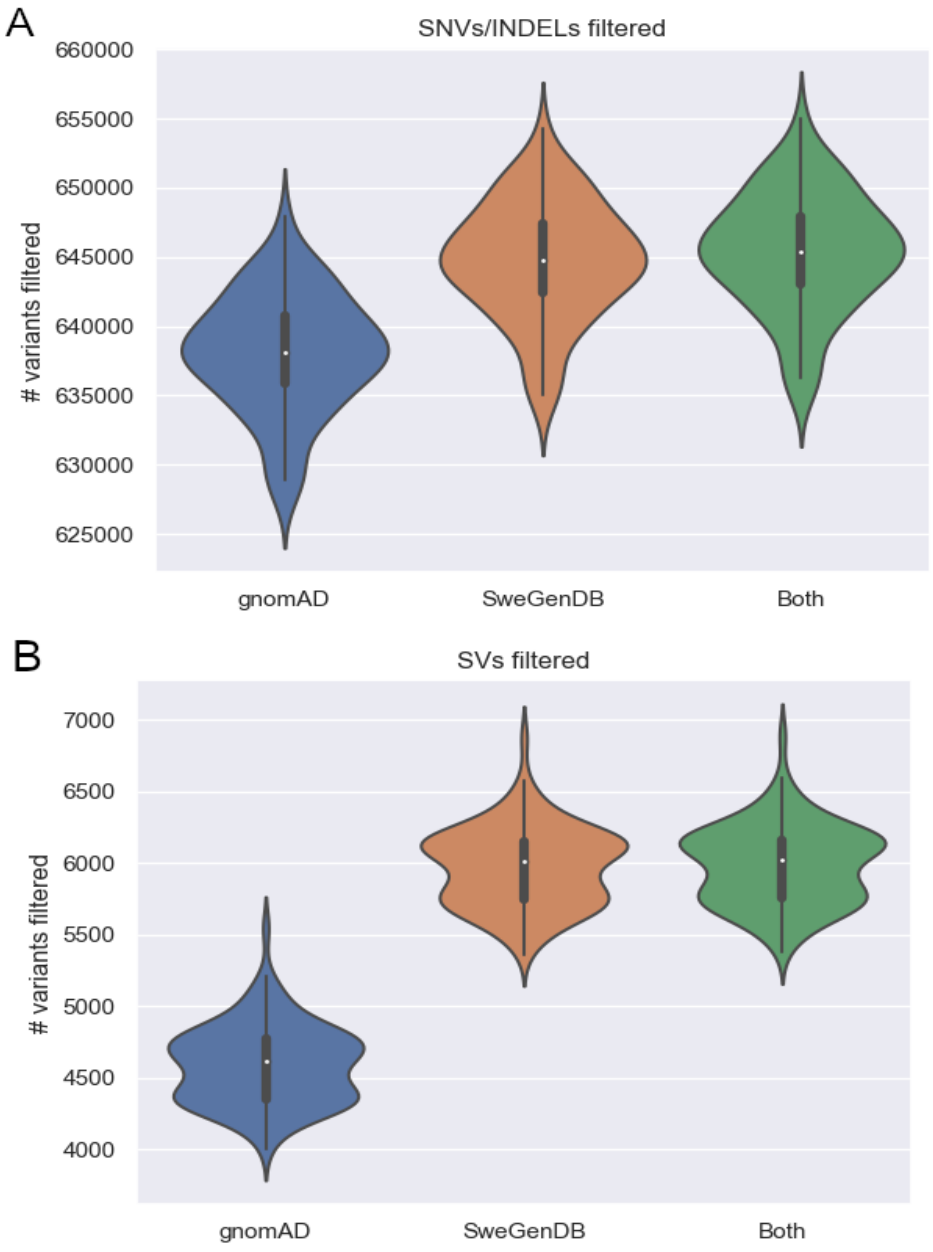

Figure 5 - Comparison of local frequency database versus gnomAD

*Violin plots presenting the number of filtered variants using a local frequeny database (SweGenDB) versus a public database (gnomAD). The local frequency database consists of 888 individuals. The frequency filter was applied on SNV calls (**a**) and SV calls (**b**)*

We found that the number of potentially disease-causing mutations can be significantly reduced in an unbiased way based on how common they are when curating a population variation database with information from local sequences using the method described in **paper II**.

4.3 PAPER III

In **paper III** we study how the shift to whole genome sequencing in the clinic impacts rare disease patients in the Stockholm region. In total 4437 individuals from the ClinGen cohort have been sequenced with WGS, where 3219 where patients and the rest were patient relatives, most often parents. In this cohort 40% (n=1285) of the patients received a molecular diagnosis between the years 2015-2019. The solve rate varies depending on the patient's phenotype group. In line with previous studies (Wright, FitzPatrick, and Firth 2018) we observed that the solve rate varied between 19-54% depending on the phenotype. These disease-causing variants originate from 754 genes indicating vast heterogeneity. In a few cases recurrent variants were observed where the top recurrent genes where *COL2A1* and *FKRP* (n=12), *MECP2* (n=11) and *DYNC1H1* (n=10). However, most of the genes with disease causing variants were singletons (n=496 or 66%). In some cases, recurrent variants were observed, both known founder mutations, such as c.826C>A in *FKRP* and a known repeat expansion in *RFC1*, and mutations previously unknown to be recurrent such as c.1969G>T in *CAPN1* that was found in two unrelated individuals from the same region outside of Sweden. SV variant calling was not introduced as a part of the clinical test up until 2017 due to the complex nature and lack of resources to annotate them with relevant information. In the beginning SVs were mostly investigated if there was an SNV variant in a gene of interest that could not by itself explain the phenotype. However, there is a non-negligible number (n=64) of cases in the ClinGen cohort that have been explained by SVs. Most of these (70%) are copy number variations (CNVs), some are STR expansions, and the remaining are genomic rearrangements.

When studying the inheritance patterns of disease-causing mutations, it was straightforward in families where both parents are sequenced (16%). However, most of the cases (84%) were single sample analyses or singletons. Whenever data was unavailable, a follow up Sanger sequencing of the region of interest was performed in both parents to confirm pattern of inheritance. This resulted in that the pattern of inheritance could be established in 870 (68%) of the disease-causing variants revealing that the most common pattern was autosomal recessive inheritance (54%) followed by autosomal dominant and x-linked *de novo* (27%), autosomal dominant (12%), inherited x-linked (5%) and mitochondrial inheritance (1%).

For some cases that were still unsolved after the initial analysis and where there was a strong suspicion of a genetic explanation for the disease the analysis was expanded to include the whole genome. Up until 2019, this has led to 17 novel disease gene discoveries and/or mechanisms for disease. This type of analysis is time consuming and new discoveries are continuously being done. The study in **paper IV** is one example of this situation, however since the finding was done after 2019 it is not counted among the 17 mentioned above.

4.4 PAPER IV

In paper IV we performed WGS of all five family members in a family with a previously unknown Mendelian condition. The proband is a girl with non-consanguineous parents who has a syndromic neurodevelopmental disorder. In the neonatal period she, in resemblance to a diseased older sister, presented with hypotonia (abnormally low muscle tone), apneas (lack of breathing), disappearance of the Moro reflex as well as *Staphylococcus aureus* parotitis (bacterial infection of the major salivary glands). At the time of the study the girl was 8 years of age and the condition had stabilized, present symptoms are severe developmental delay, hearing impairment, gastrointestinal problems, and a striking lack of tear fluid, saliva and sweat. This causes the respiratory mucosa to be dry and requires frequent inhalation of hypertonic sodium chloride to prevent life threatening breathing stops.

The routine clinical analysis of known genes from the in-house curated gene panels gave no results and the family was opened to "research analysis" expanding the study to include all genes. We found two variants inherited in a compound heterozygote fashion in *SLC12A2*, including an SNV (c.2006-1G>A) which was predicted to alter a canonical splice site and a one base pair deletion (c.1431delT) causing a frameshift in exon 8. From a bioinformatic perspective, both variants follow the pattern of highly deleterious mutations by causing severe loss-of-function of the protein product.

