UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1078

Genetics of Two Mendelian Traits and Validation of Induced Pluripotent Stem Cell (iPSC) Technology for Disease Modeling

DOROTEYA RAYKOVA

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2015

ISSN 1651-6206
ISBN 978-91-554-9184-0
urn:nbn:se:uu:diva-246228

Dissertation presented at Uppsala University to be publicly examined in Fåhraeussalen, Rudbeck Laboratoriet, Dag Hammarsjöldsväg 20, Uppsala, Friday, 24 April 2015 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English. Faculty examiner: Ann Nordgren (Karolinska Institutet, Department of Molecular Medicine and Surgery).

Abstract

Raykova, D. 2015. Genetics of Two Mendelian Traits and Validation of Induced Pluripotent Stem Cell (iPSC) Technology for Disease Modeling. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 1078. 54 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-554-9184-0.

Novel technologies for genome analysis have provided almost unlimited opportunities to uncover structural gene variants behind human disorders. Whole exome sequencing (WES) is especially useful for understanding rare Mendelian conditions, because it reduces the requirements for *a priori* clinical data, and can be applied on a small number of patients. However, supporting functional data on the effect of specific gene variants are often required to power these findings. A variety of methods and biological model systems exists for this purpose. Among those, induced pluripotent stem cells (iPSCs), which are capable of self-renewal and differentiation, stand out as an alternative to animal models.

In papers I and II we took advantage of WES to identify gene variants underlying autosomal recessive pure hair and nail ectodermal dysplasia (AR PHNED) as well as autosomal dominant familial visceral myopathy (FVM). We identified a homozygous variant c.821T>C (p.Phe274Ser) in the *KRT74* gene as the causative mutation in AR PHNED, supported by the fact that Keratin-74 was undetectable in hair follicles of an affected family member. In a family segregating FVM we found a heterozygous tandem base substitution c.806_807delinsAA (p.(Gly269Glu)) in the *ACTG2* gene in the affected members. This novel variant is associated with a broad range of visceral symptoms and a variable age of onset.

In Paper III we explored the similarity between clonally derived iPSC lines originating from a single parental fibroblast line and we highlighted the necessity to use lines originating from various donors in disease modeling because of biological variation. Paper IV focused on how the genomic integrity of iPSCs is affected by the choice of reprogramming methods. We described several novel cytogenetic rearrangements in iPSCs and we identified a chromosome 5q duplication as a candidate aberration for growth advantage.

In summary, this doctoral thesis brings novel findings on unreported disease-causing variants, as supported by extensive genetic analysis and functional data. A novel molecular mechanism behind AR PHNED is presented and the phenotypic spectrum associated with FVM is expanded. In addition, the thesis brings novel understanding of benefits and limitations of the iPSC technology to be considered for disease modeling.

Keywords: Disease modeling, Mendelian disorders, iPSC, Whole exome sequencing, Transcriptome sequencing

Doroteya Raykova, Department of Immunology, Genetics and Pathology, Medicinsk genetik och genomik, Rudbecklaboratoriet, Uppsala University, SE-751 85 Uppsala, Sweden.

ISSN 1651-6206
ISBN 978-91-554-9184-0
urn:nbn:se:uu:diva-246228 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-246228)

“Madness, in a higher sense, is the beginning of all wisdom”

— Hermann Hesse

To my family

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **Raykova D.**, Klar J., Azhar A., Khan T.N., Malik N.A., Iqbal M., Tariq M., Baig S.M., Dahl N. (2014) Autosomal Recessive Transmission of a Rare KRT74 Variant Causes Hair and Nail Ectodermal Dysplasia: Allelism with Dominant Woolly Hair/Hypotrichosis. *PLoS One* **9**(4):e93607

II Klar J.[*], **Raykova D.**[*], Gustafson E., Tóthová I., Ameur A., Wanders A., Dahl N. (2015) Phenotypic expansion of visceral myopathy associated with *ACTG2* tandem base substitution. *Eur J Hum Genet, in press*

III Schuster J., Halvardson J., Lorenzo L., Ameur A., Sobol M., **Raykova D.**, Annerén G., Feuk L., Dahl N. (2015) Transcriptome profiling reveals degree of variability in iPSC lines: Impact for human disease modeling. *Cell Reprogram, submitted*

IV Sobol M., **Raykova D.**, Cavelier L., Khalfallah A., Schuster J., Dahl N. (2015) Methods of reprogramming to iPSC associated with chromosomal integrity and delineation of a chromosome 5q candidate region for growth advantage. *Stem Cell Dev, provisionally accepted*

[*] Equal first authors

Contents

Abbreviations

3D	Three-dimensional
A	Adenine
ACTG2	Smooth muscle actin gamma-2
AD WH	Autosomal dominant wooly hair and/or hypotrichosis simplex
AR PHNED	Autosomal recessive pure hair and nail ectodermal dysplasia
bFGF	Basic fibroblast growth factor
bp	Basepair
C	Cytosine
cDNA	Complementary DNA
CGH	Comparative genomic hybridization
CNV	Copy number variation
Contig	Contiguous sequence
COSMIC	Catalogue of Somatic Mutation in Cancer
CRISPR	Clustered regularly interspaced short palindromic repeats
Cys	Cysteine
dbSNP	Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic acid
EB	Embryoid body
EdU	5-ethynyl-2'-deoxyuridine
ENCODE	Encyclopedia of DNA Elements
ESC	Embryonic stem cell
EVS	Exome Variant Server
FVM	Familial visceral myopathy
G	Guanine
Glu	Glutamic acid
Gly	Glycine
GWAS	Genome-wide association studies
hESC	Human embryonic stem cell
HGMD	Human Gene Mutation Database
HGP	Human Genome Project
HGPS	Hutchinson-Gilford progeria syndrome
Indel	Insertion or deletion
iPSC	Induced pluripotent stem cell
IVF	*In vitro* fertilization
KEGG	Kyoto Encyclopedia of Genes and Genomes
KRT74	Keratin-74
Leu	Leucine

LIF	Leukemia inhibitory factor
LINE	Long interspersed nuclear element
LOD	Logarithm of the odds
MAF	Minor allele frequency
miRNA	MicroRNA
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
OKSM	Oct4, Klf4, Sox2 and c-Myc
P	Passage
PCR	Polymerase chain reaction
PFA	Paraformaldehyde
Phe	Phenylalanine
Pro	Proline
qRT-PCR	Quantitative reverse-transcription PCR
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
RT-PCR	Reverse transcription PCR
Ser	Serine
SINE	Short interspersed nuclear element
snoRNA	Small nucleolar RNA
SNP	Single nucleotide polymorphism
STRP	Short tandem repeat polymorphism
T	Thymine
TALEN	Transcription activator-like effector nuclease
TBS	Tandem base substitution
TLDA	TaqMan Low Density Array
Tyr	Tyrosine
UPD	Uniparental disomy
WES	Whole exome sequencing
WGS	Whole genome sequencing
XCI	X chromosome inactivation

Introduction

"...the endless repetition of an ordinary miracle."
— Orhan Pamuk

Although the foundations of genetics were laid as early as the mid-nineteenth century by the pioneering experiments of Gregor Mendel, little was known and understood about the actual mechanisms of inheritance for decades on. While the basic laws were postulated, scientists had no knowledge of the molecular carrier of genetic information until 1953, when James Watson and Francis Crick discovered the structure of deoxyribonucleic acid (DNA) and hypothesized that it was the blueprint for heredity [1].

The DNA molecule is a long, double-stranded, antiparallel polymer consisting of four building blocks termed nucleotides. Each nucleotide consists of a nitrogenous base – guanine (G), adenine (A), thymine (T) or cytosine (C), – a pentose sugar (deoxyribose) and a phosphate group. G and A are termed purines, whereas T and C are called pyrimidines. The sugar/phosphate molecules form the DNA backbone, from which the nucleotides extend in a perpendicular plane. They form specific basepairs (bp) between each other, and A always couples with T via a double hydrogen bond, whereas T and C connect more tightly, forming three hydrogen bonds. The order of nucleotides in the DNA strand determines the genetic code, which in essence is the "recipe" of how the cell should make proteins.

After the discovery of DNA structure and function as the bearer of genetic information, the scientific world made an effort to "read" and understand the message encoded in it. By the 1970s, different methods were developed to determine the order of nucleotides in DNA, a procedure nowadays known as sequencing. The biggest attempt, however, was commenced in 1990 with the Human Genome Project (HGP) which had the ambition to decipher the entire human genome. The HGP was completed in 2004, uncovering that the human genome is made up of over three billion basepairs and contains approximately 20 000 protein-coding genes, along with other non-coding RNA genes, regulatory sequences, sequences of yet unassigned function, etc. [2]. Recently, 80% of the genome was linked to particular biochemical functions by the Encyclopedia of DNA Elements (ENCODE), but much remains to be elucidated [3].

Variation in the human genome

"What makes the marvelous is its peculiar way of being ordinary; what makes the ordinary is its peculiar way of being marvelous."
— Orhan Pamuk

Types of variation

Although the basic principles governing the development of organisms of the same species are essentially identical, the subtle details in our genetic blueprints contribute to the remarkable phenotypic diversity that is observed in nature [3-7]. These differences are known as genetic variation. Variation is the result of alterations (mutations) that can occur both in the coding and the non-coding elements of the genome [8, 9]. The majority of these alterations either have no effect or can produce a range of normal phenotypic variability. Occasionally however, alterations are pathogenic and become a cause of disease or increased susceptibility to disease, often in combination with other factors. Genetic changes can occur in the genome for different reasons, but are commonly due to faults in the cellular DNA repair mechanisms [10]. The effects of these variants can be neutral, beneficial or deleterious, depending on the type and location of occurrence.

Mutations, or lesions, may be small-scale, such as those that affect one or a few nucleotides (point mutations, tandem base substitutions), or large-scale, affecting bigger DNA segments (deletions, duplications, inversions, translocations).

Point mutations are lesions in which a single base has been changed for another. They are classified as transitions when a purine base is substituted with another purine (i.e. A to G or vice versa) or a pyrimidine is substituted by another pyrimidine (T to C and vice versa). Point mutations are called transversions when a purine is substituted by a pyrimidine base or vice versa (e.g. A to T or C to G). Single base substitutions, insertions or deletions give rise to common variants (frequency of 1% or above) that account for normal genetic variation and are known as single nucleotide polymorphisms (SNPs) [11, 12]. A broad collection for general genetic variation describing identified SNPs can be accessed at NCBI's dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi). These encompass silent variants that do not lead to a change of amino acid, missense variants leading to an alteration of the amino acid, and nonsense mutations

which introduce a stop codon, leading to a truncation of the protein product, as well as frameshift and splice site variants which alter the protein reading frame. SNPs with a minor allele frequency lower than 1% are considered rare and may have deleterious phenotypic effects. Those that are associated with human disease can be found at the Human Gene Mutation Database (HGMD, http://www.hgmd.org) [13].

Tandem base substitutions (TBSs) are similar to single base substitutions, but a larger number of nucleotides are affected – commonly, between two and eight. Chen et al. define TBSs as "multiple mutations that comprise two or more contiguous nucleotide substitutions without any net gain or loss of bases" [14]. This group of variants is often erroneously classified as small indels, but their role in human disease is being increasingly recognized and, according to recent data, ~0.8% of all nucleotide substitutions which occur *de novo* can be attributed to double TBS mutations [15].

Another common source of small-scale variation is repeat sequences, which are abundant in the human genome. Examples include interspersed repeats, mostly derived from transposable elements, such as long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs), as well as tandem repeats such as microsatellites (2-5 bp), minisatellites (10-50 bp) and satellites (larger than 50 bp) [16-18]. Short tandem repeat polymorphisms (STRPs), or polymorphic microsatellites, are widely used in methods for gene mapping and in forensic genetics as genetic markers. Repeated sequences are hotspots for homologous recombination and have been identified as a frequent cause of deletions/ duplications, which result from mispairing between the repeats [19]. They also play a role in the occurrence of unequal crossover events, which can give rise to copy number variations, gene duplications and other rearrangements [19].

Whenever the altered DNA segment is larger than 1 kb, one speaks of structural variations [20]. These include copy number variations (CNVs), inversions, translocations and uniparental disomies. CNVs, as their name suggests, are DNA segments most commonly between 1 and 10 kb in size that come in a different number of copies in the genome [21, 22]. Duplications, insertions and deletions fall into this category. An extensive catalogue of CNVs can be found at the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home). Inversions are segments of DNA with reverse orientation with respect to the rest of the chromosome, and translocations involve a change of the position of the segment between or within chromosomes. Deletions are aberrations associated with loss of genetic material. Insertions and deletions are collectively termed as indels [23, 24]. Uniparental disomies (UPDs) occur as a result of errors in meiosis, when the offspring receives a segment of a chromosome or a pair of homologous chromosomes from the same parent [25].

