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

GENOMIC SCREENING AND CAUSES OF RARE DISORDERS

Malin Kvarnung

Stockholm 2016

Cover: Alexander Calder, The Forest is the Best Place, 1945. © 2016 Calder Foundation, New York / BUS, Stockholm.

Published by Karolinska Institutet.
Printed by E-Print AB.
© Malin Kvarnung, 2016
ISBN 978-91-7676-151-9

Genomic screening and causes of rare disorders
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Malin Kvarnung

Principal Supervisor:
Professor Elisabeth Syk Lundberg
Karolinska Institutet
Department of Molecular Medicine and Surgery

Co-supervisor(s):
Professor Magnus Nordenskjöld
Karolinska Institutet
Department of Molecular Medicine and Surgery

Professor Ann Nordgren
Karolinska Institutet
Department of Molecular Medicine and Surgery

Dr Daniel Nilsson
Karolinska Institutet
Department of Molecular Medicine and Surgery

Dr Agne Liéden
Karolinska Institutet
Department of Molecular Medicine and Surgery

Opponent:
Dr Helen Firth
Cambridge University
Department of Medical Genetics

Examination Board:
Associate professor Cecilia Gunnarsson
Linköping University
Department of Clinical and Experimental Medicine

Associate professor Lars Feuk
Uppsala University
Department of Immunology, Genetics and Pathology

Associate professor Kristina Tedroff
Karolinska Institutet
Department of Women's and Children's Health

"Nature is nowhere accustomed more openly to display her secret mysteries than in cases where she shows tracings of her workings apart from the beaten paths; nor is there any better way to advance the proper practice of medicine than to give our minds to the discovery of the usual law of nature, by careful investigation of cases of rarer forms of disease."

Dr William Harvey, 1657

ABSTRACT

Congenital disorders affect approximately 3-4% of all children and often cause chronic disabilities with significant impact on the lives of affected individuals and their families as well as on the health-care system. These disorders constitute a large and heterogeneous group of disorders with most of them being rare (prevalence <1/2000) and having an underlying genetic basis. Understanding of the molecular etiology and phenotypic spectrum has expanded during recent years. Over the past ten years, it has been shown that different types of causative genetic variants, such as single nucleotide variants, small indels or copy number variants, can be detected in many patients with congenital disorders. However, much remain to be explored concerning the spectrum of genetic variants and phenotypes associated to these disorders.

The studies in the thesis have focused on determining the molecular etiology of rare congenital disorders and delineating the phenotypes associated with these disorders.

In order to achieve this, phenotypic investigations combined with genetic screening through clinical array-CGH and whole exome sequencing, followed by a strategy for evaluation, were performed in selected families. Twenty families with parental kinship and children affected by presumed autosomal recessive disorders and one additional family with a *de novo* dominant disorder were included in the studies. By this approach, a molecular diagnosis could be determined in 15 out of 21 families. With the results from the studies, the gene *PIGT* was established as a novel disease gene, the genes *TFG* and *KIAA1109* were confirmed as novel disease genes and additional candidate genes for congenital disorders were identified. Furthermore, the phenotypes for disorders associated with the genes *MAN1B1, RIPK4* and *FLVCR2* were expended and the spectrum of pathogenic variants in the gene *SATB2* was broadened.

The overall conclusions from the studies were that WES is a very powerful method for the identification of disease-causing variants in consanguineous families and that the diversity of AR diseases is enormous with many of the identified disorders being extremely rare. An additional conclusion is that a detailed phenotypic assessment is crucial for interpretation of data from large-scale genetic screening and for ascribing pathogenicity to the identified variants. Moreover, the full spectrum of genetic variants, including sequence alterations and CNVs, should be considered for the etiology of rare disorders.

The results altogether add detail to the clinical presentations of the given disorders and expand the number of genes and genetic variants with a presumed or established causal association to congenital disorders. Ultimately, this may increase the chances to achieve a genetic diagnosis for future patients.

LIST OF SCIENTIFIC PAPERS

I. Malin Kvarnung, Daniel Nilsson, Anna Lindstrand, Christoph Korenke, Samuel Chiang, Elisabeth Blennow, Markus Bergmann, Tommy Stödberg, Outi Mäkitie, Britt-Marie Anderlid, Yenan T. Bryceson, Magnus Nordenskjöld, Ann Nordgren
A Novel Intellectual Disability Syndrome Caused by GPI-anchor Deficiency due to Homozygous Mutations in PIGT
Journal of Medical Genetics. 2013 Aug;50(8):521-8

II. Agne Liedén, Malin Kvarnung, Daniel Nilsson, Ellika Sahlin, Elisabeth Syk Lundberg
Intragenic Duplication - A Novel Causative Mechanism for SATB2-associated Syndrome
Am J Med Genet A. 2014 Dec;164A(12):3083-7

III. Malin Kvarnung, Fulya Taylan, Daniel Nilsson, Margareta Albåge, Magnus Nordenskjöld, Britt-Marie Anderlid, Ann Nordgren, Elisabeth Syk Lundberg
Mutations in FLVCR2 associated with Fowler syndrome and survival beyond infancy
Clinical Genetics. 2015 Feb 10, Epub ahead of print

IV. Malin Kvarnung, Fulya Taylan, Daniel Nilsson, Helena Malmgren, Kristina Lagerstedt-Robinson, Eva Holmberg, Anders Helander, Alejandra Cuevas Cid, Kerstin Sars Zimmer, Suzanne Marcus, Britt-Marie Anderlid, Magnus Nordenskjöld, Ann Nordgren, Elisabeth Syk Lundberg
Whole exome sequencing in consanguineous families with rare disorders: High diagnostic yield and new disease gene identification
Manuscript

CONTENTS

LIST OF ABBREVIATIONS

AD	Autosomal Dominant
AR	Autosomal Recessive
bp	basepair
CADD	Combined Annotation Dependent Depletion
CGH	Comparative Genomic Hybridization
CNV	Copy Number Variant
dbSNP	database of Single Nucleotide Polymorphism (until 2011) database of Short Genetic Variation (from 2011)
DDD	Deciphering Developmental Disorders
DECIPHER	DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources
DGV	Database of Genomic Variants
DNA	DeoxyriboNucleic Acid
ds	double stranded
ExAc	Exome Aggregation Consortium
FADS	Fetal Akinesia Deformation Sequence
FoSTeS	Fork-Stalling and Template Switching
GPI	Glycosyl-Phosphatidyl-Inositol
HGMD	Human Gene Mutation Database
HPO	Human Phenotype Ontology
HR	Homologous Recombination
HSP	Hereditary Spastic Paraplegia
ID	Intellectual Disability
indel	insertion and/or deletion
kb	kilobase (1000 basepairs)
LCR	Low Copy Repeat
MIM	Mendelian Inheritance in Man
MLPA	Multiplex Ligation-dependent Probe Amplification
MMBIR	Microhomology-Mediated Break Induced Replication
MMEJ	Microhomology-Mediated End Joining

MPS	Massive Parallel Sequencing
MRI	Magnetic Resonance Imaging
NAHR	Non-Allelic Homologous Recombination
NGS	Next Generation Sequencing
NHEJ	Non-Homologous End Joining
OFC	Occipito-Frontal Circumference
OMIM	Online Mendelian Inheritance in Man
PCR	Polymerase Chain Reaction
PGD	Preimplantation Genetic Diagnosis
RNA	RiboNucleic Acid
SNV	Single Nucleotide Variant
Ss	single stranded
SV	Structural Variant
UPD	UniParental Disomy
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
XL	X-linked

1 INTRODUCTION

Congenital or early onset disorders affect approximately 3-4% of all children and often cause chronic disabilities with significant impact on the lives of affected individuals and their families as well as on the health-care system. These disorders constitute a large and heterogeneous group of disorders with most of them being rare. Our understanding of the etiology and phenotypic spectrum of these disorders has expanded dramatically during recent years. Through genomic screening methods, it has been shown that the etiology is genetic in the majority of the patients and that the phenotypic spectrum often is larger than previously believed.

Understanding these disorders is important and has direct clinical implications as it increases the chances of achieving an etiological diagnosis for patients with rare disorders, which in turn enables accurate genetic counseling and prenatal diagnostics as well as counseling regarding prognosis and, in some cases, treatment options. Furthermore, knowledge about the molecular etiology may open up for studies regarding treatment for some of these disorders.

Despite recent years' advances in the field, much remain to be explored concerning the full spectrum of genetic variants and phenotypes of these disorders.

1.1 RARE DISORDERS

1.1.1 Definitions

The term "rare disorders" is widely used for disorders or diseases that affect few people – as opposed to the more common disorders or diseases, like diabetes, depression or cardiovascular disease, that affect many people. There are currently two definitions or cut-off levels regarding what should be considered as rare in this context;

- In Europe, a disease or disorder is defined as rare when it affects fewer than 1 in 2000.[1]
- In the USA, a disease or disorder is defined as rare when it affects fewer than 200,000 Americans at any given time.[2] Considering a population of 319 million people in the USA, this definition can be translated into a disease or disorder that affects fewer than approximately 1 in 1600.

The terms disorder and disease will be regarded as synonyms and used interchangeably in the text.

1.1.2 Prevalence

Despite the rarity of these disorders, the total number of people affected is large. It is estimated that a rare disease affects one person out of 15 and half of them are children.[3, 4] These figures are clearly in line with the commonly given incidence figure of 3-4% for congenital or early onset disorders, which indicate that most of these disorders are rare.

The high total prevalence is explained by the large number of rare disorders, which equals nearly 10,000.[5, 6] Each individual disorder is rare, but when considered as a group, rare disorders are common.

The prevalence distribution within the group of rare disorders is skewed. A few of these disorders are relatively common with prevalence above 1/20,000, while the vast majority of the disorders are very rare (Figure 1).[7] It has been estimated that 80% of all rare disease patients are affected by approximately 350 rare diseases[8], while the rest of the patients are affected by a plethora of very rare disorders. At the extreme end, there are disorders that have been described only in one or a few patients or families.

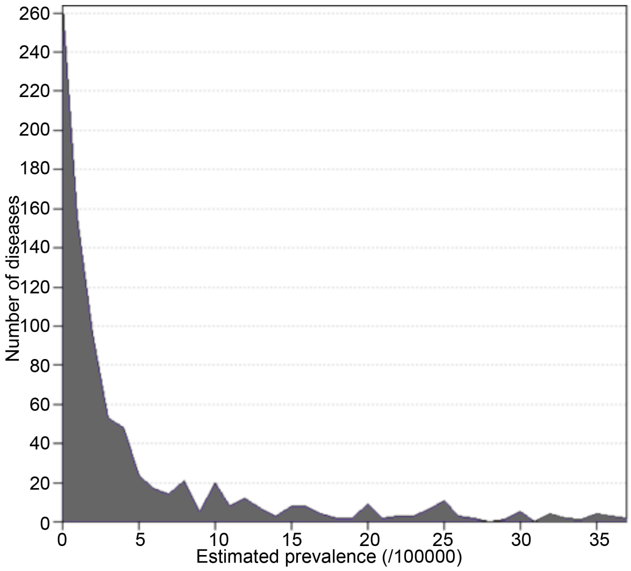

Figure 1. Prevalence of different rare diseases

Based on data from Orphanet report series, Rare diseases collection, July 2015[7] with permission from the publisher.

For some disorders, the prevalence rates are highly variable between different populations and geographical regions. These differences can be due to factors that are either genetic or environmental. Examples on the latter are rare infections or exposure to harmful substances that are more prevalent in certain regions. Regarding genetic factors, there is a variation between populations in the prevalence and inheritance of certain genetic variants. The mechanisms behind this variation are several. One mechanism is limitation of the population size where individuals marry and reproduce within the community, either due to geographical boundaries or because of traditions that encourage marriage within the ethnic group, social class, religious group etc. A restriction in population size allows for the enrichment of potentially harmful genetic variants due to a founder effect that in turn increases the risk for disease. This may be exemplified by certain disorders, such as Salla disease that is more common in regions of Finland[9] and Ellis van-Creveld syndrome that is more common in parts of the Amish population[10]. Another mechanism is natural selection for specific, potentially disease-causing, variants in certain geographical regions. This is well known for the group of disorders called hemoglobinopathies, which include sickle-cell anemia and thalassemia. These disorders are historically rare in the European Union and the USA, while prevalence in endemic malaria regions reaches >1/100. Heterozygous carriers of these autosomal recessive disorders are protected against severe malaria and therefore the carrier status is favored in the population, with the drawback of an increased prevalence of hemoglobinopathies.[11] A third mechanism that leads to variable disease prevalence is a difference in the rate of consanguineous marriages between different populations and regions. This will be discussed in more detail in chapter 1.3.