This disease gene was not described in the literature at the time of analysis. However, one patient was later reported with a phenotype referred to as the "Kilquist syndrome" (MIM:609080), resulting from a homozygous 22-kb deletion which caused a splicing defect of *SLC12A2*. No SLC12A2 (NKCC1) protein could be found in the patient, which in this case was a 5,5-year-old boy. The phenotype of this case was similar to our patient's, and there was also a close resemblance to the phenotypes of corresponding mouse models. Taken together, this establishes biallelic loss-of-function variants in *SLC12A2* as causing a distinct novel human disease.

5 DISCUSSION AND FUTURE PERSPECTIVES

A recent publication estimates that there are at least ~ 6,100 – 14,400 rare disorders remaining to be discovered (Bamshad, Nickerson, and Chong 2019). Numbers like these will increase the pressure on finding new methods to discover and communicate findings on researchers and clinicians. Even though whole genome sequencing is entering the clinic, only a small part of the data is understandable now, most variants will be labeled as VUS in the following years to come. To increase the diagnostic yield from today's ~30% there is a need for more multinational initiatives, improvements in technology and analysis tools. Initiatives such as the International Rare Disease Research Consortium (IRDiRC) are trying to push the field forward and have stated the goal that all genetic diseases will be diagnosed within a foreseeable future and suggest actions that are necessary to get there (Boycott et al. 2017). Laboratories need resources so that every patient can become a research patient, expanding the analysis from known disease genes to the whole genome. Negative cases need to be reanalyzed in a controlled way, as it has been shown that reanalysis of 1-3 years old WES data increased diagnostic yield with 10% (Wenger et al. 2017). However, the million-dollar question is:

What is explaining the last 50-70% of the undiagnosed rare disease cases?

Of course, it is not certain that all these diseases have a genetic cause so it is unlikely that diagnostic rate will ever be 100% for genetic tests in the future. In fact, it is likely that most of the unsolved genetic diseases could be explained by the types of variants that we already today have methods to find, the protein coding LoF variants. Even if laboratories do find candidate variants in an increased pace it seems like the reporting of the findings are decreasing (Bamshad et al. 2019). The reason for this change is probably that novel findings do not have the same academic value; it is not worth the effort it takes to collect evidence that is needed to publish.

5.1 TECHNOLOGIES

There are constant improvements and new innovations to the technology in the field of genomics. Current short read technologies are improved by increased read lengths and increased throughput to the same cost. At the same time other ways to access information are becoming more interesting as they take the step from research and enter the clinic. There is fascinating work being done in areas of methylation profiling and proteomics/metabolomics, however I will focus on the technologies that I believe are going to reach the clinics soon and that will have an immediate impact on discovery rates.

5.1.1 RNA-seq

Transcriptome sequencing (or RNA-seq) has a great potential to improve the interpretation of genomic variant findings by studying how RNA expression levels vary. It has been shown how RNA-seq, used as a complement to DNA sequencing, improves the diagnostic yield of rare disease patients for some phenotype groups in a clinical setting (Cummings et al. 2017). One of the advantages with transcriptome sequencing is that it can be performed using similar methods like MPS and be executed on the same instruments as regular DNA sequencing, facilitating for labs to start producing data. However, the drawback of this approach is that many genes have tissue-specific regulation so the results are dependent on what tissues the sequenced sample was taken from. It is not always clear what the relevant tissue would be for a particular syndrome and in some cases the tissue of interest might be very difficult to access.

5.1.2 Long read sequencing

There are two main players in the area of long read sequencing (LRS), Oxford Nanopore (Jain et al. 2016) and SMRT sequencing from PacBio. Short read sequencing technologies are well suited to discover SNVs and short INDELs, however when genomic variants get larger than the read length the accuracy of calling them drops significant (Mahmoud et al. 2019). The same situation applies to complex regions, such as long stretches of repeats. MPS usually have read lengths around 150 bp while the LRS technologies currently have a read length on average around 10-30 kb which makes it more well suited for discovery of SVs and to map complex regions. Challenges with current LRS include:

- Input material: Since insert sizes are much larger it is necessary to use high quality DNA, which is not always available. Also these technologies require larger amounts of DNA for library preparation.
- Error rate: error rates are much higher than MPS, meaning that every time a genomic base is read the risk of reporting the wrong nucleotide is increased.
- Cost: the cost of sequencing a genome is significantly higher with LRS than using MPS

Using LRS in combination with MPS would be an excellent solution to the error rate problem since we can infer the correct readings from the short reads and the structure of the individual genome from LRS. However, until the cost for using LRS is dropping it will most probably stay in the field of research.

5.2 VARIANT TYPES

Except from the most obvious category *unknown disease-causing genes,* there are several plausible explanations to what genetic components that are disease causative. The more complex ideas evolve around compound effects, such as pathogenic combinations of a certain methylation profile together with DNA mutations or oligogenic disease where variants in multiple genes explain the disease. There is little evidence for these suggestions today, however we will probably see an increase of reported cases. Here I will not delve further into the most complex ones and instead mention some words about current challenges.

5.2.1 Mosaicism

Mosaic mutations are mutations that are only found in a proportion of cells in the human body. These mutations arise from de novo changes in early cell divisions in the development of the embryo. There are known cases of mosaic disease causing mutations (Westenfield et al. 2018), but these are hard to detect with a regular genetic test such as WES or WGS due to the low fraction of cells that harbor the mosaic mutation if not sequencing the affected cell type. Hopefully these will be detected if more comprehensive analyses are performed where multiple tissue types are sequenced for every patient.

5.2.2 Structural Variants

Structural variants have been identified to be disease causing for many years (Stankiewicz and Lupski 2010) and, as mentioned in the text above, WGS has improved the possibility to find SVs (Willig et al. 2015). When detecting SVs using short read technologies the algorithm usually searches for three types of phenomenon after read mapping:

- Discordant read pairs: These are reads where the insert size and/or orientation of the reads differ from what was expected.
- Soft clipped reads: Reads where only a part is mapped to the reference
- Split reads: This is the case when one part of the read is mapped to one loci and the rest to another

All these situations are indications of structural variations of the genome compared to the reference. As mentioned earlier there are large parts of the genome that are repetitive, in these regions we will see the above due to the complexity of mapping short reads to these regions. It becomes hard if there are actual structural variants in the repetitive regions and even worse when looking at **repeat expansions**, short sequences that are repeated multiple times where the number of multiples determine if an individual gets affected. The most promising solution for discovery of SVs is without a doubt LRS which have the potential to read through long stretches of the genome and catch full sized SVs and repetitive regions in a single read. However, even if we find all the structural variants in the patients, they are hard to interpret since there are fewer resources, such as population frequency databases, compared to SNVs. Most likely this is one of the categories that will explain more diseases soon.

5.2.3 Silent coding mutations

Synonymous mutation in exons were previously thought to have no effect on the functional consequence. Today we know that these can affect splicing, protein folding, expression levels of the gene etc. There are tools and validation strategies to evaluate synonymous mutations but they pose a challenge to interpret (Hunt et al. 2014). Combining DNA with RNA sequencing is the straightforward strategy for discovery of pathogenic silent coding

mutations. As of today there are no known prioritization tools available that combine these two sources of information; however it is most likely that they will appear in the near future.

5.2.4 Non-Coding Variation

The non-coding part of the genome consists of sequences that have important roles in gene regulation. There are non-coding RNA genes that, unlike regular genes that code for proteins, produce functional RNA molecules. Other examples of interesting non-coding parts of the genome are enhancers, suppressors, transcription factor binding sites and untranslated regions of coding genes. Any variation that affects the binding of proteins of transcription machinery can interfere with the expression of the target gene. These noncoding regulatory sequences are cell-type specific, and identification of these regions require cell-type specific high-throughput experiments. ENCODE (Dunham et al. 2012) and Roadmap Epigenomics (Kundaje et al. 2015) projects identified many of the known regulatory regions. Identification of mutations in these sequences requires multiple layers of information from these types of studies to enable evaluation. Usually, identification of gross deletions or duplications spanning over these sequences are easier compared to identification of SNVs in these sequences. For accurate pathogenicity prediction of non-coding variants it is necessary to perform functional studies such as RNA-seq and proteomics.