Variation as a cause of disease

While human genetic variation is the basis of evolution and an intrinsic property of the genomes of healthy individuals, it can also be deleterious. Certain variants disrupt normal molecular processes and lead to disease. Defects in multiple genes, often in combination with environmental factors, cause complex (multifactorial) disorders, such as coronary heart disease, schizophrenia and certain types of cancer [26]. Multifactorial disorders do not follow obvious inheritance patterns, in contrast to Mendelian disorders.

Mendelian disorders, which are the main focus of this thesis, originate from mutations in single genes. Monogenic disorders follow typical Mendelian patterns of inheritance, i.e. follow Mendel's laws of segregation, independent assortment and dominance. Thus, these diseases fall into the following categories: autosomal dominant, autosomal recessive, or sex-linked (also subdivided into dominant or recessive) [27]. Despite being considered rare, Mendelian disorders affect a considerable amount of the population and are a major cause of child mortality [28]. 2-5% of newborns suffer from congenital malformations or genetically-predetermined diseases, which often continue throughout life [28].

Autosomal dominant diseases are usually heterozygous, caused by a pathogenic variant in a single allele of a given gene on the autosomes, which can either be inherited (an example is polycystic kidney disease (MIM 173900) caused by defects in *PKD1* [29-31]) or occur *de novo* (e.g. progeria syndrome (MIM 176670) [32-34]). The expression of a defective, truncated or dysfunctional protein from an affected allele may mediate a dominant effect through haploinsufficiency or through a dominant negative effect whenever the protein normally acts as a di- or polymer [27].

Autosomal recessive disorders are the result of mutations in both alleles of a gene on the autosomes (homozygosity for the pathogenic variant, e.g. in phenylketonuria (MIM 261600) [35, 36]) or of compound heterozygosity (as could be the case with the *TGM1* gene in congenital ichtyosis (MIM 242300) [37, 38]), abolishing the expression of a normal gene product. Rare monogenic disorders with autosomal recessive inheritance patterns are more frequent in consanguineous families due to an increased rate of homozygous variants [27].

Disorders can be caused by mutations in the sex chromosomes, and almost invariably in the X chromosome (X-linked disorders). Only very rarely the Y chromosome is affected, in which case the disease is transmitted from father to son. The latter has been proposed as an explanation for the occurrence of retinitis pigmentosa (MIM 400004) in a Chinese family [39]. X-linked diseases are commonly recessive, and in this case they mostly affect males who carry a single copy of the defective allele (since males normally only have one X-chromosome). Such state is called hemizygous, and an example for it is the inheritance of classic hemophilia (MIM 306 700) [27,

40]. There are also dominant X-linked disorders. One example of these is erythrohepatic protoporphyria (MIM 300752) caused by gain-of-function variants in the *ALAS2* gene [41, 42].

In some X-linked diseases, heterozygous females may not be affected or have milder phenotypes due to the phenomenon of random X chromosome inactivation (XCI) – a normal dosage compensation mechanism in females, which leads to the arbitrary silencing of one X [43]. XCI thus partially compensates for haploinsufficiency caused by X-linked mutations in females. Skewed X inactivation, a process in which the silencing is no longer a random event, but rather one of the X chromosomes is preferentially expressed, may occur in females suffering from X-linked diseases. Skewing may lead to complete silencing of the pathogenic allele; it can have the opposite effect as well: inactivating the wild type X selectively and causing a more severe phenotype [44, 45].

The patterns of transmission typical of monogenic disorders are not necessarily straightforward due to confounding factors such as variability, meaning that patients with the same disorder might exhibit different phenotypes, or disparate degrees of expression of the phenotype [46]. Reduced penetrance is another example: due to reduced penetrance, some patients with a given mutation may exhibit a phenotype, whereas others may not [27]. Modifier genes also play a role in "distorting" the inheritance patterns as they have quantitative effects on other genes' levels of expression (examples in [47, 48]).

Approaches to study the genetic causes of human monogenic disorders

"I read somewhere that luck is not blind, just illiterate. Luck, I mused, is a palliative for those who don't know probability and statistics."
— Orhan Pamuk

From a genetics point of view, an essential first step in understanding the mechanisms of a disease and looking into possible perspectives of finding a cure is to identify the candidate gene (or genes) that cause it. One challenge is to discriminate between normal variation and pathogenic mutations. The golden standard would be to perform functional studies and demonstrate the role of the variant *in vivo*. However, this is often a cumbersome task and may be impossible to perform in large-scale studies. A reasonable alternative is to base the assumption of pathogenicity on whether the variant affects a sequence that is known to be functional (for example the coding parts of genes, splice sites or regulatory elements) which makes the variant a more likely disease-causing candidate. Furthermore, one could look into the type of variant and in what way it affects the DNA sequence, i.e. whether it is an indel, a frameshift or a nonsense mutation disrupting a protein or modifying an exon-intron boundary and interfering with normal splicing. Missense mutations are harder to interpret, but they might affect the level of gene expression or act in a dominant-negative fashion. Two disparate missense mutations may also act in concert, such as in the case of autosomal recessive disorders caused by compound heterozygosity (i.e. each allele may harbor a different alteration, but both are defective).

Since the size of the human genome is above 3 gigabases [49] and mutations can be found in coding as well as non-coding sequences, the task to identify pathogenic variants requires robust methods. Most traditional techniques involve finding the location of the potential causative gene as a first step. Examples of such methods are briefly reviewed below.

Restriction fragment length polymorphism (RFLP) analysis

This is the classical DNA "fingerprinting" technique that was historically among the first methods for detection of disease-causing genes and linkage mapping [50, 51]. It is based on SNPs or indels that alter the recognition site of a specific restriction enzyme and relies on subjecting DNA to restriction digestion and subsequent gel separation of the resulting fragments. These are then hybridized to labeled DNA probes in Southern blotting and the variable fragment lengths are used to discriminate between individuals [50, 51].

Linkage analysis

Linkage analysis uses genetic markers such as microsatellites or SNPs and relies on estimation of recombination frequencies to map disease-causing loci that are co-inherited with (or linked to) the phenotype of interest. It is based on the assumption that markers which are in close proximity to the pathogenic variant will be passed over together, as a haplotype, since the probability for crossing-over in meiosis between closely positioned loci is low [52]. A LOD (logarithm of the odds) score is calculated to compare the probability of inheriting a trait when the causative locus and a particular genetic marker were linked with the probability of this event happening by chance. Traditionally, a LOD score higher than 3 is considered significant evidence for linkage, with a 1 in 10^3 risk of a false positive.

This method is useful to study large families or many genetically homogeneous families with unambiguous medical diagnoses.

SNP arrays

SNP arrays allow for simultaneous genotyping of millions of SNPs via the use of synthetic oligonucleotide microarrays containing unique allele-specific probes [53]. Single-stranded DNA is digested with restriction enzymes, adaptors are ligated to the resulting fragments and these are PCR-amplified and hybridized to the array. SNP arrays allow the detection not only of SNPs, but also of CNVs and loss of heterozygosity [54, 55].

This method can be used in linkage analysis and autozygosity mapping. Autozygosity mapping is a useful technique when looking for recessive variants in consanguineous families. The individuals are said to be autozygous when they share a homozygous haplotype, which is identical by descent, i.e. is inherited from a recent common ancestor [27].

Next-generation sequencing

Traditional Sanger sequencing produces long reads of high quality, however it does not allow for multiplexing and is costly [56]. Therefore it is now mostly used as a validation tool for the enormous amount of data generated by next-generation sequencing (NGS). Several different platforms for massively parallel sequencing exist, but all of them, except for the "third-generation" platforms, offer multiplexed sequencing of a genomic DNA library that has been fragmented and amplified by either solid-phase, or emulsion PCR [57-60]. The most commonly used NGS platforms in the past decade have been 454 sequencing (Roche) [61, 62], SOLiD (Applied Biosystems) [63], HiSeq and MiSeq (both by Illumina) [64], and more recently, the so-called "third-generation" sequencing platforms – Ion Torrent, Ion Proton and PacBio [65, 66].

- 454 – Single-stranded DNA library fragments linked to adaptors are mixed with capture beads, and emulsion PCR is performed in order to produce millions of clonally amplified sequence templates on each bead. These beads are then loaded onto a PicoTiter™ plate and sequencing-by-synthesis is initiated simultaneously for the entire genome. The 454 sequencing platform relies on pyrosequencing, which is based on the detection of a luciferin signal whenever a nucleotide complementary to the template is incorporated in the growing DNA strand. Over 400 000 short reads are generated per run. One advantage is the read length – around 700 bp, the time run is speedy compared to other next/second generation sequencing methods and accuracy is high, however not in the case of long stretches of repetitive bases (more than 6 bp). The cost is another disadvantage [56].
- SOLiD – Analogous to 454 sequencing, clonal amplification is achieved by emulsion PCR. Instead of using a plate, beads are then immobilized on a glass slide. On the slide, sequencing-by-ligation takes place with fluorescently labeled di-base probes. The SOLiD system provides a highly accurate read-out, particularly when coverage exceeds 30x. The read length is short however (50 bp), and this is the main drawback of this system [56].
- HiSeq 2500 and MiSeq – The HiSeq platform supports a wide range of applications, including whole genome, exome and transcriptome sequencing, creating reads of up to 250 bp in length. The MiSeq platform is faster and suitable for sequencing of small genomes or targeted gene panel sequencing, and it generates 300 bp-reads [66]. Clonally amplified DNA templates on acrylamide-coated glass flowcells are used for sequencing-by-synthesis, in which nucleotides are added base by base and a fluorescently labeled reversible terminator is imaged with the incorporation of each nucleotide. The main advantage when using the HiSeq

platform is its high throughput, but the short read assembly is a drawback [56].

- Ion Torrent/Ion Proton – This is a third-generation sequencing platform introducing some novel technological principles and solutions [66]. The first steps in the process are identical to 454 and SOLiD: adaptor-ligated DNA fragments are clonally amplified by emulsion PCR on beads known as Ion Sphere Particles. These are however loaded into proton-sensitive wells. In the course of sequencing, each of the four bases is introduced in succession. Whenever a nucleotide gets incorporated, there is a release of protons and a signal is detected. The read length is around 200 bp and the accuracy is lower than that of the Illumina platforms, but it has a very short run time of 2 hours.
- PacBio – This technology relies on fundamentally different principle compared to the others reviewed above, as it takes advantage of single molecule real time sequencing [66]. Here, the DNA template is mixed together with DNA polymerase and gamma-phosphate fluorescently labeled nucleotides in wells called zero-mode waveguides. The width of those wells does not allow light to penetrate the waveguide. However, energy can excite the fluorophores in the vicinity of the polymerase, and with the incorporation of every base, a pulse of fluorescence is detected in real time. The PacBio platform boasts an impressive read length of 1.5 kb and a comparable run time with the IonTorrent. A downside is the high error rate and price [66].

Some of the major applications for which NGS is used include *de novo* sequencing of eukaryotic genomes [67], resequencing (which is used for the identification of novel SNPs, CNVs, etc.) [68, 69], detection of genome-wide epigenetic modifications [70], etc. In addition, these platforms are suitable for RNA sequencing of whole genomes [71], but also for exomes, which are interesting in Mendelian traits where the majority of pathogenic lesions are found in protein-coding genes [72].

Ethical considerations

The acquisition of massive amounts of genomic data is now possible thanks to NGS technology, and it creates a number of ethical issues that need to be addressed. One important and sensitive question is whether and how to provide the participants in NGS studies with access to the research results. It has been suggested that participants should have the right to choose whether they want to access this information, as long the results are scientifically sound, clinically significant (suggesting serious health risk) and are disclosed according to local laws [73, 74]. Results with unclear consequences that are not likely to benefit the participants should preferably not be reported back [73]. The format of delivery needs to be considered as well, as providing the

participants with raw data will be meaningless to most of them [75]. At the same time, interpretation of the results is costly, and when incomplete, may lead to misunderstanding, which in a clinical context may have serious consequences [75]. However, having access to trained physicians capable of correctly interpreting the data for the participants remains a challenge.

The storage of the vast amount of personal data is another problem which deserves attention. An understandable fear for privacy that society has faced is the possibility for this information to be misused by potential employers, insurance companies, etc., whenever health risk variants have been detected [76]. Therefore, the data needs to be securely stored to decrease the possibility of tracing it back to each participant in the study.

Additionally, this data may be used in the future for purposes not initially outlined in the informed consent documentation, which creates possibilities for breach of confidentiality [77].