1.1.3 Etiology

Most of the rare disorders have a genetic basis, while others have non-genetic causes such as infections, auto-immunity and environmental factors. For a proportion of the disorders, the etiology is still unknown.[4]

During the course over the last 25 years there has been enormous advances in deciphering the etiology of rare genetic disorders, which is reflected in the increasing number of known disease genes and disease-causing chromosomal aberrations as well as in the number of diseases or disorders with a known molecular cause.[12-14] These data are recorded in the catalogue “Mendelian Inheritance in Man” (MIM), available online as “Online Mendelian Inheritance in Man” (OMIM), which lists more than 8000 phenotypes or diseases with a presumed genetic cause. Since 1990, the molecular etiology of more than 4500 of these disorders has been identified and the number of known disease genes is 3075 as of December 1, 2015 (Figure 2A).[5, 15] Despite the enormous progress in recent years, the basis is still unknown for nearly half of the diseases. As seen in figure 2A, the number of disorders with a known etiology is larger than the number of disease genes, which indicates that variants in the same gene can cause several different disorders. An example of this is the *ERCC5* gene, which is associated to three different disorders; xeroderma pigmentosum, Cockayne syndrome and cerebrooculofacioskeletal syndrome.[16, 17] The other way around, the same phenotype may be caused by variants in different genes, which is the case in intellectual disability for example, where an extreme heterogeneity is seen.[18, 19] (In OMIM, these are designated as separate disorders coupled to the causative gene or chromosomal aberration.)

For disorders that have a known molecular cause, the inheritance pattern is autosomal recessive in about half of the cases, autosomal dominant in 43% and X-linked in 6% (Figure 2B).[5]

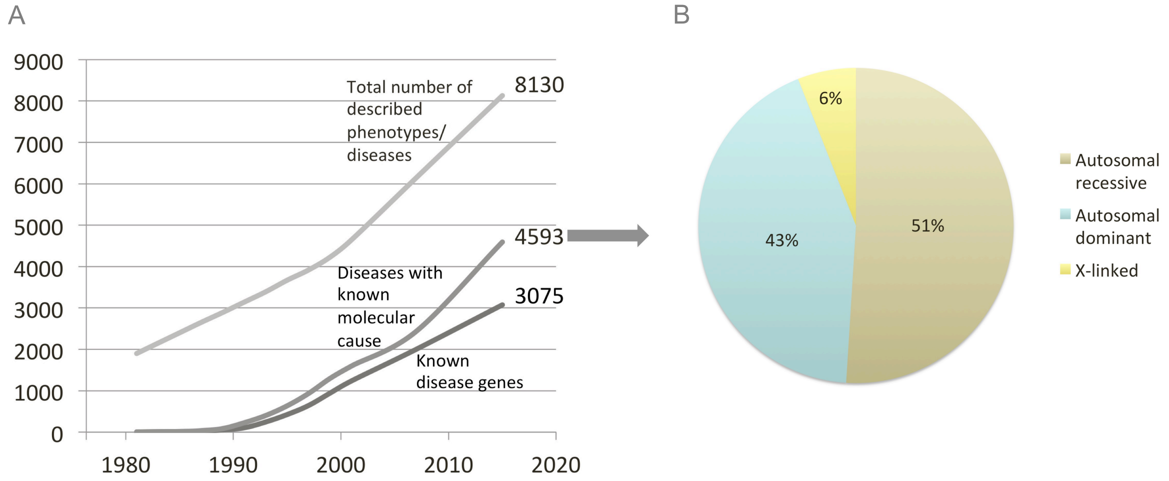

Figure 2. Number of entries in MIM/OMIM over time and inheritance of genetic diseases
A) Diagram showing the cumulative number of entries into MIM/OMIM regarding known disease genes, genetic diseases with a known molecular cause and total number of described diseases (with a presumed genetic etiology), over the last 30 years. B) Pie chart showing the inheritance patterns for diseases with a known molecular cause. Based on data from Antonarakis et al 2000[12]*, Peltonen et al 2001*[13]*, McKusick 2007*[14]*, Amberger et al 2015*[15] *and OMIM*[5]*.*

Reviewing etiology from a patient perspective, there are no comprehensive studies on the detailed etiology in cohorts of unselected rare disorder patients. However, there are several studies on different subgroups, for example patients with intellectual disability and developmental disorders. This group is of particular interest considering the high prevalence of these symptoms among rare disease patients in general[5] and also among the patients that have been studied in the thesis (see Clinical Presentation below). Recent studies indicate that up to 40% of ID patients, are affected by specific monogenic disorders. Most of these are autosomal dominant, while some are X-linked (5-10%) or autosomal recessive (2-4%). Another 20% of the patients are affected by disorders caused by deletions or duplications that span >500 bp of the genome, so called copy number variants (CNVs). In addition, 11% of the patients have larger chromosomal aberrations, including aneuploidies. The studies also show that for the vast majority of all patients with a genetic cause, the genetic variant is not inherited, but instead *de novo* in origin. (The few sporadic patients with inherited variants are those with AR disorders and approximately half of those with X-linked disorders.) The remainder of all patients, approximately 30-40%, suffer from disorders that are still of unknown etiology or due to non-genetic factors.[18-22] These figures contrast to what was known on the etiology of intellectual disability ten to fifteen years ago when 80% of the patients were considered to be affected by a disorder of unknown origin or due to non-genetic factors (Figure 3).[23]

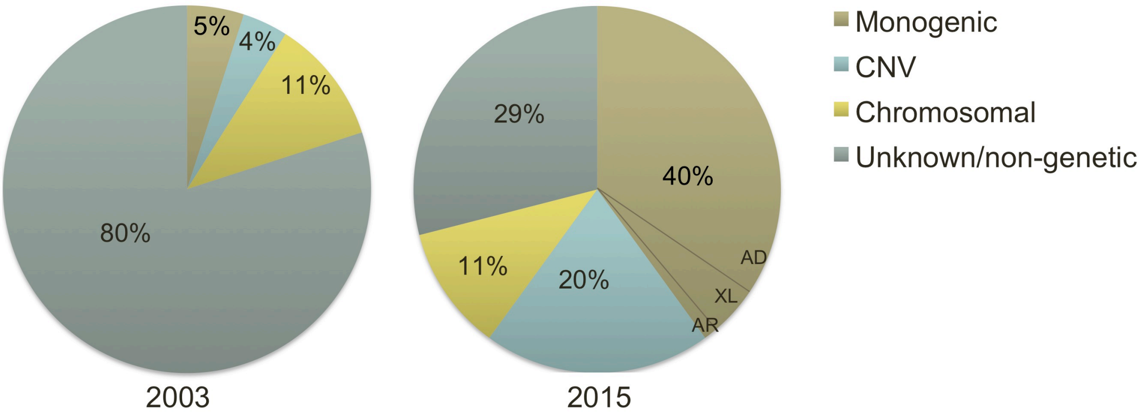

Figure 3. Established causes of intellectual disability in 2003 and 2015
Based on data from Stevenson et al 2003[23]*, Gillisen et al 2014*[21] *and Vissers et al 2015*[22]*.*

Taken together, the data from 2003 and 2015 illustrate the tremendous progress within this field, which has been enabled by the rapid advances in methodology; the introduction of microarrays and more recently massive parallel sequencing, during the same time period. The methodologies are discussed in more detail in chapter 3.

1.1.4 Clinical presentation

Rare genetic disorders are characterized by a broad diversity of symptoms and signs, ranging from mild features affecting only part of the body to severe manifestations involving multiple organ systems. The age of onset ranges from the prenatal period into late adulthood. The nervous system is commonly affected, resulting in symptoms such as intellectual disability, epilepsy, neuropsychiatric disorders and motor dysfunction. In OMIM, nearly half of the disorders with known etiology (47%) get listed when searching for disorders with "ID or epilepsy or neurologic features".[5]

The particular vulnerability of the nervous system may result from its dependence on many different proteins, within and outside the nervous system, for adequate formation and maintenance of complex structures and functions. Dependence on many proteins or genes for normal function implies a large target for genetic aberrations that may give rise to symptoms of disease.

Phenotypes that have been of particular interest in the thesis are those of early onset severe disorders. These are often characterized by ID, usually in combination with other features such as additional neurologic symptoms, congenital anomalies and dysmorphic features. Some of the disorders do not have ID as a major phenotypic finding, but instead one or several of the non-ID features mentioned above as main characteristics.

1.2 GENETIC VARIANTS

1.2.1 Definitions

The term "genetic variant" is used in this text for any alteration of the DNA-sequence or structure, when compared to a reference genome, regardless of its potential functional effect.

The terms "pathogenic variant" and "disease-causing variant" are regarded as synonyms in the text, and defined as variants that cause overt disease in an individual. However, it should be noted that pathogenic variants will not always be disease-causing, for example a pathogenic variant in a recessive gene usually does not cause a phenotype in heterozygous carriers. Furthermore, the synonymous use of these terms is applicable only for disorders that are due to fully penetrant variants, which is the case for the disorders included in the thesis.

The term "deleterious variant" is used for variants that are predicted to severely affect protein function or expression, but not necessarily lead to disease.

To avoid any confusion and in accordance with present recommendations[24], the term "mutation" is not used in the text. (If by accident the term is used, it would refer to a pathogenic variant.)

1.2.2 Spectrum of genetic variants in rare disorders

As described previously, the etiology of rare disorders is diverse with different types of genetic variants and inheritance patterns. Traditionally, disease-causing genetic variants have been divided into chromosomal abnormalities, CNVs and monogenic variants with an overall focus on variants within or including genes. Division into these groups is still useful, but with advanced understanding of the mechanisms behind genetic disorders, the boundaries between the groups have become blurred. Genetic variants could be regarded more as a continuum ranging from small changes in the DNA sequence (single nucleotide variants or insertions/deletions of a few nucleotides) and repeat expansions to structural variants of varying sizes. Structural variants can be either balanced (inversions, translocations including insertions and complex rearrangements) or unbalanced with the latter also referred to as copy number variants (deletions or duplications).[25] The size cut-off for what should be defined as a CNV was originally set at deletions or duplications >1 kb, but a more recent size definition is >50 bp.[26] Most of the rare genetic disorders are caused by variants that reside either within a protein-coding gene or include one or several such genes, but in some cases the underlying defect may be localized to a non-coding region.[27, 28] In addition, there are other types of variants such as uniparental disomy that cause some of the rare disorders. Focus in the studies included in the thesis lie on intragenic variants including sequence variants and CNVs.

1.2.3 Mechanisms underlying CNV formation

The mechanisms behind CNV formation are complex and knowledge is continuing to evolve. Frequent and relatively well-characterized mechanisms include non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ) and fork-stalling and template switching (FoSTes)/micro-homology-mediated break induced replication (MMBIR).[29]

Homologous recombination (HR) occurs naturally in cells; between homologous chromosomes in meiosis to increase genetic diversity and between or within chromatids to repair double-stranded (ds) DNA-breaks prior to mitosis.[30] NAHR can occur in either of these settings, meiosis as well as in mitosis, and is the most common mechanism behind recurrent CNVs. In the case of NAHR, there is misaligning of two DNA-sequences due to the presence of multiple highly similar DNA-stretches such as low copy repeats (LCRs), followed by recombination.[31]

NHEJ is another naturally occurring repair mechanism for ds DNA-breaks. Unlike HR, it can operate in the absence of a homologous template by “simply” joining the DNA ends. If there are two or more ds DNA breaks, errors that lead to CNVs may occur. Classical NHEJ is believed not be dependent on homology between the DNA-strands. However, presence of short homologous regions, i.e. micro-homology, between the ends of the DNA-strands may facilitate the repair process. NHEJ often results in small insertions or deletions of a few nucleotides at the ligation point (in addition to the larger CNVs that may be the result of faulty NHEJ).[31, 32] A more recently described mechanism, which is a variant of NHEJ, is the micro-homology-mediated end joining (MMEJ), which relies on a different set of repair proteins and on the presence of micro-homology.[33] The prevalence of MMEJ in humans remain to be elucidated.