5.3 THE IMPORTANCE OF SHARING DATA

Thousands of genomes are sequenced every week in labs all over the world and the pace will most probably increase for many years to come. It has been estimated that by 2025 more than 60 million patients will have had their genomes sequenced (Birney, Vamathevan, and Goodhand 2017). All this information is a potential gold mine if treated in a structured way, using quality-controlled data along with extensive phenotypic information about the patients. That situation would facilitate for sophisticated algorithms to match patients based on both genotypes and phenotypes to find evidence for candidate mutations. Unfortunately, there are some obstacles on the way, where the hardest one has to do with legal issues. Most countries apply strict regulations to protect patient data, restrictions and regulations that have not been adapted to the genomic revolution resulting in those delicate decisions end up on the clinicians table with serious consequences as a potential risk. Even though patients sign informed consents where they comply with data sharing the regulations make it hard for clinicians and researchers to share information outside of the hospital setting. It is often in the patient's interest to share their genomic information since that might improve the chance of finding an explanation to the disease and get them in contact with others in a similar situation. With the lack of solutions in place it has become increasingly popular that families reach out to the community on their own using social media (Might and Wilsey 2014). There are attempts to avoid the regulations by encouraging families to find others in a more controlled way (https://mygene2.org/MyGene2/). However, the preferred road ahead would be to direct more resources towards variant sharing and investigate if more can be done to the legal situation.

6 CONCLUDING REMARKS

I got into the field of genomics about 10 years ago, starting at a research facility today known as the National Genomics Infrastructure (NGI). It was just after WES had been announced and the whole field was in the middle of a technical revolution. It has been a fascinating journey and I feel privileged to have had the opportunity to join this revolution and hopefully been a microscopic part of it as well. My impression is that the velocity of the development and discoveries have slowed down a bit the last 3-4 years. However, the future looks bright, and I believe that we have only seen the beginning of how genomics will impact individual health and health care in the years to come. The technologies described above looks very promising, with LRS having the potential to be most disruptive in a positive way. Apart from that I sincerely hope that the scientific community drop the prestige around academic value in the form of high impact publications, leading to a protectionist way of treating patients and data. Instead, we need to always put the patient and the families first and direct resources towards intelligent data sharing which will have the greatest impact on the whole field. My prediction is that within 10 years most of us will have our genomes sequenced with full access to the data that we can analyze either on our own by utilizing commercial tools or together with a clinician in a health care setting. Cancer patients will have multiple WGS runs performed for both tumor tissue and multiple tissue types for extensive profiling etc. Overall this will be a good thing, however like with all technical advances that happens fast we need education and carefulness.

7 ACKNOWLEDGEMENTS

Unbelievable that I'm sitting here writing this, after 8 years it is coming to an end! I honestly never thought it would happen.
I want to begin by thanking my *dream lineup of supervisors*!
A huge thank you is not enough to the amazing researcher, leader and diplomat **Anna Wedell** for your patience, positivity and availability. Whenever times have been tough and I reach out, you always have meet with positive words and I walk away strengthened again. It has been a pleasure to be a PhD student for you.
Henrik Stranneheim, we have had a long and amazing journey together. Nothing of this would have been possible without you. A big thank you for all your patience and for believing in me from the very beginning. I don't know how many times I have let you down when it comes to estimating when things are going to be done, today I think you have developed a formula for calculating reality based on my estimations.
Last but not least, the walking encyclopedia of bioinformatics and genomics **Daniel Nilsson**! Always positive and ready to answer and discuss any topic. Great karaoke companion and late-night party animal as well.

There have been so many years at Scilife, and I've encountered many fantastic people so I need to do this chronologically.
It all started with **Lasse Arvestad** who, with his inspiring ways, awoke my interest in the field and I suddenly realized what I wanted to do with my life. He directed me to the toughest of schools with one of the nicest teachers **Jens Lagergren.**
Real world bioinformatics started with some shaky steps by analyzing complex HLA sequences in dogs for Mr. Dog himself, **Peter Savolainen**, among many things you convinced me that the true origin of dogs is in south-east-Asia and the danger of organized crime. Keep on fighting!

When I got my first job at NGI **Tomas Svensson** actually offered me a contract, even though I knew nothing. There I met my first mentors **Per Unneberg** and **Mikael Huss**, thank you for all the responsibilities you dared to give me, I failed so often and learned so much. The first time I succeeded was when I got to work with the marvelous **Myriam Peyrard**, it was a pleasure working with you and I got to solve my first rare disease case, it got me hooked. Can you imagine that I now work with your son **Vincent**!?

Collaboration with **CMMS** was intensified and i met so many nice people there, **Michela** even though you drove me crazy sometimes I really like you, **Nicole** thanks for all your help, **Helene**, **Christoph and Anna W**, **Martin**, all those Scout meetings where we did our best to understand each other.

Special thanks to **Tommy Kungen Stödberg** for letting me participate in that very special investigation. And to my new mentor **Virpi** for all the kindness and support.

Don't we all miss the lunches at Smittan together with **Lukas Käll** and my gymnastics friends **Kristoffer Jaojao Sahlin**, **Mattias Polly Frånberg** (you all know who was the strongest…) and **Joel Sjöstrand**, rest in peace.

At this time I also made friends with one of the most brilliant and humble persons I've ever met, **Erik The Mastermind Sjölund** I think it was fantastic that you finally joined us at CG. We will never stop learning from you.

Later on, I got to work with the unpolished diamond **Robin Andeér. I**t was great to get to spend those years together before you realized how good you are. We got help from the database wizard **Mats Dahlberg** and the legendary Scout program started to take form thanks to the two of you.

Phil golden hands Ewels! I will be forever proud in having my name mentioned in the context of MultiQC. It's mind blowing to see how everything you touch become gold. Great collaboration with nice guys **Max** and **Sverker**.

Fulya we met when you arrived to Sweden, it's been a pleasure to work with you for all this time and you have helped me a lot, I will never forget.

We all got recruited by to Clinical Genomics where **Lasse E** was running the show and could fix everything. When are we having cocktails on your balcony again??
Many off us followed on this exciting journey at CG, all the lab ladies **Sophie Sibia** (How could you leave us!?), **Anna Leinfelt** I will miss your kindness and our conversations, with your sons name you will not forget me easily. **Cissi** for super weird, amazing and fun karaoke experiences and much more. **Anna Z, Anna Lyander, Anna Gellerbring**, the fabulous triple A-team.

Jesper näsflöjten Eisfeldt you are one of a kind, keep on with that.

We intensified the collaboration with **Clinical Genetics**, thank you **Anna Lindstrand** and **Ann Nordgren** for positivity and kindness. **Maria P** you are the only reason why I will remember late afternoon vetenskapsteori in Huddinge. **Kicki** och **Helena** you were tough cookies in the beginning but after some time, among the most important and fun customers. **Bianca** you have been an invaluable source of information and friendliness. **Ellika the secret opera diva** and **Anna Hammarsjö** we need to continue teenage kids therapy talk.

Clinical genomics is such a big part of my life thanks to all you colleagues. Thank you, Emma, for years of deeptalk and smalltalk, **K1** the most Swedish Iranian guy around, **Hassan** we have been through a lot together, both ups and downs since we first met 100 years ago. **Jemten** always positive, friendly and up for a chat, when will we get to try that new sauna?? **Backman** I'm happy for the people at your new job that they will have you around and light up their environment. **Emilia** I'm still waiting for you to come back, you belong here. **Barry**, I don't know if I should thank you or hate you for introducing me to the dark world of the souls games or not. **Patrik Sokrates Grenfeldt** we want more answers than questions from such a experienced programmer! **Isak** and **Tanja** keep on rocking in the micro world. **Chiara the sincere Italian**, I hope you will stay and take care of my baby forever. **Eleonora the tall and strong Italian** I'm so glad you found your way back home. **Klättermicke** when are you taking the step into the 21st century, its IDE times now you know.

Maria Chefen Ropat you came in like a whirlwind both on the emotional and professional plane, it has been exciting to get to know you and working with you.