Another controversial point concerns the obligations of the researcher to the relatives of the participants: should the data be available to the families of the participants, whose health may also be affected by a detected pathogenic variant? McGuire et al. suggest open discussions with the participants in order to inform them of the implications and meaning of the data for their families without the requirement for informed consent from the relatives [75, 76]. Once again, the benefits should be weighed against the potential risk of information disclosure to affected relatives.

Whole exome sequencing

In whole exome sequencing (WES), only the coding parts of the genome are targeted [78]. Since it remains challenging to define with certainty which parts of the genome are truly protein-coding, "exome" in this context means targeting at least all sequences listed in RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) [79], plus a number of hypothetical proteins, with the provision that capture probes vary in efficiency and thus certain targets may be missed entirely [80]. For library preparation, genomic DNA is hybridized to oligonucleotide baits that are homologous to coding regions, while the unbound DNA is washed away. Exome-complementary DNA is then sequenced using the platform of choice (*Figure 1*).

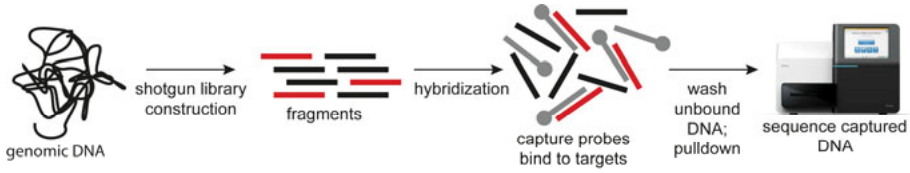

Figure 1. Principle of sample preparation for WES. Shotgun libraries are constructed from genomic DNA and then fragmented. The fragments are then hybridized to capture probes which are specific for the coding part of genome. Thus, the non-coding fragments are washed away, while the exonic fraction is enriched and can be sent for sequencing to the NGS platform of choice.

Twenty thousand or more single nucleotide variants per individual are detected by this method on average [80]. Roughly 9 000 – 11 000 of these are non-synonymous and a slightly larger proportion are synonymous [81]. To discern the possible disease-causing variants from normal variation, filtering is normally performed against publicly available databases such as dbSNP, EVS or 1000 Genomes Project, as well as against a set of healthy individuals (controls). This type of filtering is particularly valuable when studying high-penetrance rare variants, such as the ones that are usually in the basis of Mendelian traits, because it allows elimination of all common variants (i.e. variants that are unlikely to produce a phenotype on their own). This assumption, however, needs to be taken with caution, because the databases contain a small number of pathogenic variants – for example, the data may come from unaffected carriers of pathogenic alleles that only produce a phenotype when in a homozygous state [80].

Another thing to keep in mind when filtering is the minor allele frequency (MAF) of the variant – the frequency at which the less abundant allele occurs in a population. Low MAF/ high penetrance variants are the ones expected to cause Mendelian disorders (*Figure 2*). If MAF is disregarded when filtering, one risks missing actual disease-causing variants; lowering the threshold down to 0.1% for dominant disorders helps narrow down the remaining number of candidate variants. Furthermore, it is possible to stratify the data according to different criteria, such as class of mutation (frameshift, stop codon, splice-site, etc.), gene function (if known), and so on. All of this should preferably be combined with pedigree information and observed mode of inheritance. In fact, the latter should be kept in mind when conceiving the WES experiment and selecting individuals for sequencing. It is likely to confuse identity-by-state with identity-by-descent in close relatives when looking for very rare variants. Thus, in order to minimize this possibility, it is advisable to choose the most distantly related individuals in the pedigree for sequencing i.e. those who are least likely to share a haplotype based on a common ancestor [80].

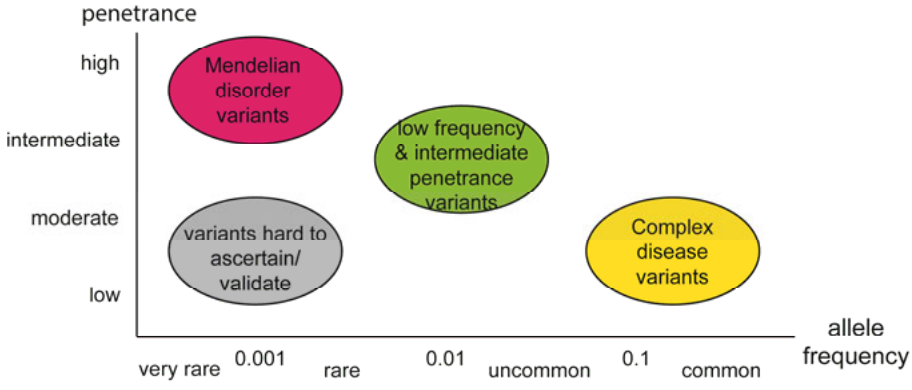

Figure 2. Dependence of variant type on allele frequency and penetrance. Mendelian disorders are normally caused by very rare high-penetrance variants, as opposed to complex/ common disorders, which are a result of common variants with low penetrance. Other frequency/penetrance correlations might be less straightforward to interpret.

Exome sequencing has, to date, been essential for the identification of numerous disease genes and for diagnostics [80, 82-85]. It is an affordable method compared to genome-wide association studies (GWAS) and whole genome sequencing (WGS) (although novel technologies in NGS may challenge the financial aspect). In addition, the resolution of WES is higher when compared to GWAS and the amount of output data is smaller and easier to interpret than that generated by WGS.

Whole genome sequencing

WGS is the most comprehensive method to estimate individual variation and mutations both in the coding and the non-coding part of the genome [86]. Thanks to the rapid development of NGS technology and the bioinformatic tools to analyze the output data, WGS is becoming more accessible. There are two approaches when performing the method. The reference-based assembly relies on mapping the reads from NGS to a reference sequence [86](for example, hg19 for humans). Variant calling and annotation follow [81]. When a reference sequence is not available, one may resort to a *de novo* assembly approach, in which the reads are screened for overlaps to build longer contiguous sequences (contigs) [86]. From a diagnostic point of view, WGS is suitable not only for Mendelian disease gene-identification like WES, but also for detection of causative variants in complex traits, *de novo* single nucleotide variants and CNVs, as it can capture nearly 3.5 mil-

lion variants [81]. This enormous amount of data is however challenging to analyze and store, and the method is still more expensive than WES [81].

Transcriptome sequencing

Looking into the exome is informative, but yet another interesting method that provides direct insight into the expression of the exome is RNA sequencing [87]. It is a way to capture the transcriptome of a cell (or tissue), i.e. its entire set of mRNA and non-coding RNA transcripts at a given time [71]. Until recently, mRNA expression was commonly measured either by qRT-PCR or by the use of microarrays. However, qRT-PCR is expensive and inapplicable to genome-wide analysis, whereas microarrays cannot compete with the sensitivity of RNA sequencing [71, 88]. RNA sequencing provides an opportunity to measure absolute transcript levels, so it is a powerful quantitative method, although the sensitivity varies depending on the NGS platform used (for example, Illumina platforms can generate millions of reads, whereas PacBio generates only around 50 000 reads; however, these are long and allow characterization of the transcript structure) [89, 90]. In addition, RNA sequencing makes it possible to detect alternative splicing; due to technical challenges, however, sequences near the 5' and the 3' ends of the transcripts are underrepresented [89, 91, 92].

There are multiple different protocols for performing RNA sequencing, but they mostly rely on the same general steps: fragmentation, RNA-to-cDNA conversion, second strand synthesis, adapter ligation and amplification (*Figure 3*). Fragmentation may be performed either on RNA level, or at the cDNA stage, but RNA fragmentation has been demonstrated to induce less bias [89, 93], and hence seems to be the more popular choice. Often, polyadenylated RNA is used, which includes mRNA, long noncoding RNAs, snoRNAs, etc. If total RNA is used as starting material, rRNA needs to be depleted beforehand, as it makes up to 80% of the RNA pool [89]. Either random hexameres, or oligo-d(T) primers are used for cDNA synthesis. However, this random priming leads to loss of strand specificity, which may be important with regard to antisense regulation [89]. Some protocol modifications overcome the problem with strand specificity [94].

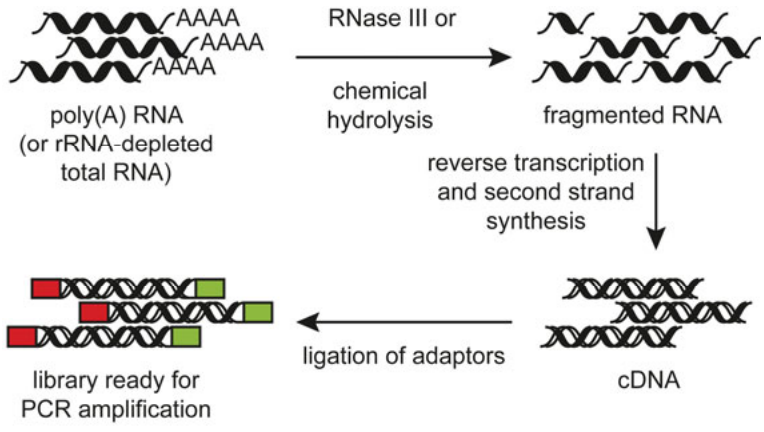

Figure 3. Principle of RNA sequencing. Either ribosomal RNA-depleted total RNA, or polyadenylated RNA is used as starting material. It is then subjected to fragmentation in order to ensure equal fragment sizes, and to reverse transcription and second strand synthesis to produce cDNA. Next, adaptors are ligated to both 5' and 3' ends and the products are PCR-amplified.

In summary, transcriptome sequencing is a highly sensitive method for capturing and quantifying absolute expression, it is suitable for genome-wide analyses and under certain conditions may detect alternative splicing (as long as certain level of coverage is reached [89]) or transcript structure (when the sequencing platform allows for long reads). Downsides of the technology include underrepresentation of 5' and 3' ends and the method is usually not strand specific.

Genotype-phenotype correlations: investigating gene function in support of variants' pathogenic role

"Who in the world am I? Ah, that's the great puzzle."
— Lewis Carroll

The identification of gene variants that might be causative for a certain disorder via next generation sequencing is powerful, but has its limitations. To prove pathogenicity of an identified variant one must resort to additional validation methods.

Validation and prediction tools

Sanger sequencing is commonly used to verify that the findings of massively parallel sequencing are not artefactual. After confirmation of the mutation, screening the segregation of the variant in the family is another test that can be done to check whether it correlates with the observed disease inheritance pattern.

Next, one commonly compares the protein sequences of orthologous genes found in different species in order to elucidate whether the position of the detected variant is highly conserved, which is a strong argument in favor of its functional importance in evolution. One widely used bioinformatics tool which provides a platform for such conservation comparisons and protein alignment is Clustal Omega (http://www.clustal.org/omega/) [95].

Variants that alter the conformation or biochemical properties of a protein are those that are most likely to be pathogenic. Therefore indels and frameshift mutations are generally considered to be the best candidates for disease-causing variants. To test this hypothesis, one can resort to a multitude of online tools such as PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/), MutationTaster (http://www.mutationtaster.org/), and SIFT (http://sift.jcvi.org/), all of which predict the effect of the identified mutation on protein function [96, 97]. Such bioinformatics-based methods can provide supporting evidence to the assumption that a variant is pathogenic. However, the golden standard to

prove pathogenicity is to run functional tests and demonstrate the role of the variant and its effect in model systems.

Functional model systems

In vivo systems

In addition to facilitating our understanding of basic gene functions, model organisms have traditionally been used in various applications in research, medicine and agriculture. Multiple model organisms exist, but among the most relevant to understanding human disease have been *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (fruit fly), *Danio rerio* (zebrafish), *Mus musculus* (mouse), and *Ratus norvegicus* (rat) [27].

Possibly the most popular invertebrate model organisms in genetics have been *C. elegans* and *D. melanogaster*. Many human genes have corresponding homologs found in these two species, which make the nematode and the fruit fly suitable models for studying human gene and protein function [98]. Both species are also easy to study and have a short life cycle, which allows for observation of aging and mutation effects within a short timeframe. *C. elegans* is transparent, which greatly facilitates the tracing of cells and cell lineages throughout development, as well as the observation of neuronal connection formation. In addition, the transparency of the nematode makes it possible to easily track fluorescently-labeled proteins and study gene function; it also provides a platform for drug testing and discovery [99, 100]. The fruit fly is an excellent model organism for studying spatially and temporally restricted expression of transgenes [27]. A *D. melanogaster* transposon called the P element can be used as a tool for mutagenesis or transgenesis [101]. The fruit fly is also a widely used model for behavioral studies, development and neuroscience [27].