Lastly, FoSTes/MMBIR is a mechanism that may occur during DNA replication due to stalling of the replication fork. One of the newly synthesized single strands may detach from the template, alternatively, the fork collapses and one of the strands breaks resulting in a “lose end” ds-DNA. The single-stranded (ss) DNA or an overhanging part of the ds-DNA may thereupon anneal to another template with micro-homology to the original template and continue to replicate. This may happen one or several times before returning to the original template.[29, 34]

1.2.4 Normal variation in the human genome

Inter-individual variation

The different types of genetic variants that may cause rare disorders are outlined above in chapter 1.2.2. During the past ten to fifteen years, it has become increasingly clear that the same types of genetic variants are present all over the genome in any human and account for normal inter-individual genetic variation.[35-37] The genomes from two individuals are 98-99% similar, while the remainder differs between the two. A large study on human genetic variation estimates that the difference between the genome of one individual and a reference genome is 0.1% due to SNVs, 1.2% due to CNV/indels and 0.3% due to inversions.[38] These figures correlate to findings that individuals carry on average 3 million SNVs and more than 1000 CNVs (>500 bp) when compared to a reference genome.[39, 40]

Deleterious genetic variants in healthy individuals

Recent studies have shown that the genome from a healthy individual may contain as many as 100 seemingly deleterious variants, mostly in a heterozygous state, but also some (0-20) bi-allelic variants.[41, 42] There are several possible explanations for the absence of a disease phenotype despite these variants. It has been shown that many human genes are haplosufficient[43], so for heterozygous variants, there may be sufficient expression from the wild type allele. Regarding bi-allelic variants, there may be residual protein function, compensation by similar genes/proteins, variants that only affect non-essential transcripts or variants in genes that are dispensable.[42]

De novo variants

Some of the variants that are seen in an individual have arisen *de novo*. All humans carry a number of SNVs that are not present in samples from the parents. The number is estimated at approximately 70 SNVs per individual genome[44] or approximately one non-synonymous SNV per individual exome[18]. These figures correlate to the age of the father with an increase of 2 SNVs per year.[45] *De novo* CNVs or indels are not as prevalent as *de novo* SNVs. Large *de novo* CNVs (>50 kb) occur in approximately one out of 50 individuals[46] while smaller *de novo* variants (indels <50 bp) occur in all individuals at a rate of approximately 9 per individual genome[44].

1.2.5 Normal variants versus disease-causing variants

To summarize the above paragraphs; each human genome contains millions of variants that are not present in a reference genome, many of these are seemingly deleterious (without strong phenotypic effects) and some of them are *de novo* variants. With this in mind, predicting the functional effect of a genetic variant is sometimes very challenging. The effect of a specific genetic variant and its impact on an individual's health is determined by several factors such as genomic localization, size of the variant, nucleotides involved and more. Making a distinction between a disease-causing variant and a normal variant is one of the major challenges in human genetics today, both in research as well as in a clinical setting. Through the introduction of massive parallel sequencing techniques millions of genetic variants can be detected in a single individual, which requires a process for filtering and interpretation. This process and the methods used are discussed in chapter 3.2.

1.3 AUTOSOMAL RECESSIVE DISORDERS AND CONSANGUINITY

The statistics on inheritance and etiology of rare disorders, as illustrated in figure 2B and 3, imply that the majority of rare disorders are inherited in an AR pattern, while only a minority of the patients have an AR disorder. This discrepancy indicates that there are numerous AR disorders that are, not only rare, but extremely rare with a very low prevalence for each disorder. Nevertheless, understanding the specific etiology of these disorders is highly relevant, as it increases the chances of achieving an etiological diagnosis for patients with AR disorders. Diagnosing these disorders is critical, since the recurrence risk is high (25%) and a specific diagnosis is a prerequisite for genetic counseling, prenatal diagnostics in future pregnancies, as well as for counseling issues regarding prognosis and, in some cases, treatment options. Understanding and diagnosing AR disorders is of particular importance for patients whose parents are consanguineous since the prevalence of AR disorders is increased in this subgroup of rare disease patients. The genetic basis for this increase and the recent advances in diagnosing AR disorders are discussed in more detail below.

Consanguinity is defined as a union between two individuals who are related as second cousins or closer and is common in many parts of the world.[47] Rare disorders are somewhat more prevalent in the offspring of consanguineous couples compared to the prevalence in an outbred population. For children born to parents that are first cousins, the estimated prevalence of a congenital or early onset disorder is 5-8%, which is approximately twice the prevalence seen in an outbred population. The observed increase is attributed to disorders that are inherited in an autosomal recessive pattern, which is a rare inheritance pattern in the general population (Figure 4).[47-49] However, it is of importance to emphasize that despite the increased prevalence of AR disorders, the vast majority of children born to consanguineous parents are completely healthy.

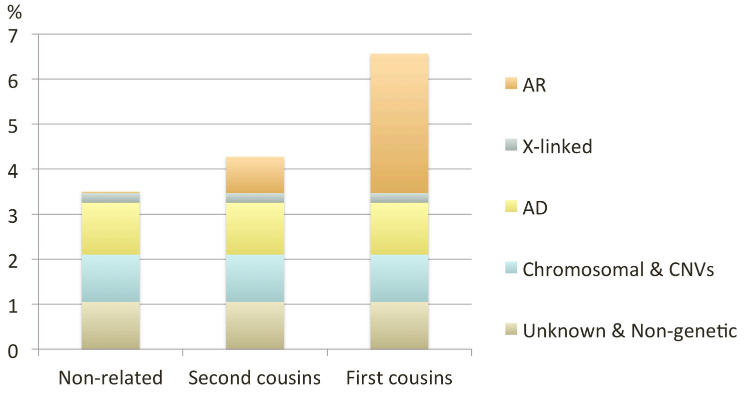

Figure 4. Prevalence and etiology of congenital disorders by parental relatedness
Based on data from Bittles et al[47], Hamamy et al[48], Zlotogora et al[49] and Powis et al[50].

The basis for an increased prevalence regarding AR disorders in the offspring of consanguineous couples is parental sharing of alleles that each has a low carrier frequency in the general population. The proportion of sharing depends on the degree of relation between two individuals. For example, first cousins share 1/8 of their genome, which implies that they will share the same allele at 1/4 of the loci. As for AR inheritance in general, the risk for offspring of parents that carry the same allele (or different variants in the same gene) to inherit both these alleles is 1/4. Taken together, this means that the offspring of first cousin parents will be homozygous at 1/16 (1/4x1/4) of their genome. In the case where homozygosity is due to inheritance from one (or several) common ancestors the terms *autozygosity* or *identity by descent* may also be used. If a pathogenic variant resides within the autozygous regions, this may give rise to an AR disorder (Figure 5).

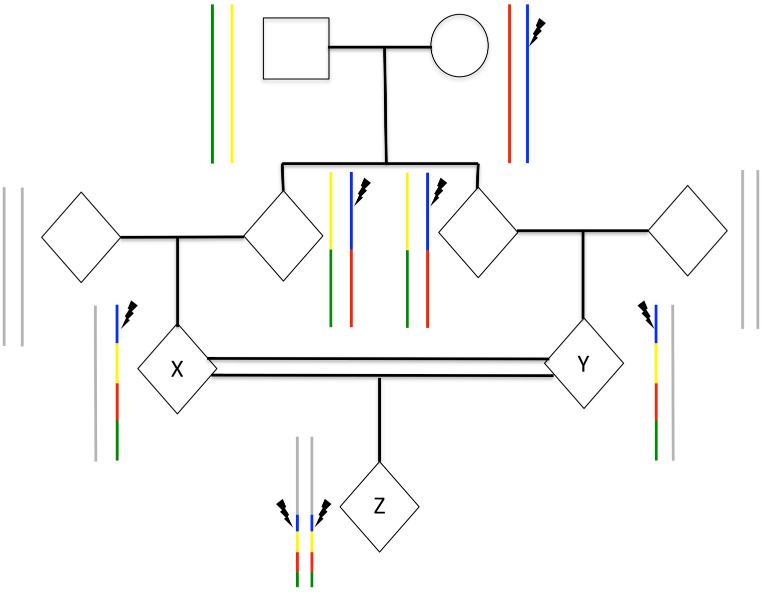

Figure 5. Allele sharing in first cousins and autozygosity in the offspring
A four-generation pedigree illustrating an individual (Z) born to parents that are first cousins (X and Y). Each vertical line next to the individuals illustrates the haploid genome of the individual. The colors represent the proportion of the genome that is inherited from the common ancestors in the first generation of the pedigree. X and Y share the same allele at 1/4 of the loci. (The probability that X and Y have the same color at a locus is one in four as illustrated by the four colors). Z is autozygous at 1/16 of the loci. (The probability that the maternal allele in a locus derive from the common ancestors is 1/2 as illustrated by half the haploid genome being colored. The probability that the paternal allele derive from the common ancestors is also 1/2. The probability that these alleles are the same is 1/4 as illustrated by the four colors. Altogether this equals 1/16). The lightning-symbol illustrates a pathogenic variant.

Obviously, the majority of the alleles that are shared between first cousins do not harbor any disease-causing variants. The likelihood of carrying an autosomal recessive disease-causing variant is not different for an individual in a consanguineous marriage than for any person in the general population. The carrier rate for autosomal recessive disease-causing variants is estimated at 2-5 per individual.[51] In one study by Bell et al., 437 recessive disease genes were sequenced in 104 healthy individuals and there were on average 2.8 (0-7) recessive deleterious variants per individual.[52] From these findings, estimating the frequency of disease-causing variants in the complete genome would clearly result in higher figures. On the other hand, large scale sequencing studies find that healthy humans carry a number of seemingly deleterious recessive variants in a homozygous or compound heterozygous state, which suggest that not all seemingly deleterious variants result in an overt disease phenotype.[41] Taken together, the estimate of each individual carrying 2-5 recessive pathogenic variants seems rational and, interestingly, it correlates very well to early theoretical estimates.[53] The recessive variants seen in healthy carriers are most frequently small sequence alterations. However, a study by Boone et al. show that approximately one out of seven individuals carry heterozygous CNVs affecting recessive disease genes.[54] In summary, for children born to consanguineous parents, the combination of inter-parental allele-sharing and the fact that humans carry a number of AR disease-causing variants lead to an increased risk for bi-parental inheritance of such variants, which in turn may manifest as an AR disorder.

Similar to what has been achieved in recent years when it comes to defining the etiology of rare disorders in general, there have been major advances in understanding the specific etiology of AR rare disorders, largely by studying consanguineous families. The first study on a large cohort of patients, by Najmabadi et al, took advantage of the massive parallel sequencing (MPS) technology (see methods 3.2.4) and identified the molecular etiology in 78 out of 136 consanguineous families. Only in this study, 50 novel genes were identified and for the majority of the diagnosed families (73/78) the identified gene was unique to that family.[55] These figures illustrate the extreme heterogeneity in AR rare disorders as well as the rapid pace of novel gene discovery made possible by the MPS technology. This method was applied also in a study of outbred families with multiple affected children and identified potential AR disease genes in 5 out of 20 families.[56] Several recent studies on the etiology of AR disorders in cohorts of consanguineous families have been successful using MPS.[57-62] In a clinical setting with an unselected group of patients, it is important to remark that, in addition to AR disorders, other etiologies should also be considered for children to consanguineous parents. A study of unselected consanguineous families analyzed with MPS in a clinical setting revealed AR inheritance in 62% of the cases with a monogenic etiology.[50]

2 AIMS

The aims of the thesis were to:

- Determine the specific etiology in patients with rare congenital disorders
- Identify new disease genes in patients with rare congenital disorders
- Delineate the phenotypes associated with these disorders

3 MATERIAL AND METHODS

3.1 PATIENTS

The patients included in the studies were referred to the Department of Clinical Genetics, Karolinska University Hospital in Stockholm, from 2008 through 2014 for assessment and investigation due to a clinical phenotype that raised suspicion of a rare congenital disorder with an underlying genetic etiology.

All patients who had at least one sibling with the same phenotype, normal findings on clinical array-CGH and parental consanguinity were included for further investigations. In addition, one sporadic patient with a rare CNV on clinical array-CGH was included for further studies.

Genomic DNA was extracted, using standard protocols, from blood samples from all patients as well as the unaffected siblings and parents. The patients further underwent thorough clinical examinations, including radiologic and biochemical investigations.

The studies were approved by the Regional Ethics Committee in Stockholm and written informed consent was obtained from each participating individual or their legal guardians.

3.2 METHODS

3.2.1 Outline of the studies

For the group of patients with consanguineous parents, an autosomal recessive rare disorder due to a homozygous genetic variant was hypothesized. DNA from all affected individuals as well as unaffected siblings and parents were analyzed with whole exome sequencing followed by a process of filtering, assessment and validation of the detected variants, including Sanger-sequencing of variants that were assessed as pathogenic or likely pathogenic.

For the case with a rare CNV on array-CGH, further analyses with MLPA and mate-pair sequencing were performed in order to delineate the exact size and location of the CNV as well as to better understand the mechanism of formation. These data combined with detailed phenotype data were analyzed in determining the etiology of this rare disorder.

The methods used and their resolution are illustrated in Figure 6. (Chromosome analysis is included for comparison.)