All the brilliant new guy's **Henning, Kalle, Ram, Khurram, Moe, Kristine, Vincent, Vadym, Aswhwini, Moa, Mei**, I'm leaving before you all pass me and start telling me what to do. And all the new people in the lab **Linnea, Sasha, Rasmus, Ida, Susanne, Marcela, Maria, Vasileios,** and others that I forgot names on or mention, please take care of CG and make it the best place to work at. Remember It's up to you!
Karolin get well soon, we think about you.

Valtteri Big Boss Wirta the true Finn who can stay up all night, be first back at work, always happy, always working, always ready for a sauna. I hope that I will get the opportunity to work with you again.

And **Maya** this is hard… I cannot really explain how much you mean to me. But I think you understand. Let's just never stop hiking and talking and hang out.

Final shout out to my family, my supporting parents **Ulla** and **Hans** I love you!

Mormis for saving our lives when times have been difficult, **Petter** for also being there when needed.
To my siblings **Mira, Jonas, Nina** you are the best family anyone could ever ask for. All the others **Zowi and Anders, Siri and Pierre, Theo, Nora, Jonas F, Robin, Janu**, I'm grateful for having you in my life.

Ann I love you with all my heart, I want you to always know that. I look forward to growing old together with you.

Iker my oldest son, I'm so proud of you and how you take care of all of us. Thank you for putting out with me and everything stupid I have done. You will be better.

Inez my little indian viking girl, you are such a brave and fantastic person with an imagination without bounds. Use it!

Abbe the star child, whatever becomes of you it will be amazing. I'm eager to follow your path.

Maji the professor, my guess is that you will be next up doing this but for you it will just be a small step on the road. We all learn from you already; I wonder where it will end…

Babis I've had your wise words in the back of my head during all the time, you are always around.

Ann-Britt Wikström how could anyone complete a PhD without your help? Thanks for all the help and patience. Your crew together with **Therese Kindåker** is invaluable to us.

Ok, now you hopefully got what you came here for, all I'm asking from you is: Please come and dance at my party!

Big hugs to all of you and thank you for the fish.

8 REFERENCES

Acuna-Hidalgo, Rocio, Joris A. Veltman, and Alexander Hoischen. 2016. "New Insights into the Generation and Role of de Novo Mutations in Health and Disease." *Genome Biology* 17(1):241. doi: 10.1186/s13059-016-1110-1.

Adzhubei, Ivan, Daniel M. Jordan, and Shamil R. Sunyaev. 2013. "Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2." *Current Protocols in Human Genetics* 76(1):7.20.1-7.20.41. doi: 10.1002/0471142905.hg0720s76.

Amberger, Joanna S., Carol A. Bocchini, François Schiettecatte, Alan F. Scott, and Ada Hamosh. 2015. "OMIM.Org: Online Mendelian Inheritance in Man (OMIM®), an Online Catalog of Human Genes and Genetic Disorders." *Nucleic Acids Research* 43(D1):D789–98. doi: 10.1093/nar/gku1205.

Ameur, Adam, Johan Dahlberg, Pall Olason, Francesco Vezzi, Robert Karlsson, Marcel Martin, Johan Viklund, Andreas Kusalananda Kähäri, Pär Lundin, Huiwen Che, Jessada Thutkawkorapin, Jesper Eisfeldt, Samuel Lampa, Mats Dahlberg, Jonas Hagberg, Niclas Jareborg, Ulrika Liljedahl, Inger Jonasson, Åsa Johansson, Lars Feuk, Joakim Lundeberg, Ann-Christine Syvänen, Sverker Lundin, Daniel Nilsson, Björn Nystedt, Patrik KE Magnusson, and Ulf Gyllensten. 2017. "SweGen: A Whole-Genome Data Resource of Genetic Variability in a Cross-Section of the Swedish Population." *European Journal of Human Genetics* 25(11):1253–60. doi: 10.1038/ejhg.2017.130.

Andeer, Robin, Mats Dalberg, Mikael Laaksonen, Måns Magnusson, Daniel Nilsson, and Chiara Rasi. [2014] 2021. *Scout* [computer program]. Version 4.39. https://github.com/Clinical-Genomics/scout.

Andeer, Robin, Måns Magnusson, Anna Wedell, and Henrik Stranneheim. 2020. "Chanjo: Clincal Grade Sequence Coverage Analysis."

Andrews, Simon, Felix Segonds.Pichon, Anne Biggins, Laura Krueger, and Christel Wingett. 2012. "Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data." Retrieved September 9, 2021 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Anon. 2014. "Whole-Genome Sequence Variation, Population Structure and Demographic History of the Dutch Population." *Nature Genetics* 46(8):818–25. doi: 10.1038/ng.3021.

Anon. 2015. "The UK10K Project Identifies Rare Variants in Health and Disease." *Nature* 526(7571):82–90. doi: 10.1038/nature14962.

Anon. n.d. "Large-Scale Whole-Genome Sequencing of the Icelandic Population | Nature Genetics." Retrieved March 31, 2020 (https://www-nature-com.proxy.kib.ki.se/articles/ng.3247).

Bainbridge, Matthew N., Wojciech Wiszniewski, David R. Murdock, Jennifer Friedman, Claudia Gonzaga-Jauregui, Irene Newsham, Jeffrey G. Reid, John K. Fink, Margaret B. Morgan, Marie-Claude Gingras, Donna M. Muzny, Linh D. Hoang, Shahed Yousaf, James R. Lupski, and Richard A. Gibbs. 2011. "Whole-Genome

Sequencing for Optimized Patient Management." *Science Translational Medicine* 3(87):87re3-87re3. doi: 10.1126/scitranslmed.3002243.

Bamshad, Michael J., Sarah B. Ng, Abigail W. Bigham, Holly K. Tabor, Mary J. Emond, Deborah A. Nickerson, and Jay Shendure. 2011. "Exome Sequencing as a Tool for Mendelian Disease Gene Discovery." *Nature Reviews Genetics* 12(11):745–55. doi: 10.1038/nrg3031.

Bamshad, Michael J., Deborah A. Nickerson, and Jessica X. Chong. 2019. "Mendelian Gene Discovery: Fast and Furious with No End in Sight." *American Journal of Human Genetics* 105(3):448–55. doi: 10.1016/j.ajhg.2019.07.011.

Birney, Ewan, Jessica Vamathevan, and Peter Goodhand. 2017. *Genomics in Healthcare: GA4GH Looks to 2022*. doi: 10.1101/203554.

Boycott, Kym M., Ana Rath, Jessica X. Chong, Taila Hartley, Fowzan S. Alkuraya, Gareth Baynam, Anthony J. Brookes, Michael Brudno, Angel Carracedo, Johan T. den Dunnen, Stephanie O. M. Dyke, Xavier Estivill, Jack Goldblatt, Catherine Gonthier, Stephen C. Groft, Ivo Gut, Ada Hamosh, Philip Hieter, Sophie Höhn, Matthew E. Hurles, Petra Kaufmann, Bartha M. Knoppers, Jeffrey P. Krischer, Milan Macek, Gert Matthijs, Annie Olry, Samantha Parker, Justin Paschall, Anthony A. Philippakis, Heidi L. Rehm, Peter N. Robinson, Pak-Chung Sham, Rumen Stefanov, Domenica Taruscio, Divya Unni, Megan R. Vanstone, Feng Zhang, Han Brunner, Michael J. Bamshad, and Hanns Lochmüller. 2017. "International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases." *American Journal of Human Genetics* 100(5):695–705. doi: 10.1016/j.ajhg.2017.04.003.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. "A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff." *Fly* 6(2):80–92. doi: 10.4161/fly.19695.

Cummings, Beryl B., Jamie L. Marshall, Taru Tukiainen, Monkol Lek, Sandra Donkervoort, A. Reghan Foley, Veronique Bolduc, Leigh B. Waddell, Sarah A. Sandaradura, Gina L. O'Grady, Elicia Estrella, Hemakumar M. Reddy, Fengmei Zhao, Ben Weisburd, Konrad J. Karczewski, Anne H. O'Donnell-Luria, Daniel Birnbaum, Anna Sarkozy, Ying Hu, Hernan Gonorazky, Kristl Claeys, Himanshu Joshi, Adam Bournazos, Emily C. Oates, Roula Ghaoui, Mark R. Davis, Nigel G. Laing, Ana Topf, Genotype-Tissue Expression Consortium, Peter B. Kang, Alan H. Beggs, Kathryn N. North, Volker Straub, James J. Dowling, Francesco Muntoni, Nigel F. Clarke, Sandra T. Cooper, Carsten G. Bönnemann, and Daniel G. MacArthur. 2017. "Improving Genetic Diagnosis in Mendelian Disease with Transcriptome Sequencing." *Science Translational Medicine* 9(386). doi: 10.1126/scitranslmed.aal5209.

Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. 2010. "Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++." *PLOS Computational Biology* 6(12):e1001025. doi: 10.1371/journal.pcbi.1001025.

Dolzhenko, Egor, Viraj Deshpande, Felix Schlesinger, Peter Krusche, Roman Petrovski, Sai Chen, Dorothea Emig-Agius, Andrew Gross, Giuseppe Narzisi, Brett Bowman, Konrad Scheffler, Joke J. F. A. van Vugt, Courtney French, Alba Sanchis-Juan,

Kristina Ibáñez, Arianna Tucci, Bryan R. Lajoie, Jan H. Veldink, F. Lucy Raymond, Ryan J. Taft, David R. Bentley, and Michael A. Eberle. 2019. “ExpansionHunter: A Sequence-Graph-Based Tool to Analyze Variation in Short Tandem Repeat Regions.” *Bioinformatics* 35(22):4754–56. doi: 10.1093/bioinformatics/btz431.

Dunham, Ian, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum-Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Ian Dunham, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Jainab Khatun, Pouya Kheradpour, Anshul Kundaje, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephen C. J. Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Ewan Birney, Ian Dunham, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A. L. Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Eric D. Green, Peter J. Good, Elise A. Feingold, Bradley E. Bernstein, Ewan Birney, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Mark Gerstein, Morgan C. Giddings, Thomas R. Gingeras, Eric D. Green, Roderic Guigó, Ross C. Hardison, Timothy J. Hubbard, Manolis Kellis, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, Michael Snyder, John A. Stamatoyannopoulos, Scott A. Tenenbaum, Zhiping Weng, Kevin P. White, Barbara Wold, Jainab Khatun, Yanbao Yu, John Wrobel, Brian A. Risk, Harsha P. Gunawardena, Heather C. Kuiper, Christopher W. Maier, Ling Xie, Xian Chen, Morgan C. Giddings, Bradley E. Bernstein, Charles B. Epstein, Noam Shoresh, Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J. Coyne, Timothy Durham, Manching Ku, Thanh Truong, Lucas D. Ward, Robert C. Altshuler, Matthew L. Eaton, Manolis Kellis, Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha P. Gunawardena, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J. Luo, Eddie Park, Jonathan B. Preall, Kimberly Presaud, Paolo Ribeca, Brian A. Risk, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Timothy J. Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory J. Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, Thomas R. Gingeras, Kate R. Rosenbloom, Cricket A. Sloan, Katrina Learned, Venkat S. Malladi, Matthew C. Wong, Galt P. Barber,

Melissa S. Cline, Timothy R. Dreszer, Steven G. Heitner, Donna Karolchik, W. James Kent, Vanessa M. Kirkup, Laurence R. Meyer, Jeffrey C. Long, Morgan Maddren, Brian J. Raney, Terrence S. Furey, Lingyun Song, Linda L. Grasfeder, Paul G. Giresi, Bum-Kyu Lee, Anna Battenhouse, Nathan C. Sheffield, Jeremy M. Simon, Kimberly A. Showers, Alexias Safi, Darin London, Akshay A. Bhinge, Christopher Shestak, Matthew R. Schaner, Seul Ki Kim, Zhuzhu Z. Zhang, Piotr A. Mieczkowski, Joanna O. Mieczkowska, Zheng Liu, Ryan M. McDaniell, Yunyun Ni, Naim U. Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Ewan Birney, Vishwanath R. Iyer, Jason D. Lieb, Gregory E. Crawford, Guoliang Li, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Oscar J. Luo, Atif Shahab, Melissa J. Fullwood, Xiaoan Ruan, Yijun Ruan, Richard M. Myers, Florencia Pauli, Brian A. Williams, Jason Gertz, Georgi K. Marinov, Timothy E. Reddy, Jost Vielmetter, E. Partridge, Diane Trout, Katherine E. Varley, Clarke Gasper, The ENCODE Project Consortium, Overall coordination (data analysis coordination), Data production leads (data production), Lead analysts (data analysis), Writing group, NHGRI project management (scientific management), Principal investigators (steering committee), Boise State University and University of North Carolina at Chapel Hill Proteomics groups (data production and analysis), Broad Institute Group (data production and analysis), University of Geneva Cold Spring Harbor Center for Genomic Regulation, Barcelona, RIKEN, Sanger Institute, University of Lausanne, Genome Institute of Singapore group (data production and analysis), Data coordination center at UC Santa Cruz (production data coordination), EBI Duke University University of Texas, Austin, University of North Carolina-Chapel Hill group (data production and analysis), Genome Institute of Singapore group (data production and analysis), and Caltech HudsonAlpha Institute UC Irvine, Stanford group (data production and analysis). 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489(7414):57–74. doi: 10.1038/nature11247.

Ece Solmaz, Asli, Huseyin Onay, Tahir Atik, Ayca Aykut, Meltem Cerrah Gunes, Ozge Ozalp Yuregir, Veysel Nijat Bas, Filiz Hazan, Ozgur Kirbiyik, and Ferda Ozkinay. 2015. “Targeted Multi-Gene Panel Testing for the Diagnosis of Bardet Biedl Syndrome: Identification of Nine Novel Mutations across BBS1, BBS2, BBS4, BBS7, BBS9, BBS10 Genes.” *European Journal of Medical Genetics* 58(12):689–94. doi: 10.1016/j.ejmg.2015.10.011.

Edwards, A. W. F. 2008. “G. H. Hardy (1908) and Hardy–Weinberg Equilibrium.” *Genetics* 179(3):1143–50. doi: 10.1534/genetics.104.92940.

Eilbeck, Karen, Suzanna E. Lewis, Christopher J. Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. 2005. “The Sequence Ontology: A Tool for the Unification of Genome Annotations.” *Genome Biology* 6(5):R44. doi: 10.1186/gb-2005-6-5-r44.

Eilbeck, Karen, Aaron Quinlan, and Mark Yandell. 2017. “Settling the Score: Variant Prioritization and Mendelian Disease.” *Nature Reviews Genetics* 18(10):599–612. doi: 10.1038/nrg.2017.52.

English, Adam C., William J. Salerno, Oliver A. Hampton, Claudia Gonzaga-Jauregui, Shruthi Ambreth, Deborah I. Ritter, Christine R. Beck, Caleb F. Davis, Mahmoud Dahdouli, Singer Ma, Andrew Carroll, Narayanan Veeraraghavan, Jeremy Bruestle, Becky Drees, Alex Hastie, Ernest T. Lam, Simon White, Pamela Mishra, Min Wang, Yi Han, Feng Zhang, Pawel Stankiewicz, David A. Wheeler, Jeffrey G.

Reid, Donna M. Muzny, Jeffrey Rogers, Aniko Sabo, Kim C. Worley, James R. Lupski, Eric Boerwinkle, and Richard A. Gibbs. 2015. “Assessing Structural Variation in a Personal Genome—towards a Human Reference Diploid Genome.” *BMC Genomics* 16(1):286. doi: 10.1186/s12864-015-1479-3.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics (Oxford, England)* 32(19):3047–48. doi: 10.1093/bioinformatics/btw354.

Exome Aggregation Consortium, Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O’Donnell-Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, Taru Tukiainen, Daniel P. Birnbaum, Jack A. Kosmicki, Laramie E. Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N. Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I. Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M. Peloso, Ryan Poplin, Manuel A. Rivas, Valentin Ruano-Rubio, Samuel A. Rose, Douglas M. Ruderfer, Khalid Shakir, Peter D. Stenson, Christine Stevens, Brett P. Thomas, Grace Tiao, Maria T. Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M. Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C. Florez, Stacey B. Gabriel, Gad Getz, Stephen J. Glatt, Christina M. Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M. Neale, Aarno Palotie, Shaun M. Purcell, Danish Saleheen, Jeremiah M. Scharf, Pamela Sklar, Patrick F. Sullivan, Jaakko Tuomilehto, Ming T. Tsuang, Hugh C. Watkins, James G. Wilson, Mark J. Daly, and Daniel G. MacArthur. 2016. “Analysis of Protein-Coding Genetic Variation in 60,706 Humans.” *Nature* 536(7616):285–91. doi: 10.1038/nature19057.