Naturally, the more evolutionarily close an organism is to humans, the more accurate system for human disease modeling it makes. Non-mammalian vertebrates provide an excellent opportunity to study vertebrate development, since they produce relatively large eggs, often fertilized and developed outside of the body, and are easy to manipulate. The zebrafish has multiple highly conserved developmental genes with orthologs in human, as well as transparent embryos which allow observation of mutants [27, 102, 103]. Gene inactivation is relatively easily achieved by the use of morpholinos [104].

Since mammals are closest to humans with respect to physiology, development and biochemistry, they are very suitable model organisms. For the same reasons however, they have the disadvantage of being more governed by ethical considerations, which may complicate study design and makes some experiments impossible. In addition, they usually produce less off-

spring and have a longer lifespan. Rodents overcome some of these limitations by having relatively short generation times and large numbers of offspring. Mice and rats have been used for decades for large-scale mutagenesis screens, human disease modeling, drug development and behavioral studies [105, 106]. The possibility of performing various genetic crosses and creating transgenic animals, conditional and constitutive gene knockouts has allowed detailed studies of gene function in health and disease using these model animals [107].

Usually, model organisms are manipulated in order to answer a particular scientific question. An exogenous gene might be integrated into the animal's germline in order to study its function or induce disease (transgenic animals); alternatively, endogenous genes may be altered with the same purpose [27].

In vitro systems

In addition to model organisms, one could use simpler biological systems for *in vitro* applications. A wide range of cell lines exist, both of animal and human origin, which can be used for knockdown, knockout or overexpression of proteins and RNA, various types of genetic manipulation, studying of molecular pathways and pharmacological response [108, 109]. A library of multiple cancer cell lines can be found at the COSMIC cell line project (http://cancer.sanger.ac.uk/cancergenome/projects/cell_lines/about).

Among the most relevant cell lines for modeling human disease phenotypes or studying development, however, have been pluripotent stem cell lines. Embryonic stem cells (ESCs) are pluripotent stem cells derived from the early embryo (the inner cell mass of the blastocyst) and can be propagated indefinitely in culture [110]. ESCs have been generated from the embryos of different species, such as mice and even humans. Murine ESCs can be used to deliver exogenous DNA or for gene inactivation by genetic manipulation and subsequent injection into blastocysts that can be re-implanted in a female mouse. This is a way to generate transgenic or knockout mice, respectively [111]. In culture, ESCs are capable of differentiating to all three germ layers: ectoderm, mesoderm, and endoderm, which makes it possible to study a particular cell type of interest *in vitro* [112]. ESCs can be subjected to directed differentiation to any specific somatic cell type and thus have an enormous potential for usage in disease modeling and regenerative medicine, although the latter is a subject of controversy due to graft-versus-host disease [27, 113]. Mouse ESCs can also contribute to the formation of an entirely new organism from cell culture [114].

An "artificially" derived alternative to ESCs are induced pluripotent stem cells (iPSCs), pioneered by Takahashi and Yamanaka in 2006 [115]. iPSCs are somatic cells that have been manipulated back to a pluripotent stage by introduction of transcription factors (*Figure 4*)[116]. Different cocktails of

transcription factors have been developed over the last decade, but the four initially identified by Yamanaka are still the most widely used ones – Oct4, Klf4, Sox2 and c-Myc (OKSM) [117]. These can, however, be replaced by various other approaches with comparable efficiency. For example, Yu et al. demonstrated that reprogramming can be achieved by replacing the oncogenes Klf4 and c-myc from the OKSM cocktail with Nanog and Lin28 [118]. Furthermore, it was shown that c-Myc is not indispensable to reprogramming; however, its presence increases the infamously low reprogramming efficiency [119]. It is possible to facilitate reprogramming by introducing small compounds [120], or to skip transgene delivery altogether and use miRNAs [121, 122] or recombinant proteins/ poly-arginine [123, 124] instead.

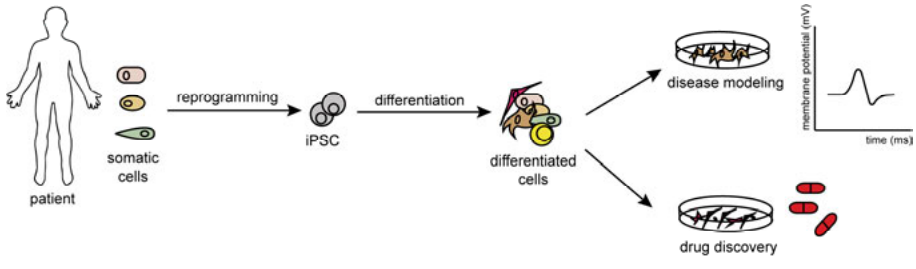

Figure 4. Schematic representation of iPSC reprogramming and some applications. Using the method of choice, somatic cells derived from the patient are induced to pluripotency. The resulting iPSCs can be differentiated to any cell type and then used, among other applications, for disease modeling or drug discovery.

The transcription factor delivery can also be done in different ways. Most commonly, it was performed with viral vectors which integrate into the host cell genome, such as retrovirus or adenovirus [117]. Improved efficiency was achieved by the use of polycystronic lentivirus cassette [125], which could also be excised from the genome with the help of Cre/LoxP sites [126]. Virus-free delivery can be done by transposons which excise themselves nearly without a trace [127]. Recently, non-integrating approaches have become a popular choice replacing the older methods, due to the fact that they do not create a risk of insertional mutagenesis. One of the most popular methods currently involves the Sendai virus vector. This technique has been reported to reprogram fibroblasts at a higher efficiency than other methods, and is nonpathogenic to humans, making it suitable for regenerative medicine applications [128, 129].

The cells that result from reprogramming have the properties of self-renewal, indefinite propagation in culture and the ability to differentiate to all germ layers, like ESCs. Similarly, mouse iPSCs can contribute to viable mice by tetraploid complementation [130]. In addition, iPSC lines circumvent the ethical problems of ESC derivation, since they do not originate from

embryonic material. Since the somatic cells used for their derivation are taken from the patients, theoretically they should not cause an immune response if transplanted (although there has been some evidence for immunogenicity [131]).

Despite being very similar to ESCs, iPSCs are distinct when it comes to their transcriptome and methylome profiles [132, 133]. Furthermore, in contrast to the *in vivo* blastomeres, human-derived female iPSCs do not reactivate both X chromosomes, which makes them an unfit model for X-linked disorders [134]. A number of authors have attempted to completely erase the epigenetic marks of the parental cells and induce the so-called "naïve" pluripotentcy in human iPSCs, denoted by hallmarks such as dome-shaped colony growth, ability to survive and maintain pluripotency as single cells, dependence on leukemia inhibitory factor (LIF) and reactivation of the two X chromosomes in females [135, 136]. These methods include treatment of the cells with telomerase [137] and culturing in defined media [136] or on specific LIF-expressing feeders [138]. Additional reports using the aforementioned technologies to model X-linked diseases are anticipated in order to support their efficacy in this setting.

Caution is also advised when it comes to the choice of suitable control lines. It has become common practice to reprogram iPSCs from healthy donors to use as controls; however, controversy has arisen due to the possibility that the different genetic backgrounds of the control lines may complicate data interpretation in Mendelian disease modeling. One way to circumvent this problem is to use isogenic lines, i.e. lines that arise from the patient cells but have been genetically corrected, so that they do not carry the patient mutation or vice versa [139, 140]. Different tools such as zinc-finger nucleases [129], TALENs [141] and recently, the CRISPR-Cas system [142] offer this possibility, though off-target insertions or deletions in the genome are a source of concern [143]. In addition, these methods are still not suitable to mimic or correct all type of lesions, specifically large structural aberrations.

In summary, the choice of method or model organism is tightly dependent on the specific research questions that one poses. Both laboratory animals and *in vitro* systems have their advantages and drawbacks, and a combination of online tools, *in vivo* and *in vitro* approaches is often the best way to study human disease in order to reflect as many of its aspects as possible.

Ethical considerations in stem cell research

Stem cells cause both enthusiasm and controversy. Their possible and promising application in cell replacement therapy is a source of hope for patients with degenerative disorders such as Parkinson and Huntington disease [144, 145], but there are some serious social, moral and ethical questions to be solved. The major ethical concerns are connected to human ESCs, the derivation of which requires that early preimplantation embryos are destroyed [144-148]. Some alternative protocols which avoid that have been published,

but the method has proven problematic [144, 149]. A common source of embryos for making human ESC lines are abortions or surplus embryos created for *in vitro* fertilization (IVF) [145-147]. Most countries require separate forms for informed consent, where parents should agree not only to the abortion, but also to donating the embryonic tissues to research; additionally, embryos for IVF can usually be used for creating ESC lines only if they would have been discarded otherwise, but they cannot be specifically created for research purposes [145]. Even those regulations, however, are met with discontent and controversy by part of society, who consider the embryo to be equivalent to a person. According to this view, destroying an embryo is tantamount to murder [145]. Others embrace a more "gradualist" approach, believing that the moral value of an embryo increases in the course of development and after passing certain milestones, such as developing a nervous system and cognition [144, 145]. Since these processes occur much later than the blastocyst stage, they see no ethical dilemma in destroying early embryos, especially with regard to the possible benefits that may come from stem cell research.

While it seems that these opposing viewpoints create an irresolvable controversy, iPSCs offer a way to circumvent the use of embryonic material and thus create a more ethical alternative to ESCs. The use of iPSCs, however, poses different types of ethical questions for society and the scientific community. Several reports have shown that both mouse iPSCs and ESCs can give rise to oocytes and sperm [150, 151], and a recent publication described the direct differentiation of human iPSC and ESCs into spermatogenic cells [152]. This discovery raises the question of whether germ lines derived from ESCs/iPSCs can be used to crate viable embryos. Although such an alternative may provide a solution for infertility, there would be considerable concern as to whether this is morally acceptable. Questions which need careful consideration include the prospect of serious health issues for children born through such an "immaculate conception" [146]; the possibility for selection of "the best" genes or a particular gender as desirable traits in the artificially created embryos brings Huxley's *Brave New World* to mind.

Despite the fact that a number of the abovementioned concerns may not be prevalent in society worldwide or may still be in the realm of remote possibilities, a careful and well-regulated balance needs to be achieved in order to allow the development of stem cell technology and cell replacement therapy while respecting ethical norms and human values.

Methods

The main methods used in the papers which comprise this thesis are briefly outlined below.

Whole exome sequencing

WES was in the basis of both papers I and II. In each case, we investigated large families clearly suffering from inherited disorders. Since the symptoms were pronounced and consistent with autosomal recessive (paper I) or autosomal dominant (paper II) inheritance typical of monogenic disorders, we reasoned that the causative mutations were likely found in the coding parts of the genome and thus resorted to WES as a comprehensive and affordable method to identify candidate variants [78]. Selected individuals were sampled for genomic DNA, and capture probes were used to enrich for the exonic part of the genome, which was then sequenced by either the SOLiD5500xl (paper I) or the Ion Proton (paper II) systems. The detected variants were then filtered against 350 "in house" exomes and several databases (as described in detail in the chapter *Approaches to study the genetic causes of human monogenic disorders*), allowing us to successfully identify the causative mutations in both families.

Immunohistochemistry/immunocytochemistry

Immunostaining of paraffin embedded tissue sections or fixed cells was used in papers I, II, III and IV. Whenever tissue sections were used, deparaffinization in xylene and ethanol was performed as a first step, followed by antigen retrieval with trypsin in humidified chamber in order to open any cross-linked epitopes and allow for antibody binding. All slides were fixed with PFA, and washed with blocking solution to ensure specificity of the subsequent staining with antibodies. This technique was applied as part of functional analysis to visualize protein expression of disease-causing variants (papers I and II) or as a way to characterize the pluripotency of iPSCs (papers III and IV).

iPSC generation and culturing

iPSCs were generated by introduction of the Yamanaka factors OKSM to fibroblasts from patients or healthy individuals [117]. The transgene delivery was achieved either via lentiviral stem cell cassette [125] (papers III and IV) or by four separate episomal Sendai virus vectors [128] (paper IV) (see *In vivo methods* section for details). Fibroblast cells were grown on mitomycin C-inactivated feeders to support self-renewal by providing extracellular matrix proteins and growth factors [153] for about one month before hand-picking the first colonies. These were then cultured clonally in bFGF-supplemented media for multiple passages until acquiring stem cell-like characteristics such as large nuclei, formation of tightly packed colonies and expression of pluripotency-associated markers.

Transcriptome sequencing

RNA (or transcriptome) sequencing was the main method used in paper III. Several iPSC lines, fibroblast lines and one ESC line were harvested to isolate total RNA, which was then depleted of the ribosomal RNA fraction, fragmented and reverse-transcribed [89]. The resulting cDNA was then ligated to adaptors and sequenced in order to determine the type and abundance of transcripts and to compare the expression profiles between samples. Analysis of transcriptome data allowed for transcript quantification and identification of differentially expressed genes and pathways between cell lines.