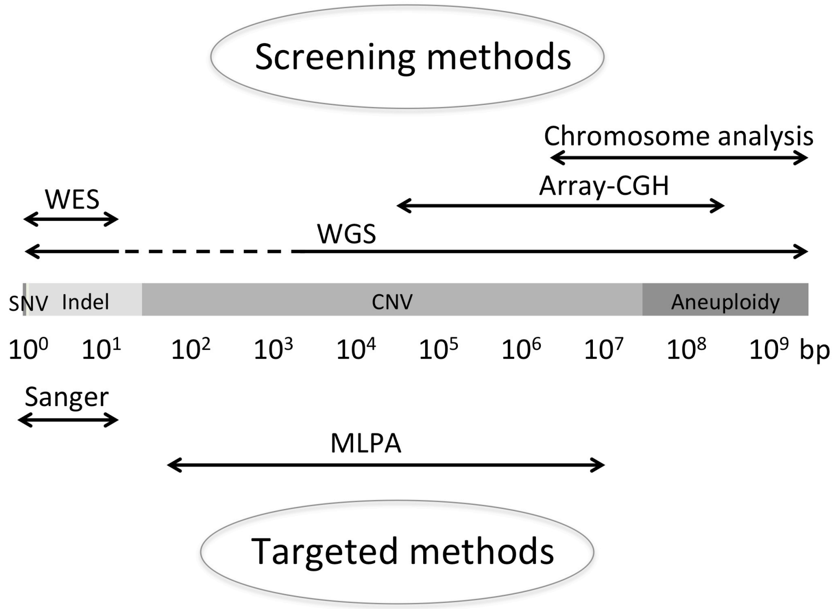

Figure 6. Resolution of different methods for detecting genetic variants

3.2.2 Array-CGH

Microarray-based comparative genomic hybridization (array-CGH) allows for a high-resolution screening regarding DNA copy number variation, i.e. CNVs. The method is based on the comparison of test and reference genomic DNA samples that are labeled with two different fluorescent colors and hybridized to unique oligonucleotides (probes) positioned on a glass slide. The amount of test-DNA versus reference-DNA in each position is measured from the intensity of fluorescent color-signal and analyzed as a ratio between the two, the log2-ratio. Resolution is dependent on the number of probes. The type of array-CGH used for the studies was a 180K SurePrint G3 Human CGH oligonucleotide microarray (Agilent Technologies, Santa Clara, CA) with an average resolution of approximately 50 kb, performed at the Department of Clinical Genetics, Karolinska University Hospital. Regions with a minimum of three consecutive probes with aberrant log2 ratios (above 0.35 for duplications and below -0.65 for deletions) were considered for further analyses.

3.2.3 MLPA

Multiplex ligation-dependent probe amplification (MLPA)[63] is another method to detect copy number variation. Unlike array-CGH, it is a targeted method that allows for the investigation of a limited number of genomic regions. The principle of MLPA is that for each locus of interest, two DNA oligonucleotides (probes) hybridize to their target sequences on the test DNA. If there is a perfect match, the two probes are ligated and subsequently amplified via a PCR reaction. The product from each probe-pair has a specific length (130-490 bp) and can thus be separated by electrophoresis and quantified. The amount of each PCR-product correlates to the amount of test DNA in that specific locus.

An algorithm for normalization within the test-DNA (via calculation of a ratio between the amount of PCR-products from each test-probe to the amount of PCR-products from control-probes that hybridize to sequences which are not subject to copy number variation) as well as between the test DNA and normal control DNA is applied to ensure stable and reliable results.

3.2.4 MPS

Massive parallel sequencing (MPS), also known as next generation sequencing (NGS) or high throughput sequencing, is a collective term for different technologies that all have in common the simultaneous sequencing of numerous genomic regions. The technology used in the thesis studies is sequencing by synthesis from Illumina (Illumina Inc, San Diego, CA, USA) based on reversible terminator chemistry.[64] The process for detecting different types of genetic variants consists of four steps; (1) library preparation, (2) cluster generation, (3) sequencing and (4) data-processing including alignment to a reference sequence for final detection of the variants. An overview of the steps is illustrated in Figure 7.

Library preparation starts with a random fragmentation of the DNA-sample and can then be performed in a number of manners with different types of data-handling and resulting output in the last step. For the studies in the thesis, whole-exome sequencing and, in one case, mate-pair sequencing have been applied. Library preparation for WES has the purpose of capturing all protein-coding exons in the genome (180.000 exons). The commercial Agilent SureSelect Human All Exon 50M kit (Agilent Technologies, Santa Clara, CA, USA) was used. This method is based on an in-solution hybridization of sample DNA to biotinylated RNA-oligonucleotides that are complementary to human exonic sequences. The RNA-DNA fragments are made magnetic by adding magnetic micro-beads covered with streptavidin that bind to biotin, whereupon magnetism may be used for capturing. Another type of library preparation is that for mate-pair sequencing. The purpose of this method is to create pairs of DNA-sequences that are thousands of bp apart, which is informative regarding genome structure and can be used for the detection of SVs. In order to achieve this, sample DNA-fragments as long as 2-5 kb are used. These are circularized and the ends are joined, followed by another round of fragmentation and purification that results in smaller pieces of DNA that correspond to the ends of the initial fragment. Notably, these pieces have an inverted orientation.

Prior to sequencing, the DNA-fragments to be analyzed are bound to the surface of a lane in a flow cell and each fragment is amplified via bridge-PCR to create a cluster of identical DNA sequences that are all bound to the surface. The number of clusters/per lane is in the range of 100-200 million.

Sequencing is performed using the DNA-fragments on the flowcell as templates and sequentially incorporating dye labeled nucleotides. Only one nucleotide at a time can be added due to a terminator linked to the nucleotide. After each addition, the fluorescent dye signal is imaged and the corresponding nucleotide is recorded. The terminator and dye is then washed away, whereupon the cycle is repeated until the sequence of around 100 bp is complete. For paired-end sequencing, a second round of cluster generation and sequencing follows, which renders sequence data from the opposite end of the DNA-fragment.[65]

The last step involves processing of all the data generated from sequencing. There are various programs for different parts of this process and for addressing different issues, such as the type of genetic variants to be detected (e.g. SNVs or SVs). In brief, the goal is to align the achieved sequences to a reference genome and detect the differences between the two.

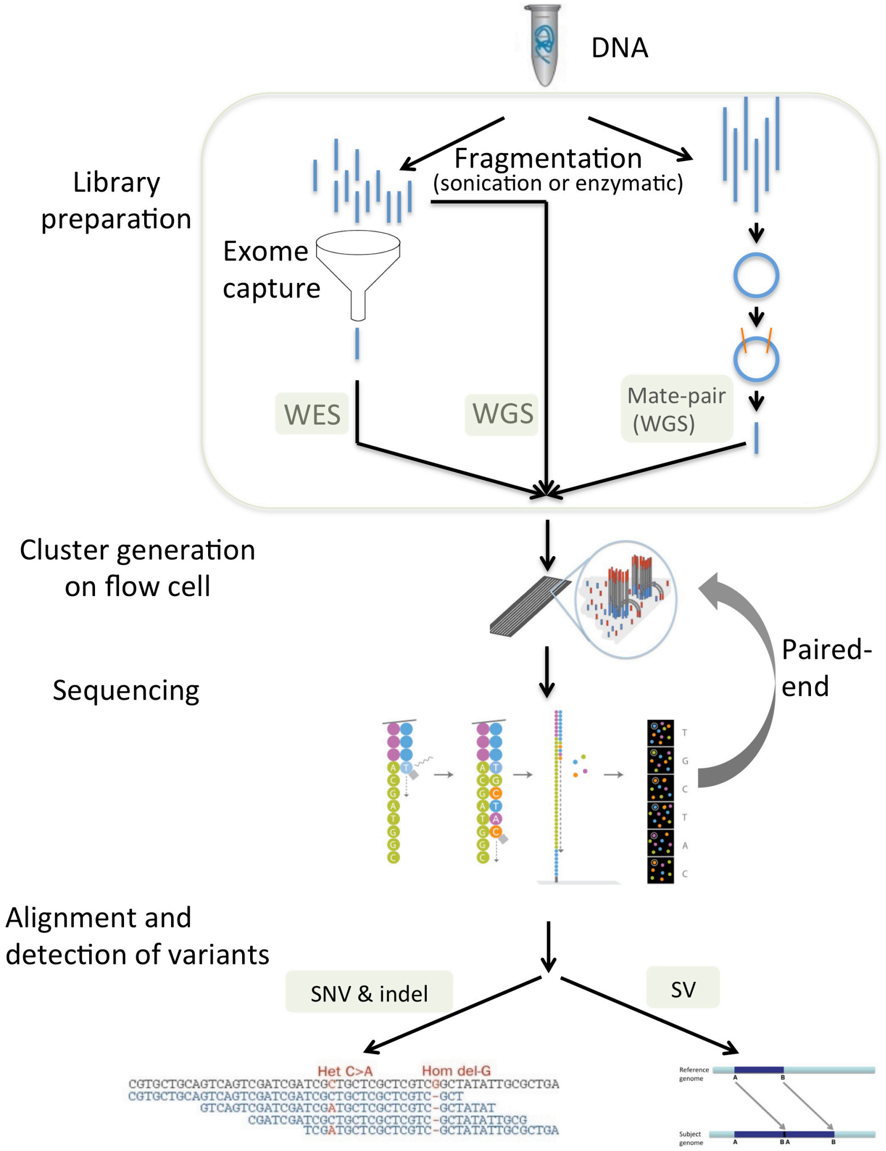
DNA
Library preparation
Fragmentation
(sonication or enzymatic)
Exome capture
WES
WGS
Mate-pair (WGS)
Cluster generation on flow cell
Sequencing
Paired-end
T
G
C
T
A
C
Alignment and detection of variants
SNV & indel
SV
Het C>A
Hom del-G
CGTGCTGCAGTCAGTCGATCGATCGCTGCTCGCTCGTCGGCTATATTGCGCTGA
CGTGCTGCAGTCAGTCGATCGATCGCTGCTCGCTCGTC-GCT
GTCAGTCGATCGATCGATGCTCGCTCGTC-GCTATAT
CGATCGATCGCTGCTCGCTCGTC-GCTATATTGCG
TCGATGCTCGCTCGTC-GCTATATTGCGCTGA
Reference genome
A
B
Subject genome
A
B A
B

Figure 7. Overview of the steps in MPS using Illumina technology[65]
The figure is a simplified overview of the steps in MPS using Illumina technology, leaving out a number of steps such as adapter-ligation, PCR-amplification and purification. For explanation of the different steps, please refer to the text.

3.2.5 Interpreting variants detected by genetic screening methods

Genetic screening methods such as array-CGH or MPS detect numerous genetic variants in an individual. For most patients with a rare disorder, only one or a few variants are pathogenic (i.e. causative of the disease-phenotype). In order to identify the disease-causing variant(s), various measures for filtering, prioritization and evaluation are available, schematically shown in figure 8.[66] Filtering and prioritization are facilitated by the use of databases and tools for predicting the functional effect of genetic variants. They are valuable also for evaluation of variants, which rely on manual assessment and sometimes additional experiments, based on the observed phenotype.