Fiume, Marc, Miroslav Cupak, Stephen Keenan, Jordi Rambla, Sabela de la Torre, Stephanie O. M. Dyke, Anthony J. Brookes, Knox Carey, David Lloyd, Peter Goodhand, Maximilian Haeussler, Michael Baudis, Heinz Stockinger, Lena Dolman, Ilkka Lappalainen, Juha Törnroos, Mikael Linden, J. Dylan Spalding, Saif Ur-Rehman, Angela Page, Paul Flicek, Stephen Sherry, David Haussler, Susheel Varma, Gary Saunders, and Serena Scollen. 2019. “Federated Discovery and Sharing of Genomic Data Using Beacons.” *Nature Biotechnology* 37(3):220–24. doi: 10.1038/s41587-019-0046-x.

Grimm, Dominik G., Chloé-Agathe Azencott, Fabian Aicheler, Udo Gieraths, Daniel G. MacArthur, Kaitlin E. Samocha, David N. Cooper, Peter D. Stenson, Mark J. Daly, Jordan W. Smoller, Laramie E. Duncan, and Karsten M. Borgwardt. 2015. “The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity.” *Human Mutation* 36(5):513–23. doi: 10.1002/humu.22768.

Gurdasani, Deepti, Tommy Carstensen, Fasil Tekola-Ayele, Luca Pagani, Ioanna Tachmazidou, Konstantinos Hatzikotoulas, Savita Karthikeyan, Louise Iles, Martin O. Pollard, Ananyo Choudhury, Graham R. S. Ritchie, Yali Xue, Jennifer Asimit, Rebecca N. Nsubuga, Elizabeth H. Young, Cristina Pomilla, Katja Kivinen, Kirk Rockett, Anatoli Kamali, Ayo P. Doumatey, Gershim Asiki, Janet Seeley, Fatoumatta Sisay-Joof, Muminatou Jallow, Stephen Tollman, Ephrem Mekonnen,

Rosemary Ekong, Tamiru Oljira, Neil Bradman, Kalifa Bojang, Michele Ramsay, Adebowale Adeyemo, Endashaw Bekele, Ayesha Motala, Shane A. Norris, Fraser Pirie, Pontiano Kaleebu, Dominic Kwiatkowski, Chris Tyler-Smith, Charles Rotimi, Eleftheria Zeggini, and Manjinder S. Sandhu. 2015. "The African Genome Variation Project Shapes Medical Genetics in Africa." *Nature* 517(7534):327–32. doi: 10.1038/nature13997.

Harrison, Steven M., Erin R. Riggs, Donna R. Maglott, Jennifer M. Lee, Danielle R. Azzariti, Annie Niehaus, Erin M. Ramos, Christa L. Martin, Melissa J. Landrum, and Heidi L. Rehm. 2016. "Using ClinVar as a Resource to Support Variant Interpretation." *Current Protocols in Human Genetics* 89(1):8.16.1-8.16.23. doi: 10.1002/0471142905.hg0816s89.

Hartley, Taila, Tuğçe B. Balcı, Samantha K. Rojas, Alison Eaton, Care4Rare Canada, David A. Dyment, and Kym M. Boycott. 2018. "The Unsolved Rare Genetic Disease Atlas? An Analysis of the Unexplained Phenotypic Descriptions in OMIM®." *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* 178(4):458–63. doi: 10.1002/ajmg.c.31662.

Hunt, Ryan C., Vijaya L. Simhadri, Matthew Iandoli, Zuben E. Sauna, and Chava Kimchi-Sarfaty. 2014. "Exposing Synonymous Mutations." *Trends in Genetics* 30(7):308–21. doi: 10.1016/j.tig.2014.04.006.

Jain, Miten, Hugh E. Olsen, Benedict Paten, and Mark Akeson. 2016. "The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community." *Genome Biology* 17(1):239. doi: 10.1186/s13059-016-1103-0.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, Kristen M. Laricchia, Andrea Ganna, Daniel P. Birnbaum, Laura D. Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A. Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M. England, Eleanor G. Seaby, Jack A. Kosmicki, Raymond K. Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X. Chong, Kaitlin E. Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H. O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S. Ware, Christopher Vittal, Irina M. Armean, Louis Bergelson, Kristian Cibulskis, Kristen M. Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E. Talkowski, The Genome Aggregation Database Consortium, Benjamin M. Neale, Mark J. Daly, and Daniel G. MacArthur. 2019. "Variation across 141,456 Human Exomes and Genomes Reveals the Spectrum of Loss-of-Function Intolerance across Human Protein-Coding Genes." *BioRxiv* 531210. doi: 10.1101/531210.

Kircher, Martin, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." *Nature Genetics* 46(3):310–15. doi: 10.1038/ng.2892.

Knudson, A. G. 1971. “Mutation and Cancer: Statistical Study of Retinoblastoma.” *Proceedings of the National Academy of Sciences of the United States of America* 68(4):820–23. doi: 10.1073/pnas.68.4.820.

Kobayashi, Yuya, Shan Yang, Keith Nykamp, John Garcia, Stephen E. Lincoln, and Scott E. Topper. 2017. “Pathogenic Variant Burden in the ExAC Database: An Empirical Approach to Evaluating Population Data for Clinical Variant Interpretation.” *Genome Medicine* 9(1):13. doi: 10.1186/s13073-017-0403-7.

Köhler, Sebastian, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, David Lewis-Smith, Nicole A. Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M. Brower, Tiffany J. Callahan, Christopher G. Chute, Johanna L. Est, Peter D. Galer, Shiva Ganesan, Matthias Griese, Matthias Haimel, Julia Pazmandi, Marc Hanauer, Nomi L. Harris, Michael J. Hartnett, Maximilian Hastreiter, Fabian Hauck, Yongqun He, Tim Jeske, Hugh Kearney, Gerhard Kindle, Christoph Klein, Katrin Knoflach, Roland Krause, David Lagorce, Julie A. McMurry, Jillian A. Miller, Monica C. Munoz-Torres, Rebecca L. Peters, Christina K. Rapp, Ana M. Rath, Shahmir A. Rind, Avi Z. Rosenberg, Michael M. Segal, Markus G. Seidel, Damian Smedley, Tomer Talmy, Yarlalu Thomas, Samuel A. Wiafe, Julie Xian, Zafer Yüksel, Ingo Helbig, Christopher J. Mungall, Melissa A. Haendel, and Peter N. Robinson. 2021. “The Human Phenotype Ontology in 2021.” *Nucleic Acids Research* 49(D1):D1207–17. doi: 10.1093/nar/gkaa1043.

Köhler, Sebastian, Marcel H. Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E. Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N. Robinson. 2009. “Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies.” *American Journal of Human Genetics* 85(4):457–64. doi: 10.1016/j.ajhg.2009.09.003.

Kundaje, Anshul, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael Stevens, Robert E. Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, and Manolis Kellis. 2015. “Integrative Analysis of 111 Reference Human Epigenomes.” *Nature* 518(7539):317–30. doi: 10.1038/nature14248.

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, Wonhee Jang, Kenneth Katz, Michael Ovetsky, George Riley, Amanjeev Sethi, Ray Tully, Ricardo Villamarin-Salomon, Wendy Rubinstein, and Donna R. Maglott. 2016. "ClinVar: Public Archive of Interpretations of Clinically Relevant Variants." *Nucleic Acids Research* 44(D1):D862–68. doi: 10.1093/nar/gkv1222.

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J. Bradley Holmes, Brandi L. Kattman, and Donna R. Maglott. 2018. "ClinVar: Improving Access to Variant Interpretations and Supporting Evidence." *Nucleic Acids Research* 46(D1):D1062–67. doi: 10.1093/nar/gkx1153.

Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 25(14):1754–60. doi: 10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25(16):2078–79. doi: 10.1093/bioinformatics/btp352.

Li, Wentian, and Jan Freudenberg. 2014. "Mappability and Read Length." *Frontiers in Genetics* 5:381. doi: 10.3389/fgene.2014.00381.

MacArthur, Daniel G., Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K. Pickrell, Stephen B. Montgomery, Cornelis A. Albers, Zhengdong D. Zhang, Donald F. Conrad, Gerton Lunter, Hancheng Zheng, Qasim Ayub, Mark A. DePristo, Eric Banks, Min Hu, Robert E. Handsaker, Jeffrey A. Rosenfeld, Menachem Fromer, Mike Jin, Xinmeng Jasmine Mu, Ekta Khurana, Kai Ye, Mike Kay, Gary Ian Saunders, Marie-Marthe Suner, Toby Hunt, If H. A. Barnes, Clara Amid, Denise R. Carvalho-Silva, Alexandra H. Bignell, Catherine Snow, Bryndis Yngvadottir, Suzannah Bumpstead, David N. Cooper, Yali Xue, Irene Gallego Romero, 1000 Genomes Project Consortium, Jun Wang, Yingrui Li, Richard A. Gibbs, Steven A. McCarroll, Emmanouil T. Dermitzakis, Jonathan K. Pritchard, Jeffrey C. Barrett, Jennifer Harrow, Matthew E. Hurles, Mark B. Gerstein, and Chris Tyler-Smith. 2012. "A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes." *Science* 335(6070):823–28. doi: 10.1126/science.1215040.

Magnusson, Måns. [2014] 2018. *genmod* [computer program]. Version 3.7.3. https://github.com/moonso/genmod.

Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. 2019. "Structural Variant Calling: The Long and the Short of It." *Genome Biology* 20(1):246. doi: 10.1186/s13059-019-1828-7.

Martin, Antonio Rueda, Eleanor Williams, Rebecca E. Foulger, Sarah Leigh, Louise C. Daugherty, Olivia Niblock, Ivone U. S. Leong, Katherine R. Smith, Oleg Gerasimenko, Eik Haraldsdottir, Ellen Thomas, Richard H. Scott, Emma Baple,

Arianna Tucci, Helen Brittain, Anna de Burca, Kristina Ibañez, Dalia Kasperaviciute, Damian Smedley, Mark Caulfield, Augusto Rendon, and Ellen M. McDonagh. 2019. “PanelApp Crowdsources Expert Knowledge to Establish Consensus Diagnostic Gene Panels.” *Nature Genetics* 51(11):1560–65. doi: 10.1038/s41588-019-0528-2.

McCarthy, Davis J., Peter Humburg, Alexander Kanapin, Manuel A. Rivas, Kyle Gaulton, Jean-Baptiste Cazier, Peter Donnelly, and The WGS500 Consortium. 2014. “Choice of Transcripts and Software Has a Large Effect on Variant Annotation.” *Genome Medicine* 6(3):26. doi: 10.1186/gm543.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. “The Ensembl Variant Effect Predictor.” *Genome Biology* 17(1):122. doi: 10.1186/s13059-016-0974-4.

Mendel, Gregor. 1866. *Versuche Über Pflanzen-Hybriden*. Brünn : Im Verlage des Vereines,.

Might, Matthew, and Matt Wilsey. 2014. “The Shifting Model in Clinical Diagnostics: How next-Generation Sequencing and Families Are Altering the Way Rare Diseases Are Discovered, Studied, and Treated.” *Genetics in Medicine* 16(10):736–37. doi: 10.1038/gim.2014.23.

Miller, Chase A., Yi Qiao, Tonya DiSera, Brian D’Astous, and Gabor T. Marth. 2014. “Bam.Iobio: A Web-Based, Real-Time, Sequence Alignment File Inspector.” *Nature Methods* 11(12):1189–1189. doi: 10.1038/nmeth.3174.

Nagasaki, Masao, Jun Yasuda, Fumiki Katsuoka, Naoki Nariai, Kaname Kojima, Yosuke Kawai, Yumi Yamaguchi-Kabata, Junji Yokozawa, Inaho Danjoh, Sakae Saito, Yukuto Sato, Takahiro Mimori, Kaoru Tsuda, Rumiko Saito, Xiaoqing Pan, Satoshi Nishikawa, Shin Ito, Yoko Kuroki, Osamu Tanabe, Nobuo Fuse, Shinichi Kuriyama, Hideyasu Kiyomoto, Atsushi Hozawa, Naoko Minegishi, James Douglas Engel, Kengo Kinoshita, Shigeo Kure, Nobuo Yaegashi, and Masayuki Yamamoto. 2015. “Rare Variant Discovery by Deep Whole-Genome Sequencing of 1,070 Japanese Individuals.” *Nature Communications* 6(1):1–13. doi: 10.1038/ncomms9018.

Narasimhan, Vagheesh M., Karen A. Hunt, Dan Mason, Christopher L. Baker, Konrad J. Karczewski, Michael R. Barnes, Anthony H. Barnett, Chris Bates, Srikanth Bellary, Nicholas A. Bockett, Kristina Giorda, Christopher J. Griffiths, Harry Hemingway, Zhilong Jia, M. Ann Kelly, Hajrah A. Khawaja, Monkol Lek, Shane McCarthy, Rosie McEachan, Anne O’Donnell-Luria, Kenneth Paigen, Constantinos A. Parisinos, Eamonn Sheridan, Laura Southgate, Louise Tee, Mark Thomas, Yali Xue, Michael Schnall-Levin, Petko M. Petkov, Chris Tyler-Smith, Eamonn R. Maher, Richard C. Trembath, Daniel G. MacArthur, John Wright, Richard Durbin, and David A. van Heel. 2016. “Health and Population Effects of Rare Gene Knockouts in Adult Humans with Related Parents.” *Science* 352(6284):474–77. doi: 10.1126/science.aac8624.

Ng, Pauline C., and Steven Henikoff. 2003. “SIFT: Predicting Amino Acid Changes That Affect Protein Function.” *Nucleic Acids Research* 31(13):3812–14. doi: 10.1093/nar/gkg509.

Ng, Sarah B., Emily H. Turner, Peggy D. Robertson, Steven D. Flygare, Abigail W. Bigham, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, Evan E. Eichler, Michael Bamshad, Deborah A. Nickerson, and Jay Shendure. 2009. "Targeted Capture and Massively Parallel Sequencing of 12 Human Exomes." *Nature* 461(7261):272–76. doi: 10.1038/nature08250.

Okonechnikov, Konstantin, Ana Conesa, and Fernando García-Alcalde. 2016. "Qualimap 2: Advanced Multi-Sample Quality Control for High-Throughput Sequencing Data." *Bioinformatics* 32(2):292–94. doi: 10.1093/bioinformatics/btv566.

Philippakis, Anthony A., Danielle R. Azzariti, Sergi Beltran, Anthony J. Brookes, Catherine A. Brownstein, Michael Brudno, Han G. Brunner, Orion J. Buske, Knox Carey, Cassie Doll, Sergiu Dumitriu, Stephanie O. M. Dyke, Johan T. den Dunnen, Helen V. Firth, Richard A. Gibbs, Marta Girdea, Michael Gonzalez, Melissa A. Haendel, Ada Hamosh, Ingrid A. Holm, Lijia Huang, Matthew E. Hurles, Ben Hutton, Joel B. Krier, Andriy Misyura, Christopher J. Mungall, Justin Paschall, Benedict Paten, Peter N. Robinson, François Schiettecatte, Nara L. Sobreira, Ganesh J. Swaminathan, Peter E. Taschner, Sharon F. Terry, Nicole L. Washington, Stephan Züchner, Kym M. Boycott, and Heidi L. Rehm. 2015. "The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery." *Human Mutation* 36(10):915–21. doi: 10.1002/humu.22858.

Richards, Sue, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L. Rehm. 2015. "Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology." *Genetics in Medicine* 17(5):405–23. doi: 10.1038/gim.2015.30.

Sayres, Melissa A. Wilson, Kirk E. Lohmueller, and Rasmus Nielsen. 2014. "Natural Selection Reduced Diversity on Human Y Chromosomes." *PLOS Genetics* 10(1):e1004064. doi: 10.1371/journal.pgen.1004064.