Karyotyping

Cytogenetic analysis was used as part of iPSC characterization in both papers III and IV, and was the focus of paper IV as a method which allows comparison of the effects of different viral delivery systems for iPSC generation on genome stability. Non-confluent iPSC (or fibroblast) cultures were treated with colcemid to arrest the cell cycle and allow for metaphase analysis. Cells were then washed with hypotonic solution and fixed on microscope slides. Giemsa-staining was performed following trypsinization and the visualized metaphase chromosomes were analyzed for aberrations.

Present investigations

"Never let the future disturb you. You will meet it, if you have to, with the same weapons of reason which today arm you against the present."
— Marcus Aurelius

Paper I. Autosomal Recessive Transmission of a Rare *KRT74* Variant Causes Hair and Nail Ectodermal Dysplasia: Allelism with Dominant Woolly Hair/Hypotrichosis

Aim:

- To identify the cause of disease in a family with ectodermal symptoms

Methods:

- WES was used to identify the putative gene candidate
- Segregation analysis was performed by bidirectional Sanger sequencing
- Probable pathogenicity of the variant was determined with PolyPhen-2 and conservation was assessed by running a prediction model in Clustal Omega
- Immunohistochemistry was performed to demonstrate how protein expression was affected by the identified variant

Results and discussion:

In this study, we re-investigated a large consanguineous Pakistani family with autosomal recessive pure hair and nail ectordermal dysplasia (AR PHNED (MIM 602032)). The patients were born to healthy related parents and suffered from mild micronychia, hypotrichosis and brittle hair. Previous studies showed a shared homozygous haplotype that was mapped to chromosome 12q12-q14.1 which harbors the cluster of type II keratin genes, but none of the two genes previously implicated in the disorder – *KRT85* and *HOXC13* – were found to deviate from the reference sequences.

In order to elucidate the genetic background behind the disease in our patients, we used WES and focused on the linked region to look for causative mutations. After filtering against common variants and Sanger sequencing

verification, we identified three homozygous variants in the 12q12-q14.1 locus: a c.821T>C transition in *KRT74* (NM_175053; p.Phe274Ser), a c.38C>T transition in *CELA1* (NM_001971; p.Pro13Leu) and a c.1037A>G transversion in *IKZF4* (NM_022465; p.Tyr346Cys). All three segregated with the phenotype, and although the c.821T>C variant in *KRT74* was reported in the Exome variant server (EVS) and dbSNP132 databases (at low frequencies, 2/10 000 and 2/1000, respectively), we reasoned that this was the most plausible among the three candidates, based on the reported functions and expression patterns.

We looked for further support for the *KRT74* variant pathogenicity by running PolyPhen-2 predictions of protein damage caused by the mutation, and in contrast to *CELA1* and *IKZF4*, the *KRT74* transition received the highest score of "probably damaging". We then made a conservation comparison between the human Keratin-74 protein sequence and its orthologs in other species, and noticed that the Phe274 residue was conserved in all of them. In addition, literature searches suggested that the coil 1B domain which harbors the mutation is important for dimerization in keratins, which are obligate heteropolymers. Finally, we compared the protein sequences in the coil 1B domains of all the 26 human type II keratins, to which Keratin-74 belongs. We discovered that the variant we identified was conserved in all except for Keratin-80, where it was substituted for the biochemically synonymous leucine.

Since the bioinformatic predictions strongly supported the c.821T>C missense mutation in *KRT74* as a cause for PHNED in our patients, we proceeded with studying the expression of Keratin-74 by immunohistochemistry. Notably, it was expressed in the nail matrix and nail bed in adult mouse claws, suggesting its role in nail formation and integrity. In addition, we compared Keratin-74 expression in human hair follicles in a normal individual and a patient. We observed strong staining in the inner root sheath of the control, but none in the patient, which additionally supports this variant as a cause of PHNED.

Paper II. Phenotypic expansion of visceral myopathy associated with *ACTG2* tandem base substitution

Aim:

- To investigate the cause of disease in a family with a variety of symptoms from the visceral organs

Methods:

- WES was used to identify the putative variant
- Segregation analysis was performed by Sanger sequencing
- PolyPhen-2 was used to predict variant pathogenicity and NNSplice Software was used to predict altered splicing
- The transcript abundance in patients and controls was assessed with qRT-PCR
- RT-PCR was further used to check for exon skipping or use of cryptic splice sites; the result was verified by Sanger sequencing of cDNA
- Protein expression was verified by immunohistochemistry
- SwissModel was used for constructing a 3D protein model

Results and discussion:

This paper attempted to elucidate the genetic cause behind the gastrointestinal problems in a large Swedish family. The patients complained from abdominal pain, distention, obstipation, abnormal peristalsis and other symptoms from the visceral organs, consistent with a diagnosis of familial visceral myopathy (FVM). Visceral myopathies (MIM155310 and MIM613834) comprise a group of heterogeneous disorders caused by impairment of smooth muscle function, associated with multiple variants in the smooth muscle actin gamma-2 gene, *ACTG2*.

We sampled four patients for WES and following confirmation with Sanger sequencing we identified a previously unreported variant in *ACTG2* – a heterozygous tandem base substitution c.806_807delinsAA (p.(Gly269Glu)) in the affected family members. The substitution is positioned at the beginning of exon 8 and was predicted by the NNSplice Software to potentially affect splicing.

We then examined the expression levels of the *ACTG2* mRNA transcript in fibroblasts and discovered a three-fold decrease in patients compared to controls. To account for the possibility of exon skipping (as suggested by prediction software) or the use of a cryptic splice site, we performed RT-PCR on cDNA of patients and controls. However, we found that the product sizes were consistent with normal splicing. To exclude nonsense mediated decay that masks abnormal splicing, we sequenced these PCR products and

were able to detect both the wild type and the mutant transcript, which confirmed that splicing was not affected by the TBS in the patients.

Next we investigated if these low transcript levels correlated with low protein levels by staining patient ileum sections for ACTG2 and observed no difference relative to healthy individuals. This may be explained by focal abnormalities not detected in our tissue sections or lack of sensitivity of light microscopy for quantitation. Our 3D protein modeling suggested that the substitution of the uncharged wild type glycine in position 269 with the polar and bulkier glutamic acid alters the physical distances to adjacent amino acids and may have an effect on actin polymerization. Taken together, these data suggest a dominant negative effect of the variant we detected. Our study also expands the knowledge of the clinical phenotypes associated with FVM, which may facilitate diagnosis and treatment in the future.

Paper III. Transcriptome profiling reveals degree of variability in iPSC lines: Impact for human disease modeling

Aim:

- To determine the level of variability between sister iPSC lines (i.e. lines originating from the same donor's somatic cells)

Methods:

- iPSCs were reprogrammed from two healthy donors' fibroblasts using a lentiviral vector with OKSM transcription factors
- Pluripotency was assessed by immunofluorescent staining and TLDA assay for pluripotency markers, and EB differentiation capacity
- Chromosomal integrity was assessed by karyotyping
- Click-it EdU assay was used to estimate cell proliferation
- RNA sequencing was used to compare expression profiles between cell lines

Results and discussion:

In an attempt to broaden the technical knowledge on iPSCs, we compared different iPSC lines in order to elucidate the minimal number of clones required and how to best account for biological variability.

To this end, we established fibroblast lines from two healthy donors, HDF-K1 and HDF-K2. Three sister iPSC lines and one iPSC line were derived from these fibroblasts respectively. A hESC line was used as a pluripotent control.

We first verified that we had produced high-quality iPSC lines, as only few cells in the reprogramming process are known to reach the pluripotent and unlimited self-renewal stage. We subjected our lines to stringent criteria, which they met by displaying 1) characteristic morphology, 2) positive staining of the markers Nanog, TRA-1-60, TRA-1-81, SSEA-3 and -4, 3) expression of pluripotency-associated genes similar to the hESC line and 4) ability to differentiate into all three germ layers under withdrawal of growth factors during culturing. In addition, all our lines were cytogenetically normal.

We resorted to RNA sequencing and rigorous transcriptome analysis of all lines, and observed that the expression profiles of the three sister iPSC lines were highly similar to one another but distinct from the profiles of the iPSCs derived from HDK-K2 and the hESC line. As expected, fibroblasts showed a very different expression profile from all pluripotent lines. A similar observation could be made regarding whole transcriptome profiles of EBs

derived from each pluripotent line: lines which originated from the same parental fibroblasts clustered together and separately from the other lines.

Gene ontology and KEGG pathway comparison within the iPSC lines revealed that there was no significant difference in expression between sister lines; however, there were more than 3000 differentially expressed genes (but not pathways) when sister lines were compared to the remaining iPSC line, and these were primarily involved in gene regulation. We then compared all iPSC lines to the hESC line and detected three significantly differentially expressed KEGG pathways which were involved in RNA transport, ribosome biogenesis and spliceosome formation. This observation suggested a difference in proliferation, thus we performed a proliferation assay based on EdU-labelling. However, our results indicated similar generation rates among all pluripotent lines.

Taken together, our results demonstrate a pronounced similarity between clonally derived lines which originate from the same donors. In this work we also highlight the importance of picking several fibroblast lines as starting material for reprogramming rather than simply relying upon multiple sister iPSC lines.

Paper IV. Methods of reprogramming to iPSC associated with chromosomal integrity and delineation of a chromosome 5q candidate region for growth advantage

Aim:

- To compare the effects of two different methods of viral transgene delivery on chromosomal integrity in iPSCs

Methods:

- iPSCs were reprogrammed using the Yamanaka factors by either introducing them with a lentiviral vector, or with Sendai virus
- iPSCs were characterized as pluripotent by 1) immunocytochemical staining for pluripotency-associated markers, 2) stem cell marker TaqMan array and 3) their ability to differentiate into all germ layers by forming embryoid bodies
- Karyotyping was performed in order to estimate frequency and types of genomic abnormalities in iPSCs

Results and discussion:

Ever since iPSCs were pioneered as a method by Takahashi and Yamanaka in 2006, their popularity as an *in vitro* model system has been rising and they show great potential for human disease modeling. However, in order to understand disease, one first needs to understand the advantages and limitations of the methods one employs. Since previous studies have pointed towards frequent loss of genomic integrity in iPSCs, we chose to investigate two widely-used strategies for iPSC reprogramming and their effects on the occurrence of chromosomal aberrations.

We derived sixteen lines using genome-integrating lentivirus and sixteen lines by introducing the episomal Sendai virus. In both cases we used the Yamanaka transcription factor cocktail OKSM and cultured the cells under comparable conditions. We confirmed that we have obtained high-quality pluripotent iPSCs by staining them for expression of widely known pluripotency-associated factors such as Nanog and TRA-1-60, by running an RNA-based expression array for a battery of stem cell markers and by confirming their ability to differentiate in the three germ layers in an embryoid body formation assay. We then estimated the genomic integrity of these cell lines by karyotyping.

Our cytogenetic analysis showed that at P10-21 there were seven out of sixteen lentivirus (L-) reprogrammed lines and only one out of sixteen Sen-

dai virus (S-) reprogrammed lines which were abnormal. This difference is significant ($p = 0.04$). We observed both structural and numerical aberrations and the most common ones involved chromosomes 5 and 12.

We decided to focus on these abnormal karyotypes and determine at what stage they arose. One of the most abnormal iPSC lines was L-K2A with a (51,XY,+5,12,16,17,20) karyotype detected at P17. We thawed an earlier stock of this line (P13), but it was also abnormal, with a (50,XY,+5,12,16,20[1]/50,XY,+9,12,16,20[1]/49,XY,12,16,20[10]) mosaic karyotype. This observation implied that the aberration occurred at an even earlier stage. Another aberrant sister line L-K2B (46,XY,dup(5)(q13q33)[9]/46,XY[11]) at P15 was then rethawed and the best colonies according to morphological criteria were subcloned to give rise to four daughter sub-lines. When their karyotypes were analyzed, no genomic instability was detected, which suggests that mosaic iPSC cultures can be "rescued" by a hand-picking approach as described here. Finally, we looked into the L-Liss2A line, (47,XY,+12[4]/47,XY,+mar[1]/46,XY[25]) at P17. When rethawed at P12, these cells were cytogenetically normal, i.e. the mosaic phenotype was acquired after this stage. Our observations suggest that most aberrant lines acquired their abnormal karyotypes at early stages and after reprogramming with lentivirus.