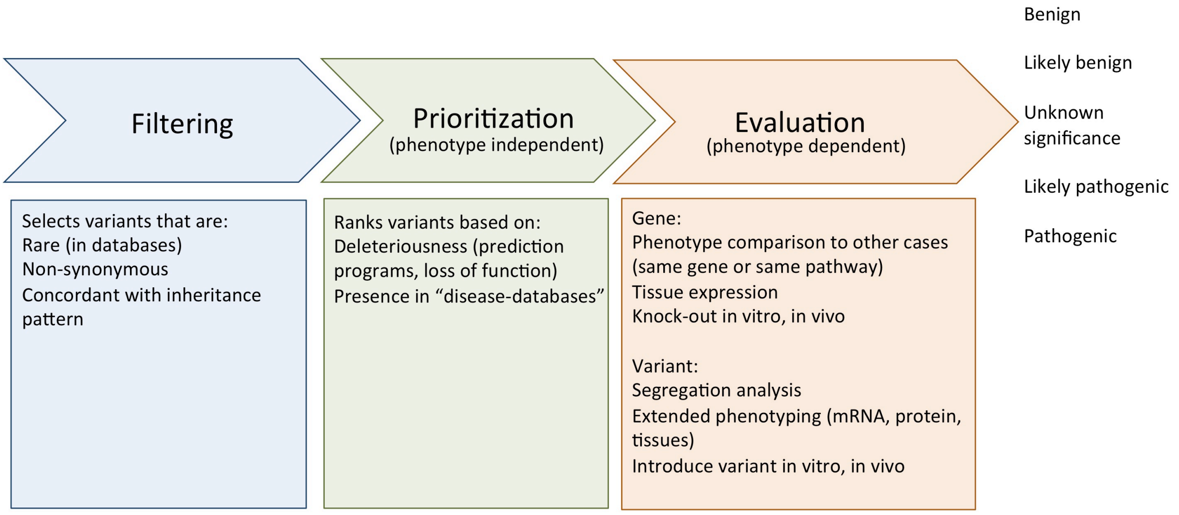
Filtering
Prioritization
(phenotype independent)
Evaluation
(phenotype dependent)
Benign
Likely benign
Unknown significance
Likely pathogenic
Pathogenic
Selects variants that are:
Rare (in databases)
Non-synonymous
Concordant with inheritance pattern
Ranks variants based on:
Deleteriousness (prediction programs, loss of function)
Presence in "disease-databases"
Gene:
Phenotype comparison to other cases (same gene or same pathway)
Tissue expression
Knock-out in vitro, in vivo
Variant:
Segregation analysis
Extended phenotyping (mRNA, protein, tissues)
Introduce variant in vitro, in vivo

Figure 8. Overview of the process for interpreting genetic variants detected by genomic screening

Databases for genetic variants

Great efforts have been made in creating useful databases with collections of normal variants and/or disease-causing variants to aid in the interpretation of variants identified in patients. Databases that collect disease-causing variants are for example DECIPHER[67], which traditionally have focused on CNVs and the Human Gene Mutation Database (HGMD)[68], whose main focus has been on SNVs. However, both databases now include different types of variants. Regarding normal variants, these are recorded in, for example, Database of Genomic Variants (DGV)[26] with main focus on CNVs, and dbSNP[69] or ExAc[70], who both focus on SNVs. Recording of phenotype data in databases has become increasingly important for assisting in the interpretation of variants and assigning pathogenicity to variants. The comparison of phenotypes in different patients who have variants affecting the same gene or genes is highly informative in the process of assessing genetic variants. Many databases, such as DECIPHER, have included phenotype data in a standardized format based on the Human Phenotype Ontology (HPO).[71] Other databases such as OMIM, include phenotype data in a less strict manner.

Tools for predicting the functional effect of genetic variants

There are numerous tools for predicting pathogenicity of genetic variants. These programs use different algorithms and hence, the outcome may differ depending on program. For the thesis studies, different tools have been used with Combined Annotation Dependent Depletion (CADD)[72] as the major one. It has the advantage of integrating a number of annotations to produce a combined score. The scaled CADD score relates the variant of interest to all possible theoretical variants in the genome resulting in log-scaled number. For example, a score of 20 means that the variant is in the top 1% of the most deleterious variants, a score of 30 in the top 0.1% etc.

Evaluation and scoring of genetic variants

After narrowing down the number of potential pathogenic variants by filtering and prioritization, manual evaluation of the remaining variants is possible. The number of variants to evaluate depends on filtering and cut-off level for prioritization. In the thesis, filtering on a minor allele frequency of 1% and segregation according to an autosomal recessive pattern was applied and all remaining variants, regardless of priority, were evaluated. The variants were evaluated with respect to their relevance for the observed disease phenotype with a resulting score (benign, likely benign, unknown significance, likely pathogenic or pathogenic). Notably, for a given variant these scores may be subject to change after achievement of results from additional investigations. In order to score a variant as pathogenic, the following criteria was used; (a) a phenotype that is identical or similar to previously described cases with pathogenic variants affecting the same gene and segregation in the family, in combination with either a variant that was previously reported as pathogenic or a non-sense/frame-shift variant or a variant with scaled CADD-score >20 or (b) a variant in a novel gene with several lines of evidence for functional effects and causality.

Evaluating the potential pathogenicity of a variant largely depend on the phenotype observed in the individual as well as in other members of the family. Comparison of the observed phenotype to other cases with variants affecting the same gene or genes in the same pathway is informative. Additional information on a gene level can be achieved by data on expression in the tissue of interest and functional assays in "knock-out" cell-lines or animal models. The latter may be used also for assessing the effect when introducing a specific genetic variant. Evaluation of the functional effect of a specific variant can likewise be performed by analyses in the individual itself, which can be considered as an extended phenotype characterization or molecular phenotyping. These analyses may also include family members as part of a segregation analysis (Figure 8).[66]

4 RESULTS AND DISCUSSION

4.1 RARE DUPLICATION IN SPORADIC SYNDROMIC INTELLECTUAL DISABILITY

(PAPER II)

Array-CGH identified a small *de novo* duplication on chromosome 2 in a male patient with intellectual disability, speech and language impairment, cleft palate, malformed teeth, and oligodontia. The finding was confirmed with MLPA, which showed that the duplication included exon 5, 6 and 7 of the *SATB2* gene. Further analysis with WGS using mate-pairs proved that the duplicated region (≈35 kb) was intragenic and arranged in tandem. Closer examination with Sanger-sequencing of the breakpoint junction revealed a 3 bp sequence of micro-homology shared between the distal and proximal breakpoints (Figure 9).

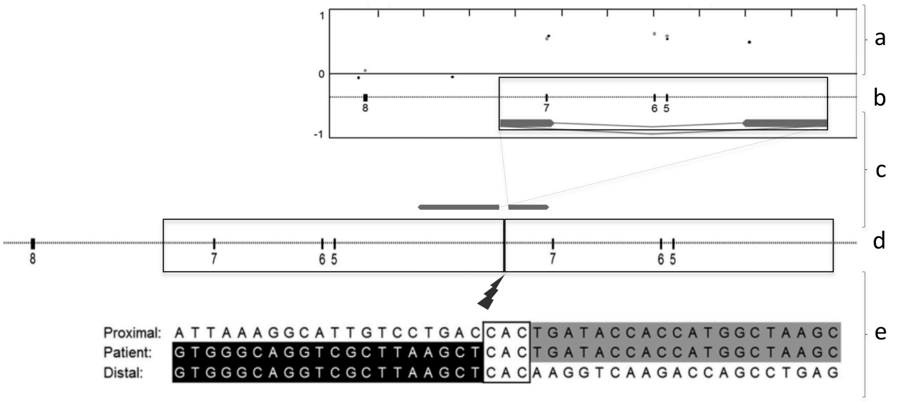
1
0
-1
a
b
8
7
6 5
c
d
8
7
6 5
7
6 5
e
Proximal: ATTAAAGGCATTGTCCTGAC CAC TGATACCACCATGGCTAAGC
Patient: GTGGGCAGGTCGCTTAAGCT CAC TGATACCACCATGGCTAAGC
Distal: GTGGGCAGGTCGCTTAAGCT CAC AAGGTCAAGACCAGCCTGAG

Figure 9. Schematic illustration of the intragenic duplication in SATB2
a) Results from array-CGH and MLPA showing probes consistent with duplication of exon 5-7. b) Reference gene c) Mate-pair sequencing reads from DNA-fragments that span the breakpoint junction in DNA from the patient (lower) showing an inverted (forward-reverse) mapping-orientation within the reference gene (upper) d) Illustration of the tandem duplication in the patient. e) Sanger-sequencing of the break point junction in the patient showing that 3 basepairs (CAC) are common to the proximal and distal breakpoints, i.e. micro-homology between the regions. Adapted from figure 2 in paper II.

The findings of a direct tandem orientation of the duplication and micro-homology in the breakpoint junction were in line with recent studies showing the prevalent occurrence of these findings for duplications. Studies of rare, non-recurrent duplications show that 80-90% are arranged in tandem. (The remaining duplications are part of complex rearrangements or due to an insertional translocation.)[73] Furthermore, micro-homology (2-70 bp) in breakpoint junctions of rare, non-recurrent, CNVs is a common finding seen in approximately 70-80% of these CNVs. The frequency is similar for deletions and duplications as well as for normal variants and pathogenic variants.[74, 75]

The conclusions drawn from the genetic results in the patient were that, based on the finding of micro-homology, the most likely underlying mechanism of formation was either a replication based mechanism (i.e. FoSTes/MMBIR) or NHEJ/MMEJ, which both depend on micro-homology (see chapter 1.2.3). Furthermore, the intragenic location of the duplication, strongly suggested a functional effect on the SATB2 protein. This conclusion was supported also from the phenotypic findings, which were very similar to the phenotype described in other cases with pathogenic *SATB2* variants. At the time of the study, there were only six patients reported with deleterious variants in (and confined to) *SATB2;* two cases with SNVs and four cases with intragenic deletions.[76-79] Our case was the first description of an intragenic *SATB2* duplication. As of today, there are two additional reports of cases with intragenic duplications in the same gene (exon 3 and exon 4, respectively).[80, 81]

With the introduction of MPS and screening of large cohorts of patients with rare disorders, variants in *SATB2* have shown to be a prevalent cause of syndromic intellectual disability. Two recent large MPS screening studies have reported a total of 12 cases with *de novo* SNVs or small indels in *SATB2.*[19, 82] In fact, *SATB2* appeared as one of the top five causative genes in the large DDD-study, when reporting the findings in more than 1000 children with developmental disorders.[19] A distinct and recognizable phenotype has emerged over time and together with results from the recent MPS-studies, the phenotype is now further confirmed. All patients seem to have intellectual disability (often moderate-severe), limited or absent speech and visible dental abnormalities (oligodontia, abnormal shape, crowding). Nearly all patients have cleft palate, a happy and jovial personality, micrognathia and distinct facial features. Osteopenia/osteoporosis is recorded in many patients. Seizures and abnormalities on MRI of the brain are rare, but can be part of the phenotype. Growth parameters including OFC are typically within the normal range.

There are yet no phenotypic findings that can differentiate between patients with SNVs versus intragenic CNVs, suggesting that there is a common underlying pathogenic mechanism that results from haploinsufficiency.

4.2 RARE DISORDERS IN CONSANGUINEOUS FAMILIES (PAPER/MANUSCRIPT I, III, IV)

4.2.1 Overall results in the cohort

A total of 20 families fulfilled the criteria of at least two siblings with the same rare disorder, consanguineous parents and normal findings on clinical array-CGH. These were included for further studies with detailed phenotypic investigations and WES of affected as well as unaffected siblings and parents, followed by a process of filtering and evaluation as described in chapter 3.

Genetic results

In 14 out of 20 families, a pathogenic variant causative of the observed phenotype was identified. Out of these “solved cases”, 11 families had variants in known disease genes and 3 families had variants in genes that were not associated to a disease phenotype at the time for initial evaluation (*PIGT, TFG* and *KIAA1109*). Through functional evaluation and/or additional independent cases that were subsequently published, the variants in *PIGT, TFG* and *KIAA1109* have proved to be pathogenic and we could thus confirm these genes as novel disease genes. In addition, 4 families had variants that were considered as possibly causative of their disease phenotype, although causal associations remain to be established. The detected pathogenic variants were all present in a homozygous state in the affected individuals while heterozygous in their parents and the variants were of various types; missense variants (n=6), nonsense variants (n=4), indels or frameshift variants (n=3) and large intragenic deletions (n=1) (Table 1). The finding of a homozygous deletion seen in 1 out of 14 families correlates to recent findings by Boone et al. who analyzed a cohort of >20000 individuals and estimated the carrier frequency of heterozygous recessive SNVs to be 13.5 times higher than the carrier frequency of heterozygous recessive CNVs.[54]

Phenotypic results

The most common presentation among the patients was intellectual disability (ID). On detailed clinical assessment, all patients with ID manifested additional features, ranging from mild traits such as dysmorphic facial features and abnormalities on biochemical testing to severe manifestations like intractable seizures and gross intracranial malformations. In many cases, identification of these findings were essential for evaluation of the genetic data from WES and proved to be crucial for ascribing pathogenicity to the detected genetic variants and reaching a diagnosis for the patient. Six families presented with a phenotype where ID was not seen as a main feature; in two families, there were severe disorders with prenatal onset and intrauterine or neonatal lethality, another two families were diagnosed with spastic paraparesis, one additional family presented a phenotype of severe myopathy and, lastly, one family had symptoms compatible with ectodermal dysplasia (Table 1).