Schaffner, Stephen F. 2004. "The X Chromosome in Population Genetics." *Nature Reviews Genetics* 5(1):43–51. doi: 10.1038/nrg1247.

Shendure, Jay, Gregory J. Porreca, Nikos B. Reppas, Xiaoxia Lin, John P. McCutcheon, Abraham M. Rosenbaum, Michael D. Wang, Kun Zhang, Robi D. Mitra, and George M. Church. 2005. "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome." *Science*.

Slatkin, Montgomery. 2008. "Linkage Disequilibrium — Understanding the Evolutionary Past and Mapping the Medical Future." *Nature Reviews Genetics* 9(6):477–85. doi: 10.1038/nrg2361.

Sohn, Jang-il, and Jin-Wu Nam. 2018. "The Present and Future of de Novo Whole-Genome Assembly." *Briefings in Bioinformatics* 19(1):23–40. doi: 10.1093/bib/bbw096.

Stankiewicz, Paweł, and James R. Lupski. 2010. "Structural Variation in the Human Genome and Its Role in Disease." *Annual Review of Medicine* 61(1):437–55. doi: 10.1146/annurev-med-100708-204735.

Stenson, Peter D., Matthew Mort, Edward V. Ball, Katy Evans, Matthew Hayden, Sally Heywood, Michelle Hussain, Andrew D. Phillips, and David N. Cooper. 2017. "The

Human Gene Mutation Database: Towards a Comprehensive Repository of Inherited Mutation Data for Medical Research, Genetic Diagnosis and next-Generation Sequencing Studies." *Human Genetics* 136(6):665–77. doi: 10.1007/s00439-017-1779-6.

Stranneheim, Henrik, Martin Engvall, Karin Naess, Nicole Lesko, Pontus Larsson, Mats Dahlberg, Robin Andeer, Anna Wredenberg, Chris Freyer, Michela Barbaro, Helene Bruhn, Tesfail Emahazion, Måns Magnusson, Rolf Wibom, Rolf H. Zetterström, Valtteri Wirta, Ulrika von Döbeln, and Anna Wedell. 2014. "Rapid Pulsed Whole Genome Sequencing for Comprehensive Acute Diagnostics of Inborn Errors of Metabolism." *BMC Genomics* 15(1):1090. doi: 10.1186/1471-2164-15-1090.

Tan, Adrian, Gonçalo R. Abecasis, and Hyun Min Kang. 2015. "Unified Representation of Genetic Variants." *Bioinformatics* 31(13):2202–4. doi: 10.1093/bioinformatics/btv112.

Tattini, Lorenzo, Romina D'Aurizio, and Alberto Magi. 2015. "Detection of Genomic Structural Variants from Next-Generation Sequencing Data." *Frontiers in Bioengineering and Biotechnology* 3. doi: 10.3389/fbioe.2015.00092.

Telenti, Amalio, Levi C. T. Pierce, William H. Biggs, Julia di Iulio, Emily H. M. Wong, Martin M. Fabani, Ewen F. Kirkness, Ahmed Moustafa, Naisha Shah, Chao Xie, Suzanne C. Brewerton, Nadeem Bulsara, Chad Garner, Gary Metzker, Efren Sandoval, Brad A. Perkins, Franz J. Och, Yaron Turpaz, and J. Craig Venter. 2016. "Deep Sequencing of 10,000 Human Genomes." *Proceedings of the National Academy of Sciences* 113(42):11901–6. doi: 10.1073/pnas.1613365113.

Tennessen, Jacob A., Abigail W. Bigham, Timothy D. O'Connor, Wenqing Fu, Eimear E. Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, Hyun Min Kang, Daniel Jordan, Suzanne M. Leal, Stacey Gabriel, Mark J. Rieder, Goncalo Abecasis, David Altshuler, Deborah A. Nickerson, Eric Boerwinkle, Shamil Sunyaev, Carlos D. Bustamante, Michael J. Bamshad, Joshua M. Akey, Broad Go, Seattle Go, and on behalf of the NHLBI Exome Sequencing Project. 2012. "Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes." *Science* 337(6090):64–69. doi: 10.1126/science.1219240.

Tiepolo, L., and O. Zuffardi. 1976. "Localization of Factors Controlling Spermatogenesis in the Nonfluorescent Portion of the Human Y Chromosome Long Arm." *Human Genetics* 34(2):119–24. doi: 10.1007/BF00278879.

Warr, Amanda, Christelle Robert, David Hume, Alan Archibald, Nader Deeb, and Mick Watson. 2015. "Exome Sequencing: Current and Future Perspectives." *G3 Genes|Genomes|Genetics* 5(8):1543–50. doi: 10.1534/g3.115.018564.

Wenger, Aaron M., Harendra Guturu, Jonathan A. Bernstein, and Gill Bejerano. 2017. "Systematic Reanalysis of Clinical Exome Data Yields Additional Diagnoses: Implications for Providers." *Genetics in Medicine* 19(2):209–14. doi: 10.1038/gim.2016.88.

Westenfield, Kristen, Kyriakie Sarafoglou, Laura C. Speltz, Elizabeth I. Pierpont, Joan Steyermark, David Nascene, Matthew Bower, and Mary Ella Pierpont. 2018. "Mosaicism of the UDP-Galactose Transporter SLC35A2 in a Female Causing a

Congenital Disorder of Glycosylation: A Case Report." *BMC Medical Genetics* 19:100. doi: 10.1186/s12881-018-0617-6.

Whiffin, Nicola, Eric Minikel, Roddy Walsh, Anne H. O'Donnell-Luria, Konrad Karczewski, Alexander Y. Ing, Paul J. R. Barton, Birgit Funke, Stuart A. Cook, Daniel MacArthur, and James S. Ware. 2017. "Using High-Resolution Variant Frequencies to Empower Clinical Genome Interpretation." *Genetics in Medicine* 19(10):1151–58. doi: 10.1038/gim.2017.26.

Willig, Laurel K., Josh E. Petrikin, Laurie D. Smith, Carol J. Saunders, Isabelle Thiffault, Neil A. Miller, Sarah E. Soden, Julie A. Cakici, Suzanne M. Herd, Greyson Twist, Aaron Noll, Mitchell Creed, Patria M. Alba, Shannon L. Carpenter, Mark A. Clements, Ryan T. Fischer, J. Allyson Hays, Howard Kilbride, Ryan J. McDonough, Jamie L. Rosterman, Sarah L. Tsai, Lee Zellmer, Emily G. Farrow, and Stephen F. Kingsmore. 2015. "Whole-Genome Sequencing for Identification of Mendelian Disorders in Critically Ill Infants: A Retrospective Analysis of Diagnostic and Clinical Findings." *The Lancet Respiratory Medicine* 3(5):377–87. doi: 10.1016/S2213-2600(15)00139-3.

Wright, Caroline F., Tomas W. Fitzgerald, Wendy D. Jones, Stephen Clayton, Jeremy F. McRae, Margriet van Kogelenberg, Daniel A. King, Kirsty Ambridge, Daniel M. Barrett, Tanya Bayzetinova, A. Paul Bevan, Eugene Bragin, Eleni A. Chatzimichali, Susan Gribble, Philip Jones, Netravathi Krishnappa, Laura E. Mason, Ray Miller, Katherine I. Morley, Vijaya Parthiban, Elena Prigmore, Diana Rajan, Alejandro Sifrim, G. Jawahar Swaminathan, Adrian R. Tivey, Anna Middleton, Michael Parker, Nigel P. Carter, Jeffrey C. Barrett, Matthew E. Hurles, David R. FitzPatrick, and Helen V. Firth. 2015. "Genetic Diagnosis of Developmental Disorders in the DDD Study: A Scalable Analysis of Genome-Wide Research Data." *The Lancet* 385(9975):1305–14. doi: 10.1016/S0140-6736(14)61705-0.

Wright, Caroline F., David R. FitzPatrick, and Helen V. Firth. 2018. "Paediatric Genomics: Diagnosing Rare Disease in Children." *Nature Reviews Genetics* 19(5):253–68. doi: 10.1038/nrg.2017.116.