We were also interested to discover how prolonged passaging influences genomic integrity, so we karyotyped the normal lines at a late passage (P22-37). Our analysis demonstrated that the majority of lines maintained their normal karyotypes, except for S-K2A which acquired a trisomy of chromosome 5. In fact, we detected a total of three lines with aberrations involving chromosome 5, including one dup(5)(q13;q33). This observation and data from previous reports suggest that the restricted region on chromosome 5 may be implicated in growth advantage.

Taken together, our data demonstrates that non-integrating methods of reprogramming (specifically Sendai virus) are preferable to integrating ones not only because they do not carry a risk for insertional mutagenesis, but also because they tend to cause less instability in the genome of the obtained iPSC lines. Additionally, we expand the list of karyotypic abnormalities associated with iPSCs and identify a putative locus responsible for iPSC survival.

Discussion and future perspectives

"The present is theirs; the future, for which I really worked, is mine."
— Nikola Tesla

The field of human genetics is extremely broad and intricate, and understanding disease is just a piece of the puzzle. Yet by adding ever-so-small pieces one by one, we gradually get a picture of the whole, and this provides us not only with knowledge about the fundamental molecular principles of heredity and human biology, but also with a possible means to correct the "mistakes of Nature", treat disease and alleviate suffering.

Mendelian disorders are rare as a rule and only affect a small fraction of the population. Thus, one might ask – why focus on a rare disease, when multifactorial disorders are much more common and present the most significant burden on healthcare? One reason is that in monogenic disorders, there is often only a single pathway or molecular interaction affected, i.e. genetic diseases have a modular nature [154]. In addition, the genetic effects are strong and pronounced, and therefore easier to distinguish from lifestyle and environmental effects. Studying these disorders therefore provides information on the role of the gene product in signaling pathways and cascades, protein interactions, regulation of other genes, function in a specific organelle, etc. [155]. Knowing and understanding these basic interactions is the backbone of understanding disease and providing better diagnostics, which in its turn allows for treating the cause, not just the symptoms. An optimistic example is the treatment of Hutchinson-Gilford progeria syndrome (HGPS), a very rare disease causing premature senescence and early death. HGPS is caused by mutations in the *LMNA* gene which produce an abnormally farnesylated form of the lamin A protein called progerin [32, 33, 156]. The treatment with lonafarnib, a farnesyltransferase inhibitor, came only 9 years after identifying the disease cause [157, 158], which would have been impossible without the achievements of modern genetics. In addition, understanding progeria and establishing a meticulously described phenotype provided insights into the normal process of aging [159, 160]. Similarly, in paper I of this thesis we expanded the knowledge on the genetic cause of AR PHNED and our findings were intriguing not only because of the identification of a novel locus and a novel gene causing ectodermal dysplasia, but also because the inheritance pattern was recessive, which is atypical of this group of diseases. Based on crystallographic studies on similar keratins, we hypothesized

that the loss of Keratin-74 will interfere with heteropolymer formation, which is important for formation of intermediate filaments in the cell. Since AD WH, a disease without nail involvement, has been associated with other heterozygous variants in *KRT74* [161, 162], it is plausible that the variant we detected interferes specifically with interaction partners of Keratin-74 that are expressed in the nail. Thus, its loss of function produces a nail phenotype in our patients. Defining the crystallographic structure of the protein and its interaction partners will provide insights and possible explanations for the diverse phenotypes observed as a result of *KRT74* mutations.

In a similar fashion, in paper II we investigated a family suffering from symptoms consistent with FVM diagnosis. At the time when we initiated this project, the cause of this disorder was unknown; by the time we identified a lesion in *ACTG2*, this gene had already been implicated in several studies. Yet, all reported variants thus far have been single missense mutations, whereas we were the first to associate FVM with a TBS, thus contributing to the rising recognition of the importance of such aberrations in human disease. In addition, we observed wider variety of symptoms than what has been described so far. By expanding the phenotypic description, we provide a means for improved diagnosis of patients with FVM. In addition, we propose that the variant we detected acts in a dominant-negative fashion, which would make it an excellent candidate for gene therapy. The future is therefore promising for the treatment of patients suffering from this often lethal disease.

Since complex disorders also have a genetic component, identifying the causes of Mendelian disease can also help understand pathways shared between multifactorial and monogenic traits [163]. These advances are facilitated by the development of methods for detection of pathogenic variants, genome-wide studies of expression and novel model systems that provide both ethical and illustrative platforms to mimic disease. While papers I and II were focused primarily on detection of novel pathogenic variants, with papers III and IV we took a step further into functional analysis and investigated iPSCs as an *in vitro* method for human disease modeling. Paper III explored transcriptional variability between iPSC lines and how to design an iPSC-based experiment in order to best account for biological variability. Since 2006, when Takahashi and Yamanaka first published their method for iPSC generation, there has been a continuous rise in the use of this technology for disease modeling. Even though multiple studies have to date shown what a brilliant model iPSCs can make, this method was perhaps sometimes applied too hastily, and exploited a little too soon, not allowing us to fully grasp “the rules of the game”. Thus, we are still in the dark when it comes to understanding many of the limitations and possibilities of this method. By demonstrating the importance of using multiple donors for iPSC line generation, we attempted to add to this knowledge. A drawback of our study is the small sample size, which we hope to overcome in the future, but based on

the results we have, a clear trend can be observed that may facilitate iPSC study designs.

Our results in paper IV highlighted the advantages of Sendai virus reprogramming, a non-integrative method which may have implications for cell replacement therapy and regenerative medicine. Furthermore, we expanded the list of cytogenetic abnormalities associated with iPSCs and we noted chromosome 5 aberrations recurring in several of the lines we investigated. It is thus plausible that chromosome 5 harbors a locus which confers growth or survival advantage. Array-CGH would be our next step in order to investigate this in more detail. The data presented in our paper contribute to a better understanding of iPSCs as a model system and underscore the importance of routine screens for the genomic stability of iPSC lines used for disease modeling, since rearranged lines clearly do not make good models. In the future, it would be interesting to make an even wider comparison, including more non-integrative reprogramming approaches.

In summary, this thesis contributes in various ways to the existing knowledge on two monogenic disorders by identifying disease-causing variants and arguing about their potential mechanisms of action. It also expands the understanding of the cutting-edge iPSC technology as a method. My hope is that these tiny pieces of the grand puzzle which I present herein will be part of the foundation on which future studies will be based in order to develop treatment and cure of disease and build on our current understanding of the biological mechanism of life.

Acknowledgements

I would like to express my gratitude to everyone who walked along with me during the exciting roller-coaster-like journey that was my PhD: to my supervisor **Prof. Niklas Dahl**, for teaching me a lot not only about genetics but also for giving me some valuable life lessons – thank you for having me in your group where I could grow and develop as a scientist and as a person.

To my co-supervisors, who joined or left throughout the years: **Jens**, **Maria**, **Joakim**, **Marie-Louise**. Thank you for helping me add a couple of nifty methods to my toolbox, for the discussions during our numerous iPS meetings, group meetings, journal clubs and the informal conversations in the office. It has been a fun and quite unique experience to work with you!

To **Ayda**, for being a great companion when the road felt too bumpy, for changing media on weekends, for commiserating and celebrating with me, and for our discussions. I'm so glad you joined our lab and made it a sunnier place! I value our friendship immensely. But you know that ☺

To **Tahir**, who gave me the first lessons on how to use Adobe Illustrator. Who knew that stuff would turn out to be so useful?! ☺ Thank you for your patience, your help, your refreshing sense of humor and for our endless late-night conversations in the office. I miss these great times!

To our project students **Feria**, **Loora**, **Sanna** and **Alberto** – I'm glad I got to know you guys, your enthusiasm has been contagious and you made the work environment so much more fun and pleasant.

To **Katrin** – I miss your kindness, your unfailing smile and your fun company! Thanks for showing me the basics of Ensembl and Primer 3, for the Helsinki trip, for the intimate talks over a cup of tea (and for the German Christmas tea itself!), and for the wonderful friendship.

To **Margarita**, who also started as a project student in the Dahl lab, but unexpectedly became one of my best, closest, most wonderful friends. Thank you for always being there for me, for listening to me going about the same topics over and over again ☺, for confiding in me. Thank you for the useful work-related discussions and for our awesome mini-vacation trips. May we have many more! You make Sweden feel like home; you're practically like my family here.

Thanks to all the **past and present members of the Dahl group** who crossed paths with me – it has been an enlightening experience to work with you.

I would also like to say many thanks to the multiple people who do not belong to our group, but who made my PhD student life incomparably better and more fun: thank you **Ammar**, for being supportive, for listening to me and giving advice at rough times, and for inspiring me to be a fighter.

Thanks to the awesome PhD students and co-members of the Rudbeck masquerade organization committee: **Anna S**., **Chiara**, **Sofia**, **Anna W**., **Viktor**, **Diego**, **Johanna**, **Leire**, **Sara B**., **Gabriel**. It has been so much fun doing this with you guys, making PBS (yeah, "PBS"…), having after-works and simply chatting. I have missed you since I moved to BMC (yes, yes, Sofia, I moved). Thanks, **Lucy**, for proofreading my thesis in an impressively expedite fashion. You rock ☺. Thanks to all the other fellow PhD students at Rudbeck, it has been so great getting to know you!

A big "thank you" goes to **Sara P.G**. Firstly, thanks for the fun iPS journal clubs, but most of all for our fikas and lunches and for the possibility to talk and share. I always feel better after a chat with you, Sara! Best of luck on your own dissertation!

I wholeheartedly thank **Vasil** for being the person I could always rely on (well, except for when I foolishly expect you to come for lunch on time). You have been a great friend, honest, supportive and critical (and annoying as hell) at the same time. Thanks for all the (late!) lunches, for the stimulating scientific discussions and the totally non-scientific goofy talks, for understanding and not judging me. I don't know what I would do without you.

Thanks to **Antonia** for the (even later) lunches and for the shopping sprees (where you buy ten dresses, Vicky gets a black T-shirt and I walk out empty-handed), for the dinners and for the trips. Thanks for your level-headed advice, your understanding and your awesome company. Thanks to **Manoj** for calling me "чичо" and for cooking less-spicy Indian food (which in reality turns one into a fire-breathing dragon) especially for me!

I thank my cousin **Victoria** and her family, **Carl-Olof** and little **Kai**. Thanks for the fun times we've had, the karaoke nights ("you're hot and you're cold"…), for kanelbullar, for introducing me to sushi, for the walks in the rain, for inspiring me to come to Sweden. It's been quite the journey.

My heartfelt gratitude goes also to **Christina Magnusson**. You are a fantastic, big-hearted, extremely efficient and helpful person! I am sure that all the IGP PhD students will agree with me when I say that without you, we're lost. Thanks for your kindness and support, Christina. I also want to say thank you to **Prof. Anna Dimberg** – your patience and help meant a lot to me. Thank you for listening to me, advising me and being there in a tough period.

My thanks extend outside of IGP and Uppsala University as well. I would like to acknowledge my first and biggest teacher who introduced me to genetics, **D-r Svetoslav Dimov**. Without you I would never be where I am now, and any success in science that I have is your success as well. You didn't just introduce me to the basics – you infected me with your enthusi-

asm and creativity, and ignited a lifelong love for genetics in me. Thank you for believing in me and for your friendship throughout the years! Thanks also to **Slavil**, my example for a bright and inquisitive mind. Every conversation with you has been inspiring and stimulating (even if sometimes that meant I was stimulated to throw a rock at you :P). I have learned a great deal from you and the years we worked together created some of the very best memories in my life ("forever and ever").

Thanks to my bestie **Yana**, for being the one constant in my friends circle since the 8th grade. Thanks for all the support, patience and understanding, dear. It's been a long ride but even thousands of kilometers between us could not destroy our friendship. Thank you for laughing and crying with me, for the secrets we shared, for growing up together, for accepting me as I am and not trying to change me. I love you.

And last but not least, my biggest gratitude goes to **my incredible family**. Мамо, тате, бате, в най-буквален смисъл вие сте стълбовете на живота ми и всичко, което имам и което съм, е благодарение на вас. Вие сте моят пример за честност, човечност, интелект и мъдрост, вие ми дадохте жажда да надникна "зад тежката черна завеса" и пак вие ми давате сили да продължавам напред и в най-трудните моменти. Благодаря ви. Гордея се с вас и безкрайно много ви обичам.