Joint analysis of genotype-phenotype results

For the 14 families in whom a molecular etiology was established, the majority were affected by disorders that have been reported in only a handful of cases, in single families or by disorders with no previous patients reported (Table 1). The latter was the case for the family in whom a homozygous variant in *PIGT* was detected. Pathogenicity could be confirmed by functional validation, described in more detail below. Two families were affected by disorders that had previously been described only in single families. One of these families included three fetuses affected by fetal akinesia deformation sequence (FADS) in whom we identified a homozygous intragenic deletion (exon 28-55) in *KIAA1109*. This gene is not yet annotated as a disease gene in OMIM, but has recently been suggested as a candidate gene by Alazami et al. who identified a variant in *KIAA1109* in a family with a phenotype similar to that of our patients.[60] The second family was affected by spastic paraplegia and a homozygous pathogenic variant in the *TFG* gene was identified. Bi-allelic disease-causing variants in this gene have been previously reported in one family with complicated spastic paraplegia.[83]

Fam #	Results	Phenotype	Gene	Mutation type	Previously reported cases*
3	P (novel gene)	ID syndrome	PIGT	missense	0
4	P (novel phenotype)	ID syndrome	FLVCR2	missense	0 (40)
6	P	FADS	KIAA1109	CNV (del)	1
11	P	HSP	TFG	missense	1
17	P	FADS	ERCC5	nonsense	5 (>50)
18	P	Myopathy	MEGF10	frameshift	10
7	P	HSP	CYP7B1	nonsense	>10
10	P	ID syndrome	EXOSC3	missense	>10
12	P	ID syndrome	ASAH1	missense	>10
19	P	Ectodermal dysplasia	RIPK4	missense	>10
20	P	ID syndrome	MAN1B1	nonsense	>10
14	P	ID syndrome	ASPM	frameshift	>50
13	P	ID syndrome	WDR62	frameshift	>50
1	P	ID syndrome	ALMS	nonsense	>50
2	C	ID syndrome	GPX7, SUSD4	missense	
15	C	ID syndrome		intronic	
9	C	ID syndrome		intronic	
5	C	ID syndrome/HSP		intronic	
8	-	ID syndrome			
16	-	ID syndrome			

Table 1. Results of MPS-analysis and evaluation in the cohort of 20 families
*P, pathogenic variant (causative of the observed disease phenotype) identified; C, candidate gene/s identified; -, no candidate gene or pathogenic variant was identified; ID, intellectual disability; HSP, hereditary spastic paraparesis; FADS, fetal akinesia deformation sequence; *, numbers indicate how many cases have been reported with a pathogenic variant in the same gene and a phenotype that is identical or similar to the study case or, in parenthesis, a phenotype that is distinctly different from the study case*

A resemblance in clinical presentation between our patients and previously reported cases was true for most, but not all patients with an identified disease-causing variant. In the families with pathogenic variants in *MAN1B1, RIPK4* and *FLVCR2,* respectively, the observed phenotypes differed in some aspects, compared to the majority of previously reported cases. The family with variants in *FLVCR2* is described in more detail below.

For those cases in whom we were not able to establish the molecular etiology with certainty, there may be several reasons for this. The disease-causing variant may reside in a genomic region that was not captured by the method used. Furthermore, variants may be overlooked due to their position in genes for which current knowledge is insufficient regarding gene function and phenotypic effects of variants. Considering the families with variants that were scored as likely pathogenic, it may be very difficult to prove pathogenicity, partly due to lack of additional cases with variants affecting the same gene. Studies in order to evaluate these variants are ongoing. Thus, future results and re-evaluation of the data may increase the diagnostic yield.

The diagnostic yield of 70% in this study is comparable to the yield in other similar studies that have applied WES in families with several affected individuals and kinship between the parents. The yield varies from 36% to 95% between different studies.[57-62] One explanation to these differences is that some studies have included cases with likely pathogenic variants in novel disease genes among the positive cases, while others have only considered cases with variants in known disease genes.

4.2.2 PIGT – a novel disease gene associated to AR syndromic ID

In one of the families from the cohort described above, there were four patients with the same congenital disorder characterized by intellectual disability, hypotonia and seizures, in combination with abnormal skeletal and ophthalmologic findings. Results from WES identified a homozygous variant, c.547A>C (p.Thr183Pro), in the gene *PIGT* as the most likely disease-causing variant. The predicted protein alteration affects a highly conserved amino acid and several prediction programs scored the variant as deleterious (scaled CADD score 26). In addition, the variant segregated with the disease on analysis with Sanger sequencing of eight family members in total. *PIGT* encodes phosphatidylinositol-glycan biosynthesis class T protein, which is part of the glycosylphosphatidylinositol (GPI) anchor pathway. The gene was not previously reported as a disease gene. However, several other genes in the same biochemical pathway were associated to disorders with a common core phenotype that includes ID, seizures and abnormal levels of alkaline phosphatase – all of which were present in the patients from the study family. In order to functionally validate the detected variant in *PIGT,* we wanted to measure the level of GPI-linked proteins on the cell surface. This was achieved by flow cytometry, which showed that granulocytes from the patients had reduced levels of the GPI-anchored protein CD16b, supporting pathogenicity of the variant. Further functional *in-vivo* validation via morpholino-mediated knockdown of the *PIGT* ortholog in zebrafish (*pigt*) showed that, unlike human wildtype *PIGT* mRNA, the p.Thr183Pro encoding mRNA failed to rescue gastrulation defects induced by the suppression of *pigt*. When summarizing the results, we concluded that the detected homozygous variant in *PIGT* was causative of the observed phenotype and thus, *PIGT* represents a novel disease gene associated to syndromic ID.

Two additional families have subsequently been published[85, 86], further confirming *PIGT* as a disease gene, causative of multiple congenital anomalies-hypotonia-seizures syndrome 3 (MIM 615398).

4.2.3 Pathogenic variants in FLVCR2 are compatible with survival beyond infancy

In another family from the cohort described above, a brother and a sister were affected by a disorder of severe intellectual and neurologic disabilities. They had no functional movements, nor any means of communication and they suffered from seizures. Imaging of the brain showed calcifications, profound ventriculomegaly with only a thin edging of the cerebral cortex and hypoplastic cerebellum. WES revealed, in both patients, a homozygous variant, c.1289C>T (p.Thr430Met), in the gene *FLVCR2.* The variant was predicted to be deleterious upon analysis with several prediction programs and pathogenicity was further supported by segregation in the family with neither of five healthy members carrying the variant in a homozygous state. Additional support of pathogenicity came from a previous report of the variant being detected in a compound heterozygous state in a fetus with Fowler syndrome.[87]

FLVCR2 is a known disease gene causative of proliferative vasculopathy and hydranencephaly-hydrocephaly syndrome (MIM 225790), also known as Fowler syndrome, which was previously considered prenatally lethal. The features described in prenatal cases are glomerular vasculopathy in the central nervous system, severe hydrocephaly, hypokinesia and arthrogryposis. These features and the findings in the study patients are similar. However, there is a striking difference in survival which prove that Fowler syndrome is not always prenatally lethal, but may be compatible with survival beyond infancy.

5 CONCLUDING REMARKS

The studies in the thesis have focused on determining the molecular etiology for rare congenital disorders and delineating the phenotypes associated with these disorders.

The specific conclusions from the results are summarized below. In brief, the results:

- Expand the spectrum of pathogenic variants in *SATB2* and confirm the presence of a distinct and recognizable *SATB2*-deficiency phenotype
- Establish *PIGT* as a novel disease gene
- Confirm *TFG* and *KIAA1109* as novel disease genes
- Expand the phenotypic spectrum for disorders associated with variants in *MAN1B1, RIPK4* and *FLVCR2*
- Identify novel candidate genes for congenital disorders

There are several overall conclusions to be drawn from the studies. First of all, WES is a very powerful method for the identification of disease-causing variants in consanguineous families. Furthermore, the diversity of AR diseases among these families is enormous with many of the identified disorders being extremely rare. An additional conclusion is that a detailed phenotypic assessment is crucial for interpretation of data from large-scale genetic screening and for ascribing pathogenicity to the identified variants. Moreover, the full spectrum of genetic variants, including sequence alterations and CNVs, should be considered for the etiology of rare disorders.

The results altogether add detail to the clinical presentations of the given disorders and expand the number of genes and genetic variants with a presumed or established causal association to congenital disorders. Ultimately, this may increase the chances to achieve a genetic diagnosis for future patients.

6 FUTURE PERSPECTIVES

There are several challenges and also great opportunities for future studies of congenital disorders and their etiology. One of the challenges will be to improve the genome wide detection rate for different types of genetic variants as well as to improve interpretation of these variants. A related future perspective regards potential prevention of congenital disorders by carrier screening for couples at increased risk of having a child with an autosomal recessive disorder or by prenatal screening for pathogenic *de novo* variants and the interpretation of variants detected. A further important field of research that may expand in the future concerns understanding of the molecular pathomechanisms behind these disorders and development of treatment options.

Improved detection rates for common genetic variants (SNVs, indels and CNVs)

Despite the advances in technology over the last years and the increase in diagnostic yield for patients with rare disorders, there is still a large proportion of the patients in whom the etiologic diagnosis remains unknown.[22] By applying WGS instead of WES, the diagnostic yield increases significantly. For a population of patients in whom no etiology was established by a combination of micro-array and WES, the molecular etiology could be identified in 42% by WGS. The etiologies detected by WGS were small CNVs (38% of the diagnosed cases) and SNVs/small indels in coding regions (62% of the diagnosed cases).[21] In other words, some of the pathogenic variants in coding regions are missed by WES and small CNVs are difficult to detect on micro-array or WES. If cost was not an issue, WGS would therefore be the method of choice in both research and clinical setting. In the future, costs are likely to drop, enabling a more widespread use of WGS.

Detection of "alternative" genetic variants and mechanisms

Even with an improved detection rate of SNVs, small indels and CNVs in coding regions of the genome, there is still a proportion (30-40%) of the rare disease patients in whom an etiologic diagnosis can not be established. Some of these disorders may be caused by alternative types of genetic variants and mechanisms while others may be due to any genetic variant that escape recognition or pathogenicity establishment because of currently insufficient data for interpretation.

Genome wide screening for alternative variants or mechanisms include search for somatic mosaicism, variants in non-coding regions of the genome, balanced structural variants, repeat expansions, epigenetic aberrations such as imprinting defects and uniparental disomy (UPD). For some of these, there are numbers on their frequency in cohorts of patients with congenital disorders, e.g. mosaicism for CNVs in 0.5-2% of the patients[19, 88] and UPD in <1% of the patients[19]. A recently described mechanism for disease is structural variants that disrupt certain regions called topologically associated domains (TADs). These domains can be regarded as regulatory units within which enhancers and promoters can interact. Disruption of a TAD can lead to altered gene-expression and thereby cause disease.[89]

Development of methods, including bioinformatic methods, to detect all of these variants, interpret them (see below) and determine their relative contribution to the etiology of rare disorders are ongoing and will be an important future field of research.

Improved interpretation of genetic variants

Another future challenge, in addition to detecting all types of genetic variants, is interpreting these variants and establishing a causal relation to a specific disease phenotype. This applies to both research and clinical settings. Major challenges concern interpretation of variants in non-coding regions of the genome including variants that affect genomic structure and transcription as well as understanding the interaction between several co-occurring variants and their common contribution to a specific phenotype.

Also for variants that intuitively are easier to grasp, like those within genes, there are challenges in prioritization among a number of variants and interpretation. Further development and use of databases may facilitate this process and improve the outcome. Some of the existing databases have a function of finding other patients based on phenotypic and/or genotypic similarities and this may become increasingly important for interpreting variants and understanding rare disorders. Furthermore, "phenotype-programs" are emerging, which use input regarding the patient's phenotype, either for prioritization of variants detected by MPS or for assessment of a candidate gene's relevance to the phenotype observed in a patient.[90]

Genetic screening for couples at increased risk of having a child with an AR disorder

Pre-conception carrier screening by MPS for consanguineous couples is in theory a good method for identifying common recessive disease alleles. This would enable individualized genetic counseling and more accurate figures regarding the risk of having a child with an AR disorder as well as enable prenatal diagnostics or preimplantation genetic diagnosis (PGD) for couples at high (25%) risk. However, as of today there may be difficulties in applying this method in a clinical setting, much due to insufficient knowledge about the full spectrum of AR disorders and interpretation of genetic variants. These issues may lead to false negative results and, worse, false positive results if a variant is misinterpreted as pathogenic. As mentioned in previous chapters, seemingly deleterious variants may prove to be benign and variants in disease databases may sometimes be incorrectly annotated as pathogenic.[91] Large studies to address the clinical feasibility of pre-conception carrier screening by MPS has not yet been performed. The current issues of false negative and false positive rates may decrease in the future, opening up for such studies.

Prenatal screening for pathogenic de novo variants

In theory, prenatal screening with MPS in order to detect pathogenic *de novo* variants would enable prevention of many congenital disorders. The issues (ethical issues not included) with this method are similar to those described above with false negative and false positive results. In addition, the time frame is much shorter in the case of prenatal screening. Common to all types of prospective genetic screening is the lack of a phenotype to aid in the interpretation of the genetic variants detected. Despite these issues, studies to explore the opportunities with this type of screening may be undertaken in the future.

Treatment of rare disorders

Only small fractions of the rare disorders are treatable today. Recent years' progress in deciphering the etiology for rare genetic disorders will hopefully be followed by advancements in understanding of the pathophysiology that underlie these disorders and ultimately development of treatments in the future. However, this is a challenging task, not only due to difficulties in understanding the biology, but also due to issues regarding financing and clinical trials, which in turn is a consequence of the rarity of these disorders.

Future perspectives and reflections on "genotype versus phenotype"

Medical research is often focused on understanding human disease and the development of treatments. In a clinical setting, treatment and care largely depend on the specific disorder/diagnosis. Therefore, questions of relevance for both research and clinical care are how we define a disorder and how we establish a specific diagnosis in a patient (in order to choose the most effective treatment and care, give correct information on prognosis etc.).

Regarding genetic disorders, the answers to these questions are now shifting from "by the phenotype" to "by the genotype". With new technologies, such as MPS, that enables us to define many of these disorders based on the etiology, these questions are more relevant than ever before. The relevance is not restricted to rare disorders, but also applies to what we usually denote as common disorders. Many of the common disorders have traditionally been defined based on their phenotype, e.g. diabetes, cancer and schizophrenia. As of today and for the future, some of these disorders will rather be defined by their etiology. Data from genome wide screening studies has proved that many common disorders have a heterogeneous etiology and are actually collections of rare disorders. What is learned from studies of rare disorders may thus be applicable also to common disorders. These findings lead to another question of how much breakdown into etiological subgroups that is feasible in a clinical setting and for treatment development. Is definition of a disorder by a specific gene (or genes) enough, or should the exact genetic variant be considered, or even the exact variant in the context of *n* additional factors? At the extreme end, this would mean that each patient has a unique disorder that may require unique care and treatment. This may bring us to a new era of 'personalized medicine' with influences on care and treatments for both rare and common disorders.

What about the phenotype? If diagnosis, care and treatment will be based on the genotype, is there no need to define a phenotype for future patients with genetic disorders? Most likely, clinical assessment of patients will be just as important in the future as it was previously. However, there will probably be (and already is) a shift in the way phenotypic data is used for establishing a diagnosis in patients. Historically, time and money were spent on gathering clinical information that could be used to group patients together, sometimes followed by targeted genetic analyses, in order to establish a diagnosis. As of today and in the future, clinical data may instead be used to facilitate the interpretation of variants generated by genomic screening methods, to achieve a diagnosis. Targeted genetic analyses based on an extensive phenotype would thereby be replaced by targeted clinical investigations based on an extensive genetic analysis.

7 POPULÄRVETENSKAPLIG SAMMANFATTNING

Medfödda sjukdomar, skador eller funktionsnedsättningar drabbar ca 3-4% av alla födda barn. Vissa diagnoser märks redan i fosterlivet medan andra upptäcks först efter födslen eller längre fram. De är ofta förknippade med kroniska symptom och en betydande inverkan på livet för drabbade individer och deras familjer. Vanligt förekommande symptom är olika typer av missbildningar och/eller utvecklingsstörning. Detta förekommer dock inte hos alla, utan vissa individer kan ha andra typer av symptom såsom allvarlig muskelsvaghet, blindhet, dövhet etc. De flesta av diagnoserna är sällsynta (<1/2000) medan andra, som till exempel Down syndrom, är vanligare. Man har uppskattat att det finns närmare 10.000 sällsynta diagnoser, vilket innebär att varje enskild diagnos drabbar få individer, ibland bara en handfull personer i hela världen. Kunskapen är därför begränsad vad gäller många av dessa diagnoser. Orsaken är ofta genetisk, men även andra faktorer såsom infektioner, läkemedel och skadliga ämnen under eller efter fosterlivet samt autoimmunitet kan ge upphov till medfödda eller tidigt debuterande sjukdomar.

Forskningen kring dessa diagnoser och deras orsaker har avancerat snabbt under de senaste tio åren. Vad gäller utvecklingsstörning till exempel har det visats att upp till 70% av patienterna har detta på grund av en genetisk orsak. Det finns olika typer av genetiska förändringar som kan orsaka sjukdom och i fallen med utvecklingsstörning sågs hos ca 10% en kromosom-avvikelse (tex en extra kromosom 21 som vid Down syndrom), hos ca 20% små avvikelser i mängden arvsmassa och hos ca 40% enstaka "stavfel" i den genetiska koden. De sistnämnda två benämns kopietals-varianter (copy number variants, CNVs) respektive enbaspars-varianter (single nucleotide variants, SNVs). Vidare har man sett att hos de flesta individerna hade den genetiska förändringen uppstått *de novo,* vilket innebär att den inte var nedärvd från föräldrarna utan sågs bara hos individen själv.

Många av de medfödda sjukdomarna beror dock på genetiska förändringar som ärvs från båda föräldrarna via så kallad autosomalt recessiv (AR) nedärvning. Dessa sjukdomar förekommer över hela världen och kan drabba alla, men är ofta väldigt sällsynta, vilket kan förklara varför man inte fann så många AR sjukdomar i gruppen med utvecklingsstörning. Om man fått ett barn med en AR sjukdom är risken 25% att eventuella framtida syskon drabbas av samma sjukdom. Fullständig kunskap om dessa sjukdomar saknas, men är viktig ur flera aspekter - inte minst för att kunna fastställa en korrekt diagnos och därmed erbjuda rätt information och vård till drabbade individer och deras föräldrar samt för att kunna erbjuda fosterdiagnostik eller andra alternativ för de som så önskar. AR sjukdomar är mer frekventa i vissa regioner, till exempel regioner som varit geografiskt isolerade, eller då det finns släktskap mellan föräldrarna, men även i dessa fall är förekomsten låg (ca 2-4%).

Syftet med studierna i avhandlingen har varit att undersöka och fastställa den genetiska bakgrunden till olika sällsynta sjukdomar med fokus på AR sjukdomar samt att undersöka och beskriva den kliniska bilden vid dessa sjukdomar.

I studierna har totalt 21 familjer med barn som drabbats av en medfödd sjukdom genomgått undersökning. Majoriteten av familjerna (20 st) inkluderades på basen av en mycket stark misstanke om AR sjukdom. Denna misstanke grundades på förekomst av minst två barn med samma symptombild i familjen samt släktskap mellan föräldrarna. Dessutom studerades ytterligare en familj med en sällsynt *de novo* CNV.

De drabbade individerna har dels genomgått en noggrann klinisk undersökning för att fastställa vilka symptom de har samt dels flera genetiska undersökningar och ibland ytterligare uppföljande undersökningar. De genetiska undersökningar har inneburit att arvsmassan har analyserats för att identifiera genetiska förändringar av de typer som nämnts ovan (kromosomavvikelser, CNVs och SNVs). Inom ramen för sjukvården har sk array-CGH utförts, vilken kan detektera de två förstnämnda typerna av förändringar. Därefter har komplettering skett med helexomsekvensering, vilket innebär att DNA-sekvensen i alla gener "läses av" och olika "stavfel" kan detekteras. Vid båda metoderna jämförs arvsmassan med ett "facit" från en eller några friska individer. Arvsmassan är till 98-99% identisk om man jämför två individer, medan 1-2% av arvsmassan naturligt skiljer sig åt mellan individerna. En frisk individ har ca 30000 "stavfel" (SNVs) i sina gener och totalt ca 1000 CNVs om man jämför med ett "facit" från en annan frisk individ, vilket innebär att de förändringar som ses hos en drabbad individ inte alls behöver vara orsaken till sjukdomen utan bara en normalvariant som inte finns hos just den eller de individer som "facit" baseras på. Att skilja mellan genetiska normalvarianter och sjukdomsorsakande varianter är en av de största utmaningarna inom både genetik-forskning och sjukvård. I studierna har detta gjorts genom bland annat användande av databaser för att filtrera bort kända normalvarianter, jämförelse av arvsmassan från andra familjemedlemmar (framförallt syskon med eller utan sjukdom) samt jämförelse av den kliniska bilden mot vad som finns känt sedan tidigare om den gen som kan vara skadad. I utvalda fall har kompletterande undersökningar bidragit till att fastställa om förändringen varit orsaken till sjukdom eller ej.

Resultaten av dessa undersökningar ledde till att man i 15 av de 21 familjerna kunde fastställa sjukdomsorsaken och rapportera detta tillbaka till familjen och sjukvården. I endast tre av familjerna var den påvisade sjukdomen en relativt välbeskriven sjukdom med över femtio olika fall beskrivna i världen. Resten av familjerna visade sig vara drabbade av mycket sällsynta sjukdomar. Ytterligheten var en familj med flera barn drabbade av svår utvecklingsförsening och epilepsi där en genetisk variant i genen *PIGT* detekterades. Att denna variant var sjukdomsorsakande kunde bekräftas med en rad olika analyser. Dessa fall var de första beskrivna i världen med "*PIGT*-sjukdom". Idag finns ytterligare ett par familjer rapporterade med samma sjukdom som nu kallas "multiple congenital anomalies-hypotonia-seizures syndrome-3 (MCAHS3)". I två andra studie-familjer fastställdes diagnoser som endast beskrivits en gång tidigare. Med resultaten från denna studie kan orsaken till dessa två allvarliga sjukdomar bekräftas samt den kliniska bilden ytterligare beskrivas. Beskrivning av den kliniska bilden är värdefull i de flesta fall av dessa sällsynta sjukdomar. För de flesta av studie-patienterna stämde den kliniska bilden väl med de få fall som beskrivits tidigare, medan individer från tre av familjerna uppvisade en “icke-klassisk” symptombild. Det mest påtagliga fallet rörde en familj med två syskon som hade svår utvecklingsstörning samt stora avvikelser på hjärnröntgen där man bland annat såg kraftigt vidgade hålrum. Diagnosen som fastställdes, Fowler syndrom, hade tidigare beskrivits endast hos foster och ansetts icke-förenlig med överlevnad. Trots sin svåra sjukdom är äldsta syskonet nu sju år gammal. Slutligen, för den studie-familj som inte hade en AR sjukdom utan istället en sällsynt *de novo* CNV visade resultaten att det rörde sig om en dubblering av arvsmassa inom genen *SATB2*. Denna gen är kopplad till en specifik diagnos med bland annat gomspalt och utvecklingsstörning. Nytt i detta fall var att den här typen av förändring inte rapporterats bland något av de få tidigare fallen (som istället hade små “stavfel” eller förlust av arvsmassa inom samma gen).

De övergripande slutsatserna från studierna är att helexomsekvensering är en mycket effektiv metod för identifiering av sjukdomsorsakande genetiska varianter i familjer med stark misstanke om AR sjukdom samt att en noggrann klinisk undersökning i många fall är avgörande för att kunna tolka resultaten från de genetiska undersökningarna. En ytterligare slutsats är att analys av samtliga typer av genetiska varianter, inklusive SNVs och CNVs, bör ingå vid diagnostik av sällsynta sjukdomar.

Resultaten från denna studie har breddat kunskapen om den kliniska bilden vid de diagnoser som beskrivits samt utökat antalet gener och genetiska varianter som har betydelse för uppkomsten av medfödda sjukdomar. Förhoppningsvis kan detta bidra till att öka chansen för framtida patienter att erhålla en diagnos och därmed förbättra deras och familjernas vård och omhändertagande.

8 REFERENCES

1. *Programme of community action on rare diseases (1999 – 2003) Decision No 1295/99/EC of the European Parliament and of the Council of 29 April 1999* Available from: http://ec.europa.eu/health/archive/ph_overview/previous_programme/rare_diseases/raredis_wpgm99_en.pdf.

2. *Rare Diseases Act of 2002*. Available from: https://history.nih.gov/research/downloads/PL107-280.pdf.

3. de Vrueh, R., E.R.F. Baekelandt, and J.M.H. de Haan. *Priority Medicines for Europe and the World "A Public Health Approach to Innovation", Update on 2004 Background Paper, BP 6.19 Rare Diseases*. 2013; Available from: http://www.who.int/medicines/areas/priority_medicines/Ch6_19Rare.pdf.

4. *What is a rare disease?* 2014; Available from: http://www.eurordis.org/about-rare-diseases.

5. *Online Mendelian Inheritance in Man OMIM®*. 2015; Available from: http://omim.org/.

6. *About Rare Diseases*. 2012; Available from: http://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=EN.

7. Olry, A. and A. Rath. *Prevalence of rare diseases: Bibliographic data, Orphanet report series, Rare diseases collection*. 2015; Available from: http://www.orpha.net/orphacom/cahiers/docs/GB/Prevalence_of_rare_diseases_by_alphabetical_list.pdf.

8. *Statistics and Figures on Prevalence of Genetic and Rare Disease*. Available from: https://globalgenes.org/rare-diseases-facts-statistics/.

9. Norio, R., *The Finnish Disease Heritage III: the individual diseases.* Hum Genet, 2003. **112**(5-6): p. 470-526.

10. Muensterer, O.J., et al., *Ellis-van Creveld syndrome: its history.* Pediatr Radiol, 2013. **43**(8): p. 1030-6.

11. Modell, B. and M. Darlison, *Global epidemiology of haemoglobin disorders and derived service indicators.* Bull World Health Organ, 2008. **86**(6): p. 480-7.