References

1. Watson, J.D. and F.H. Crick, *The structure of DNA.* Cold Spring Harb Symp Quant Biol, 1953. **18**: p. 123-31.
2. Consortium, I.H.G.S., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
3. Bernstein, B.E., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
4. Lewontin, R.C., *The Genetic Basis of Evolutionary Change*. 1974: Columbia University Press.
5. Abecasis, G.R., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
6. Abecasis, G.R., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
7. Jakobsson, M., et al., *Genotype, haplotype and copy-number variation in worldwide human populations.* Nature, 2008. **451**(7181): p. 998-1003.
8. Mu, X.J., et al., *Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project.* Nucleic Acids Res, 2011. **39**(16): p. 7058-76.
9. Shen, H., et al., *Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians.* PLoS One, 2013. **8**(4): p. e59494.
10. Barnes, D.E. and T. Lindahl, *Repair and genetic consequences of endogenous DNA base damage in mammalian cells.* Annu Rev Genet, 2004. **38**: p. 445-76.
11. Altshuler, D.M., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.
12. Day, I.N., *dbSNP in the detail and copy number complexities.* Hum Mutat, 2010. **31**(1): p. 2-4.
13. 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.
14. Chen, J.M., C. Ferec, and D.N. Cooper, *Patterns and mutational signatures of tandem base substitutions causing human inherited disease.* Hum Mutat, 2013. **34**(8): p. 1119-30.
15. Chen, J.M., D.N. Cooper, and C. Ferec, *A new and more accurate estimate of the rate of concurrent tandem-base substitution mutations in the human germline: approximately 0.4% of the single-nucleotide substitution mutation rate.* Hum Mutat, 2014. **35**(3): p. 392-4.
16. Grandi, F.C. and W. An, *Non-LTR retrotransposons and microsatellites: Partners in genomic variation.* Mob Genet Elements, 2013. **3**(4): p. e25674.
17. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.

18. Dewannieux, M. and T. Heidmann, *LINEs, SINEs and processed pseudogenes: parasitic strategies for genome modeling.* Cytogenet Genome Res, 2005. **110**(1-4): p. 35-48.
19. Purandare, S.M. and P.I. Patel, *Recombination hot spots and human disease.* Genome Res, 1997. **7**(8): p. 773-86.
20. Pang, A.W., et al., *Towards a comprehensive structural variation map of an individual human genome.* Genome Biol, 2010. **11**(5): p. R52.
21. Katju, V. and U. Bergthorsson, *Copy-number changes in evolution: rates, fitness effects and adaptive significance.* Front Genet, 2013. **4**: p. 273.
22. Krumm, N., et al., *Copy number variation detection and genotyping from exome sequence data.* Genome Res, 2012. **22**(8): p. 1525-32.
23. Mills, R.E., et al., *Natural genetic variation caused by small insertions and deletions in the human genome.* Genome Res, 2011. **21**(6): p. 830-9.
24. Mullaney, J.M., et al., *Small insertions and deletions (INDELs) in human genomes.* Hum Mol Genet, 2010. **19**(R2): p. R131-6.
25. Robinson, W.P., *Mechanisms leading to uniparental disomy and their clinical consequences.* Bioessays, 2000. **22**(5): p. 452-9.
26. Dempfle, A., et al., *Gene-environment interactions for complex traits: definitions, methodological requirements and challenges.* Eur J Hum Genet, 2008. **16**(10): p. 1164-72.
27. Stachan, T.R., A. P., *Human Molecular Genetics.* 2nd ed. 1999, New York: Wiley-Liss.
28. Kennedy, M.A., *Mendelian Genetic Disorders*, in *eLS*. 2001, John Wiley & Sons, Ltd.
29. Brasier, J.L. and E.P. Henske, *Loss of the polycystic kidney disease (PKD1) region of chromosome 16p13 in renal cyst cells supports a loss-of-function model for cyst pathogenesis.* J Clin Invest, 1997. **99**(2): p. 194-9.
30. Peral, B., et al., *A stable, nonsense mutation associated with a case of infantile onset polycystic kidney disease 1 (PKD1).* Hum Mol Genet, 1996. **5**(4): p. 539-42.
31. Qian, F., et al., *The molecular basis of focal cyst formation in human autosomal dominant polycystic kidney disease type I.* Cell, 1996. **87**(6): p. 979-87.
32. Cao, H. and R.A. Hegele, *LMNA is mutated in Hutchinson-Gilford progeria (MIM 176670) but not in Wiedemann-Rautenstrauch progeroid syndrome (MIM 264090).* J Hum Genet, 2003. **48**(5): p. 271-4.
33. Eriksson, M., et al., *Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome.* Nature, 2003. **423**(6937): p. 293-8.
34. Wuyts, W., et al., *Somatic and gonadal mosaicism in Hutchinson-Gilford progeria.* Am J Med Genet A, 2005. **135**(1): p. 66-8.
35. DiLella, A.G., et al., *Molecular structure and polymorphic map of the human phenylalanine hydroxylase gene.* Biochemistry, 1986. **25**(4): p. 743-9.
36. Eisensmith, R.C. and S.L. Woo, *Molecular basis of phenylketonuria and related hyperphenylalaninemias: mutations and polymorphisms in the human phenylalanine hydroxylase gene.* Hum Mutat, 1992. **1**(1): p. 13-23.
37. Akiyama, M., et al., *Novel mutations of TGM1 in a child with congenital ichthyosiform erythroderma.* Br J Dermatol, 2001. **144**(2): p. 401-7.
38. Russell, L.J., et al., *Mutations in the gene for transglutaminase 1 in autosomal recessive lamellar ichthyosis.* Nat Genet, 1995. **9**(3): p. 279-83.
39. Zhao, G.Y., et al., *Chinese family with retinitis pigmentosa.* Ophthalmic Genet, 1995. **16**(2): p. 75-6.

40. Mannucci, P.M. and E.G. Tuddenham, *The hemophilias--from royal genes to gene therapy.* N Engl J Med, 2001. **344**(23): p. 1773-9.
41. Ducamp, S., et al., *Molecular and functional analysis of the C-terminal region of human erythroid-specific 5-aminolevulinic synthase associated with X-linked dominant protoporphyria (XLDPP).* Hum Mol Genet, 2013. **22**(7): p. 1280-8.
42. Whatley, S.D., et al., *C-terminal deletions in the ALAS2 gene lead to gain of function and cause X-linked dominant protoporphyria without anemia or iron overload.* Am J Hum Genet, 2008. **83**(3): p. 408-14.
43. Lyon, M.F., *Gene action in the X-chromosome of the mouse (Mus musculus L.).* Nature, 1961. **190**: p. 372-3.
44. Shirahama, S., et al., *Skewed X-chromosome inactivation causes intra-familial phenotypic variation of an EBP mutation in a family with X-linked dominant chondrodysplasia punctata.* Hum Genet, 2003. **112**(1): p. 78-83.
45. Willard, H.F., *X chromosome inactivation and X-linked mental retardation.* Am J Med Genet, 1996. **64**(1): p. 21-6.
46. Wills, C., *Genetic variability*. 1981, Oxford: Clarendon Press; New York: Oxford University Press.
47. Talsma, A.D., et al., *Genome-wide screen for modifiers of Na + /K + ATPase alleles identifies critical genetic loci.* Mol Brain, 2014. **7**(1): p. 89.
48. Zingg, D., et al., *The epigenetic modifier EZH2 controls melanoma growth and metastasis through silencing of distinct tumour suppressors.* Nat Commun, 2015. **6**: p. 6051.
49. Morton, N.E., *Parameters of the human genome.* Proc Natl Acad Sci U S A, 1991. **88**(17): p. 7474-6.
50. Botstein, D., et al., *Construction of a Genetic-Linkage Map in Man Using Restriction Fragment Length Polymorphisms.* American Journal of Human Genetics, 1980. **32**(3): p. 314-331.
51. Green, P., et al., *Construction of a linkage map of the human genome, and its application to mapping genetic diseases.* Clin Chem, 1989. **35**(7 Suppl): p. B33-7.
52. Morton, N.E., *Sequential tests for the detection of linkage.* Am J Hum Genet, 1955. **7**(3): p. 277-318.
53. Wiszniewska, J., et al., *Combined array CGH plus SNP genome analyses in a single assay for optimized clinical testing.* Eur J Hum Genet, 2014. **22**(1): p. 79-87.
54. LaFramboise, T., *Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances.* Nucleic Acids Res, 2009. **37**(13): p. 4181-93.
55. Ragoussis, J., *Genotyping technologies for genetic research.* Annu Rev Genomics Hum Genet, 2009. **10**: p. 117-33.
56. Liu, L., et al., *Comparison of next-generation sequencing systems.* J Biomed Biotechnol, 2012. **2012**: p. 251364.
57. Ansorge, W.J., *Next-generation DNA sequencing techniques.* N Biotechnol, 2009. **25**(4): p. 195-203.
58. Morozova, O. and M.A. Marra, *Applications of next-generation sequencing technologies in functional genomics.* Genomics, 2008. **92**(5): p. 255-64.
59. Gupta, P.K., *Single-molecule DNA sequencing technologies for future genomics research.* Trends Biotechnol, 2008. **26**(11): p. 602-11.
60. Su, Z., et al., *Next-generation sequencing and its applications in molecular diagnostics.* Expert Rev Mol Diagn, 2011. **11**(3): p. 333-43.

61. Rothberg, J.M. and J.H. Leamon, *The development and impact of 454 sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1117-24.
62. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
63. Voelkerding, K.V., S.A. Dames, and J.D. Durtschi, *Next-generation sequencing: from basic research to diagnostics.* Clin Chem, 2009. **55**(4): p. 641-58.
64. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
65. Zhao, J. and S.F. Grant, *Advances in whole genome sequencing technology.* Curr Pharm Biotechnol, 2011. **12**(2): p. 293-305.
66. Quail, M.A., et al., *A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.* BMC Genomics, 2012. **13**: p. 341.
67. Zhou, X., et al., *The next-generation sequencing technology and application.* Protein Cell, 2010. **1**(6): p. 520-36.
68. Nowrousian, M., *Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems.* Eukaryot Cell, 2010. **9**(9): p. 1300-10.
69. Jia, P. and Z. Zhao, *Personalized pathway enrichment map of putative cancer genes from next generation sequencing data.* PLoS One, 2012. **7**(5): p. e37595.
70. Cullum, R., O. Alder, and P.A. Hoodless, *The next generation: using new sequencing technologies to analyse gene regulation.* Respirology, 2011. **16**(2): p. 210-22.
71. Mutz, K.O., et al., *Transcriptome analysis using next-generation sequencing.* Curr Opin Biotechnol, 2013. **24**(1): p. 22-30.
72. Stenson, P.D., et al., *The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics.* Hum Genomics, 2009. **4**(2): p. 69-72.
73. Prucka, S.K., et al., *An update to returning genetic research results to individuals: perspectives of the industry pharmacogenomics working group.* Bioethics, 2015. **29**(2): p. 82-90.
74. Ayuso, C., et al., *Informed consent for whole-genome sequencing studies in the clinical setting. Proposed recommendations on essential content and process.* Eur J Hum Genet, 2013. **21**(10): p. 1054-9.
75. McGuire, A.L., T. Caulfield, and M.K. Cho, *Research ethics and the challenge of whole-genome sequencing.* Nat Rev Genet, 2008. **9**(2): p. 152-6.
76. van El, C.G., et al., *Whole-genome sequencing in health care: recommendations of the European Society of Human Genetics.* Eur J Hum Genet, 2013. **21**(6): p. 580-4.
77. Greely, H.T., *Breaking the stalemate: a prospective regulatory framework for unforseen research uses of human tissue samples and health information.* Wake Forest Law Rev, 1999. **34**(3): p. 737-66.
78. Majewski, J., et al., *What can exome sequencing do for you?* J Med Genet, 2011. **48**(9): p. 580-9.
79. Pruitt, K.D., et al., *NCBI Reference Sequences: current status, policy and new initiatives.* Nucleic Acids Res, 2009. **37**(Database issue): p. D32-6.
80. Bamshad, M.J., et al., *Exome sequencing as a tool for Mendelian disease gene discovery.* Nat Rev Genet, 2011. **12**(11): p. 745-55.
81. Gonzaga-Jauregui, C., J.R. Lupski, and R.A. Gibbs, *Human genome sequencing in health and disease.* Annu Rev Med, 2012. **63**: p. 35-61.