12. Antonarakis, S.E. and V.A. McKusick, *OMIM passes the 1,000-disease-gene mark.* Nat Genet, 2000. **25**(1): p. 11.

13. Peltonen, L. and V.A. McKusick, *Genomics and medicine. Dissecting human disease in the postgenomic era.* Science, 2001. **291**(5507): p. 1224-9.

14. McKusick, V.A., *Mendelian Inheritance in Man and its online version, OMIM.* Am J Hum Genet, 2007. **80**(4): p. 588-604.

15. Amberger, J.S., et al., *OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders.* Nucleic Acids Res, 2015. **43**(Database issue): p. D789-98.

16. Nouspikel, T., et al., *A common mutational pattern in Cockayne syndrome patients from xeroderma pigmentosum group G: implications for a second XPG function.* Proc Natl Acad Sci U S A, 1997. **94**(7): p. 3116-21.

17. Drury, S., et al., *A novel homozygous ERCC5 truncating mutation in a family with prenatal arthrogryposis--further evidence of genotype-phenotype correlation.* Am J Med Genet A, 2014. **164A**(7): p. 1777-83.

18. Rauch, A., et al., *Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study.* Lancet, 2012. **380**(9854): p. 1674-82.

19. Deciphering Developmental Disorders, S., *Large-scale discovery of novel genetic causes of developmental disorders.* Nature, 2015. **519**(7542): p. 223-8.

20. de Ligt, J., et al., *Diagnostic exome sequencing in persons with severe intellectual disability.* N Engl J Med, 2012. **367**(20): p. 1921-9.

21. Gilissen, C., et al., *Genome sequencing identifies major causes of severe intellectual disability.* Nature, 2014. **511**(7509): p. 344-7.

22. Vissers, L.E., C. Gilissen, and J.A. Veltman, *Genetic studies in intellectual disability and related disorders.* Nat Rev Genet, 2016. **17**(1): p. 9-18.

23. Stevenson, R.E., et al., *Genetic syndromes among individuals with mental retardation.* Am J Med Genet A, 2003. **123A**(1): p. 29-32.

24. Richards, S., et al., *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.* Genet Med, 2015. **17**(5): p. 405-24.

25. Stenson, P.D., et al., *The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine.* Hum Genet, 2014. **133**(1): p. 1-9.

26. MacDonald, J.R., et al., *The Database of Genomic Variants: a curated collection of structural variation in the human genome.* Nucleic Acids Res, 2014. **42**(Database issue): p. D986-92.

27. Talkowski, M.E., et al., *Disruption of a large intergenic noncoding RNA in subjects with neurodevelopmental disabilities.* Am J Hum Genet, 2012. **91**(6): p. 1128-34.

28. Rainger, J.K., et al., *Disruption of SATB2 or its long-range cis-regulation by SOX9 causes a syndromic form of Pierre Robin sequence.* Hum Mol Genet, 2014. **23**(10): p. 2569-79.

29. Zhang, F., et al., *Copy number variation in human health, disease, and evolution.* Annu Rev Genomics Hum Genet, 2009. **10**: p. 451-81.

30. Heyer, W.D., K.T. Ehmsen, and J. Liu, *Regulation of homologous recombination in eukaryotes.* Annu Rev Genet, 2010. **44**: p. 113-39.

31. Lupski, J.R. and P. Stankiewicz, *Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes.* PLoS Genet, 2005. **1**(6): p. e49.

32. Lieber, M.R., *The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway.* Annu Rev Biochem, 2010. **79**: p. 181-211.

33. Ottaviani, D., M. LeCain, and D. Sheer, *The role of microhomology in genomic structural variation.* Trends Genet, 2014. **30**(3): p. 85-94.

34. Lee, J.A., C.M. Carvalho, and J.R. Lupski, *A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders.* Cell, 2007. **131**(7): p. 1235-47.

35. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nat Rev Genet, 2006. **7**(2): p. 85-97.

36. Lee, C. and S.W. Scherer, *The clinical context of copy number variation in the human genome.* Expert Rev Mol Med, 2010. **12**: p. e8.

37. Wheeler, D.A., et al., *The complete genome of an individual by massively parallel DNA sequencing.* Nature, 2008. **452**(7189): p. 872-6.

38. Pang, A.W., et al., *Towards a comprehensive structural variation map of an individual human genome.* Genome Biol, 2010. **11**(5): p. R52.

39. Levy, S., et al., *The diploid genome sequence of an individual human.* PLoS Biol, 2007. **5**(10): p. e254.

40. Conrad, D.F., et al., *Origins and functional impact of copy number variation in the human genome.* Nature, 2010. **464**(7289): p. 704-12.

41. MacArthur, D.G., et al., *A systematic survey of loss-of-function variants in human protein-coding genes.* Science, 2012. **335**(6070): p. 823-8.

42. Kaiser, V.B., et al., *Homozygous loss-of-function variants in European cosmopolitan and isolate populations.* Hum Mol Genet, 2015. **24**(19): p. 5464-74.

43. Huang, N., et al., *Characterising and predicting haploinsufficiency in the human genome.* PLoS Genet, 2010. **6**(10): p. e1001154.

44. Besenbacher, S., et al., *Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios.* Nat Commun, 2015. **6**: p. 5969.

45. Kong, A., et al., *Rate of de novo mutations and the importance of father's age to disease risk.* Nature, 2012. **488**(7412): p. 471-5.

46. Itsara, A., et al., *De novo rates and selection of large copy number variation.* Genome Res, 2010. **20**(11): p. 1469-81.

47. Bittles, A.H. and M.L. Black, *The impact of consanguinity on neonatal and infant health.* Early Hum Dev, 2010. **86**(11): p. 737-41.

48. Hamamy, H., et al., *Consanguineous marriages, pearls and perils: Geneva International Consanguinity Workshop Report.* Genetics in medicine : official journal of the American College of Medical Genetics, 2011. **13**(9): p. 841-847.

49. Zlotogora, J. and S.A. Shalev, *The consequences of consanguinity on the rates of malformations and major medical conditions at birth and in early childhood in inbred populations.* American journal of medical genetics. Part A, 2010. **152A**(8): p. 2023-8.

50. Powis, Z., et al., *Diagnostic exome sequencing for patients with a family history of consanguinity: over 38% of positive results are not autosomal recessive pattern.* J Hum Genet, 2015.

51. Xue, Y., et al., *Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing.* Am J Hum Genet, 2012. **91**(6): p. 1022-32.

52. Bell, C.J., et al., *Carrier testing for severe childhood recessive diseases by next-generation sequencing.* Sci Transl Med, 2011. **3**(65): p. 65ra4.

53. Morton, N.E., J.F. Crow, and H.J. Muller, *An Estimate of the Mutational Damage in Man from Data on Consanguineous Marriages.* Proc Natl Acad Sci U S A, 1956. **42**(11): p. 855-63.

54. Boone, P.M., et al., *Deletions of recessive disease genes: CNV contribution to carrier states and disease-causing alleles.* Genome Res, 2013. **23**(9): p. 1383-94.

55. Najmabadi, H., et al., *Deep sequencing reveals 50 novel genes for recessive cognitive disorders.* Nature, 2011. **478**(7367): p. 57-63.

56. Schuurs-Hoeijmakers, J.H., et al., *Identification of pathogenic gene variants in small families with intellectually disabled siblings by exome sequencing.* J Med Genet, 2013. **50**(12): p. 802-11.

57. Yavarna, T., et al., *High diagnostic yield of clinical exome sequencing in Middle Eastern patients with Mendelian disorders.* Hum Genet, 2015. **134**(9): p. 967-80.

58. Shamseldin, H.E., et al., *Identification of embryonic lethal genes in humans by autozygosity mapping and exome sequencing in consanguineous families.* Genome Biol, 2015. **16**: p. 116.

59. Shaheen, R., et al., *Accelerating matchmaking of novel dysmorphology syndromes through clinical and genomic characterization of a large cohort.* Genet Med, 2015.

60. Alazami, A.M., et al., *Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families.* Cell Rep, 2015. **10**(2): p. 148-61.

61. Makrythanasis, P., et al., *Diagnostic exome sequencing to elucidate the genetic basis of likely recessive disorders in consanguineous families.* Hum Mutat, 2014. **35**(10): p. 1203-10.

62. Fahiminiya, S., et al., *Whole exome sequencing unravels disease-causing genes in consanguineous families in Qatar.* Clin Genet, 2014. **86**(2): p. 134-41.

63. Schouten, J.P., et al., *Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification.* Nucleic Acids Res, 2002. **30**(12): p. e57.

64. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.

65. *Sequencing by Synthesis (SBS) Technology*. 2015; Available from: http://www.illumina.com/technology/next-generation-sequencing.html.

66. MacArthur, D.G., et al., *Guidelines for investigating causality of sequence variants in human disease.* Nature, 2014. **508**(7497): p. 469-76.

67. Firth, H.V., et al., *DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources.* Am J Hum Genet, 2009. **84**(4): p. 524-33.

68. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-81.

69. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.

70. Exome-Aggregation-Consortium, *Analysis of protein-coding genetic variation in 60,706 humans.* bioRxiv preprint, 2015.

71. Kohler, S., et al., *The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data.* Nucleic Acids Res, 2014. **42**(Database issue): p. D966-74.

72. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants.* Nat Genet, 2014. **46**(3): p. 310-5.

73. Newman, S., et al., *Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints.* Am J Hum Genet, 2015. **96**(2): p. 208-20.

74. Vissers, L.E., et al., *Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture.* Hum Mol Genet, 2009. **18**(19): p. 3579-93.

75. Mills, R.E., et al., *Mapping copy number variation by population-scale genome sequencing.* Nature, 2011. **470**(7332): p. 59-65.

76. Leoyklang, P., et al., *Heterozygous nonsense mutation SATB2 associated with cleft palate, osteoporosis, and cognitive defects.* Hum Mutat, 2007. **28**(7): p. 732-8.

77. Rosenfeld, J.A., et al., *Small deletions of SATB2 cause some of the clinical features of the 2q33.1 microdeletion syndrome.* PLoS One, 2009. **4**(8): p. e6568.

78. Balasubramanian, M., et al., *Case series: 2q33.1 microdeletion syndrome--further delineation of the phenotype.* J Med Genet, 2011. **48**(5): p. 290-8.

79. Docker, D., et al., *Further delineation of the SATB2 phenotype.* European journal of human genetics : EJHG, 2013.

80. Asadollahi, R., et al., *The clinical significance of small copy number variants in neurodevelopmental disorders.* J Med Genet, 2014. **51**(10): p. 677-88.

81. Kaiser, A.S., et al., *Characterization of the first intragenic SATB2 duplication in a girl with intellectual disability, nearly absent speech and suspected hypodontia.* Eur J Hum Genet, 2015. **23**(5): p. 704-7.

82. Zarate, Y.A., et al., *Further supporting evidence for the SATB2-associated syndrome found through whole exome sequencing.* Am J Med Genet A, 2015. **167A**(5): p. 1026-32.

83. Beetz, C., et al., *Inhibition of TFG function causes hereditary axon degeneration by impairing endoplasmic reticulum structure.* Proc Natl Acad Sci U S A, 2013. **110**(13): p. 5091-6.

84. Najmabadi, H., et al., *Deep sequencing reveals 50 novel genes for recessive cognitive disorders.* Nature, 2011.

85. Nakashima, M., et al., *Novel compound heterozygous PIGT mutations caused multiple congenital anomalies-hypotonia-seizures syndrome 3.* Neurogenetics, 2014. **15**(3): p. 193-200.

86. Lam, C., et al., *Expanding the clinical and molecular characteristics of PIGT-CDG, a disorder of glycosylphosphatidylinositol anchors.* Mol Genet Metab, 2015. **115**(2-3): p. 128-40.

87. Thomas, S., et al., *High-throughput sequencing of a 4.1 Mb linkage interval reveals FLVCR2 deletions and mutations in lethal cerebral vasculopathy.* Hum Mutat, 2010. **31**(10): p. 1134-41.

88. Biesecker, L.G. and N.B. Spinner, *A genomic view of mosaicism and human disease.* Nat Rev Genet, 2013. **14**(5): p. 307-20.

89. Lupianez, D.G., et al., *Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.* Cell, 2015. **161**(5): p. 1012-25.

90. Smedley, D. and P.N. Robinson, *Phenotype-driven strategies for exome prioritization of human Mendelian disease genes.* Genome Med, 2015. **7**(1): p. 81.

91. Wang, J. and Y. Shen, *When a "disease-causing mutation" is not a pathogenic variant.* Clin Chem, 2014. **60**(5): p. 711-3.