82. Gilissen, C., et al., *Disease gene identification strategies for exome sequencing.* Eur J Hum Genet, 2012. **20**(5): p. 490-7.
83. Ku, C.S., N. Naidoo, and Y. Pawitan, *Revisiting Mendelian disorders through exome sequencing.* Hum Genet, 2011. **129**(4): p. 351-70.
84. Ku, C.S., et al., *Exome sequencing: dual role as a discovery and diagnostic tool.* Ann Neurol, 2012. **71**(1): p. 5-14.
85. Ng, S.B., et al., *Exome sequencing identifies the cause of a mendelian disorder.* Nat Genet, 2010. **42**(1): p. 30-5.
86. Ng, P.C. and E.F. Kirkness, *Whole genome sequencing.* Methods Mol Biol, 2010. **628**: p. 215-26.
87. Nagalakshmi, U., K. Waern, and M. Snyder, *RNA-Seq: a method for comprehensive transcriptome analysis.* Curr Protoc Mol Biol, 2010. **Chapter 4**: p. Unit 4 11 1-13.
88. Mardis, E.R., *Next-generation DNA sequencing methods.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 387-402.
89. de Klerk, E., J.T. den Dunnen, and P.A. t Hoen, *RNA sequencing: from tag-based profiling to resolving complete transcript structure.* Cell Mol Life Sci, 2014. **71**(18): p. 3537-51.
90. Sharon, D., et al., *A single-molecule long-read survey of the human transcriptome.* Nat Biotechnol, 2013. **31**(11): p. 1009-14.
91. Hansen, K.D., S.E. Brenner, and S. Dudoit, *Biases in Illumina transcriptome sequencing caused by random hexamer priming.* Nucleic Acids Res, 2010. **38**(12): p. e131.
92. Roberts, A., et al., *Improving RNA-Seq expression estimates by correcting for fragment bias.* Genome Biol, 2011. **12**(3): p. R22.
93. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.
94. Levin, J.Z., et al., *Comprehensive comparative analysis of strand-specific RNA sequencing methods.* Nat Methods, 2010. **7**(9): p. 709-15.
95. Sievers, F., et al., *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Mol Syst Biol, 2011. **7**: p. 539.
96. Flanagan, S.E., A.M. Patch, and S. Ellard, *Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations.* Genet Test Mol Biomarkers, 2010. **14**(4): p. 533-7.
97. Schwarz, J.M., et al., *MutationTaster evaluates disease-causing potential of sequence alterations.* Nat Methods, 2010. **7**(8): p. 575-6.
98. Krijgsveld, J., et al., *Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics.* Nat Biotechnol, 2003. **21**(8): p. 927-31.
99. Kaletta, T. and M.O. Hengartner, *Finding function in novel targets: C. elegans as a model organism.* Nat Rev Drug Discov, 2006. **5**(5): p. 387-98.
100. Artal-Sanz, M., L. de Jong, and N. Tavernarakis, *Caenorhabditis elegans: a versatile platform for drug discovery.* Biotechnol J, 2006. **1**(12): p. 1405-18.
101. Bingham, P.M., M.G. Kidwell, and G.M. Rubin, *The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family.* Cell, 1982. **29**(3): p. 995-1004.
102. White, R.M., et al., *Transparent adult zebrafish as a tool for in vivo transplantation analysis.* Cell Stem Cell, 2008. **2**(2): p. 183-9.
103. Mayden, R.L., et al., *Phylogenetic relationships of Danio within the order Cypriniformes: a framework for comparative and evolutionary studies of a model species.* J Exp Zool B Mol Dev Evol, 2007. **308**(5): p. 642-54.
104. Rosen, J.N., M.F. Sweeney, and J.D. Mably, *Microinjection of zebrafish embryos to analyze gene function.* J Vis Exp, 2009(25).

105. Hardouin, S.N. and A. Nagy, *Mouse models for human disease.* Clin Genet, 2000. **57**(4): p. 237-44.
106. Erickson, R.P., *Mouse models of human genetic disease: which mouse is more like a man?* Bioessays, 1996. **18**(12): p. 993-8.
107. Harper, A.J., *Production of transgenic and mutant mouse models.* Methods Mol Med, 2005. **104**: p. 185-202.
108. Ferreira, D., F. Adega, and R. Chaves, *The Importance of Cancer Cell Lines as in vitro Models in Cancer Methylome Analysis and Anticancer Drugs Testing.* 2013.
109. Domcke, S., et al., *Evaluating cell lines as tumour models by comparison of genomic profiles.* Nat Commun, 2013. **4**: p. 2126.
110. Thomson, J.A., et al., *Embryonic stem cell lines derived from human blastocysts.* Science, 1998. **282**(5391): p. 1145-7.
111. Thomas, K.R. and M.R. Capecchi, *Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells.* Cell, 1987. **51**(3): p. 503-12.
112. Odorico, J.S., D.S. Kaufman, and J.A. Thomson, *Multilineage differentiation from human embryonic stem cell lines.* Stem Cells, 2001. **19**(3): p. 193-204.
113. Weissman, I.L., *Translating stem and progenitor cell biology to the clinic: barriers and opportunities.* Science, 2000. **287**(5457): p. 1442-6.
114. Nagy, A., et al., *Derivation of completely cell culture-derived mice from early-passage embryonic stem cells.* Proc Natl Acad Sci U S A, 1993. **90**(18): p. 8424-8.
115. Takahashi, K. and S. Yamanaka, *Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.* Cell, 2006. **126**(4): p. 663-76.
116. Yamanaka, S. and H.M. Blau, *Nuclear reprogramming to a pluripotent state by three approaches.* Nature, 2010. **465**(7299): p. 704-12.
117. Takahashi, K., et al., *Induction of pluripotent stem cells from adult human fibroblasts by defined factors.* Cell, 2007. **131**(5): p. 861-72.
118. Yu, J., et al., *Induced pluripotent stem cell lines derived from human somatic cells.* Science, 2007. **318**(5858): p. 1917-20.
119. Nakagawa, M., et al., *Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts.* Nat Biotechnol, 2008. **26**(1): p. 101-6.
120. Huangfu, D., et al., *Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds.* Nat Biotechnol, 2008. **26**(7): p. 795-7.
121. Kamata, M., et al., *Live cell monitoring of hiPSC generation and differentiation using differential expression of endogenous microRNAs.* PLoS One, 2010. **5**(7): p. e11834.
122. Anokye-Danso, F., et al., *Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency.* Cell Stem Cell, 2011. **8**(4): p. 376-88.
123. Zhou, H., et al., *Generation of induced pluripotent stem cells using recombinant proteins.* Cell Stem Cell, 2009. **4**(5): p. 381-4.
124. Kim, D., et al., *Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins.* Cell Stem Cell, 2009. **4**(6): p. 472-6.
125. Sommer, C.A., et al., *Induced pluripotent stem cell generation using a single lentiviral stem cell cassette.* Stem Cells, 2009. **27**(3): p. 543-9.
126. Sommer, C.A., et al., *Excision of reprogramming transgenes improves the differentiation potential of iPS cells generated with a single excisable vector.* Stem Cells, 2010. **28**(1): p. 64-74.

127. Belay, E., et al., *Transposon-mediated gene transfer into adult and induced pluripotent stem cells.* Curr Gene Ther, 2011. **11**(5): p. 406-13.
128. Lieu, P.T., et al., *Generation of induced pluripotent stem cells with CytoTune, a non-integrating Sendai virus.* Methods Mol Biol, 2013. **997**: p. 45-56.
129. Fusaki, N., et al., *Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome.* Proc Jpn Acad Ser B Phys Biol Sci, 2009. **85**(8): p. 348-62.
130. Zhao, X.Y., et al., *iPS cells produce viable mice through tetraploid complementation.* Nature, 2009. **461**(7260): p. 86-90.
131. Zhao, T., et al., *Immunogenicity of induced pluripotent stem cells.* Nature, 2011. **474**(7350): p. 212-5.
132. Bock, C., et al., *Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines.* Cell, 2011. **144**(3): p. 439-52.
133. Doi, A., et al., *Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts.* Nat Genet, 2009. **41**(12): p. 1350-3.
134. Tchieu, J., et al., *Female human iPSCs retain an inactive X chromosome.* Cell Stem Cell, 2010. **7**(3): p. 329-42.
135. Miyanari, Y. and M.E. Torres-Padilla, *Control of ground-state pluripotency by allelic regulation of Nanog.* Nature, 2012. **483**(7390): p. 470-3.
136. Gafni, O., et al., *Derivation of novel human ground state naive pluripotent stem cells.* Nature, 2013. **504**(7479): p. 282-6.
137. Pomp, O., et al., *Unexpected X chromosome skewing during culture and reprogramming of human somatic cells can be alleviated by exogenous telomerase.* Cell Stem Cell, 2011. **9**(2): p. 156-65.
138. Tomoda, K., et al., *Derivation conditions impact X-inactivation status in female human induced pluripotent stem cells.* Cell Stem Cell, 2012. **11**(1): p. 91-9.
139. Soldner, F., et al., *Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations.* Cell, 2011. **146**(2): p. 318-31.
140. Murray, A., et al., *Isogenic Induced Pluripotent Stem Cell Lines from an Adult with Mosaic Down Syndrome Model Accelerated neuronal Ageing and Neurodegeneration.* Stem Cells, 2015.
141. Miyaoka, Y., et al., *Isolation of single-base genome-edited human iPS cells without antibiotic selection.* Nat Methods, 2014. **11**(3): p. 291-3.
142. Yang, L., et al., *CRISPR-Cas-mediated targeted genome editing in human cells.* Methods Mol Biol, 2014. **1114**: p. 245-67.
143. Cai, M. and Y. Yang, *Targeted genome editing tools for disease modeling and gene therapy.* Curr Gene Ther, 2014. **14**(1): p. 2-9.
144. Barker, R.A. and I. de Beaufort, *Scientific and ethical issues related to stem cell research and interventions in neurodegenerative disorders of the brain.* Prog Neurobiol, 2013. **110**: p. 63-73.
145. Robertson, J.A., *Human embryonic stem cell research: ethical and legal issues.* Nat Rev Genet, 2001. **2**(1): p. 74-8.
146. Ishii, T., R.A. Pera, and H.T. Greely, *Ethical and legal issues arising in research on inducing human germ cells from pluripotent stem cells.* Cell Stem Cell, 2013. **13**(2): p. 145-8.
147. McLaren, A., *Ethical and social considerations of stem cell research.* Nature, 2001. **414**(6859): p. 129-31.

148. Robertson, J.A., *Embryo stem cell research: ten years of controversy.* J Law Med Ethics, 2010. **38**(2): p. 191-203.
149. Klimanskaya, I., et al., *Human embryonic stem cell lines derived from single blastomeres.* Nature, 2006. **444**(7118): p. 481-5.
150. Hayashi, K., et al., *Offspring from oocytes derived from in vitro primordial germ cell-like cells in mice.* Science, 2012. **338**(6109): p. 971-5.
151. Hayashi, K., et al., *Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells.* Cell, 2011. **146**(4): p. 519-32.
152. Easley, C.A.t., et al., *Direct differentiation of human pluripotent stem cells into haploid spermatogenic cells.* Cell Rep, 2012. **2**(3): p. 440-6.
153. Hongisto, H., et al., *Laminin-511 expression is associated with the functionality of feeder cells in human embryonic stem cell culture.* Stem Cell Res, 2012. **8**(1): p. 97-108.
154. Oti, M. and H.G. Brunner, *The modular nature of genetic diseases.* Clin Genet, 2007. **71**(1): p. 1-11.
155. Gilissen, C., et al., *Unlocking Mendelian disease using exome sequencing.* Genome Biol, 2011. **12**(9): p. 228.
156. De Sandre-Giovannoli, A., et al., *Lamin a truncation in Hutchinson-Gilford progeria.* Science, 2003. **300**(5628): p. 2055.
157. Gordon, L.B., et al., *Clinical trial of a farnesyltransferase inhibitor in children with Hutchinson-Gilford progeria syndrome.* Proc Natl Acad Sci U S A, 2012. **109**(41): p. 16666-71.
158. Ullrich, N.J., et al., *Neurologic features of Hutchinson-Gilford progeria syndrome after lonafarnib treatment.* Neurology, 2013. **81**(5): p. 427-30.
159. McClintock, D., et al., *The mutant form of lamin A that causes Hutchinson-Gilford progeria is a biomarker of cellular aging in human skin.* PLoS One, 2007. **2**(12): p. e1269.
160. Merideth, M.A., et al., *Phenotype and course of Hutchinson-Gilford progeria syndrome.* N Engl J Med, 2008. **358**(6): p. 592-604.
161. Wasif, N., et al., *Novel mutations in the keratin-74 (KRT74) gene underlie autosomal dominant woolly hair/hypotrichosis in Pakistani families.* Hum Genet, 2011. **129**(4): p. 419-24.
162. Shimomura, Y., et al., *Autosomal-dominant woolly hair resulting from disruption of keratin 74 (KRT74), a potential determinant of human hair texture.* Am J Hum Genet, 2010. **86**(4): p. 632-8.
163. Peltonen, L., et al., *Lessons from studying monogenic disease for common disease.* Hum Mol Genet, 2006. **15 Spec No 1**: p. R67-74.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1078

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-246228

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2015