1

General Introduction

General Introduction

The human genome consists of ~ 6 billion base pairs, which is divided over 23 pairs of chromosomes. It is estimated that there are 20000-25000 protein coding genes. The DNA sequence in our genome is on average 99.9% identical to any other human being [1]. The more closely related two people are, the more similar their genomes. However, every human being is genetically unique and variations in composition and structure of the DNA are found throughout the genome. Human genetic variation refers to genetic differences between individuals and is important for diversity in a population. Without genetic variability a population cannot adapt to changes in the environment. The differences in genotype can cause differences in phenotype with subtle or sometimes quite obvious effects. Despite the large amount of variation in the human genome, most sequence variants have no obvious functional consequences. Determining if a specific variant has an effect is an important part of genetic analysis. The identification of genomic variation in large numbers of individuals helps to distinguish neutral variants (not involved in disease, or 'non-pathogenic' variants) from variants disrupting gene function (involved in disease, or 'pathogenic variants'). DNA analysis of patients with Mendelian disorders has resulted in the identification of a broad range of variants in genes, from definitely pathogenic mutations, to unclassified variants, to neutral polymorphisms. The relation between genomic variation and complex and quantitative human traits (i.e. obesity, height, multifactorial disease) has also been studied extensively.

As new methods for studying DNA are developed, different types of genomic variation have been discovered. Detection of large numbers of unclassified variants (UVs), with unclear significance for disease, have further emphasized the importance of cataloging genomic variation and studying the functional effects. The major challenge for molecular geneticists, cytogeneticists, clinical geneticists, and genetic counsellors is assessing the impact of all types of genomic variation on monogenic and complex disorders, along with the effective communication of findings to counselees.

Genomic variation; definition of sequence variation and structural variation

Genomic variation is a general term that covers all types of possible DNA variants, ranging from alterations affecting entire chromosomes to single nucleotide changes.

Differences in classification and terminology of genomic variation often cause confusion, therefore recommendations for the description of DNA variants have been proposed by the Human Genome Variation Society (HGVS).

Two main groups of variation can be identified; 1) sequence variation and 2) structural variation. An overview of the different types of variation, definitions and the spectrum in which they operate are given in Table 1a+b. Examples are shown in Figure 1a+b.

Sequence variants can be classified as single or multi-nucleotide changes and range from single nucleotide differences to 1 kilobase (kb)-sized changes to full chromosome or even full genome changes of a segment of DNA [2]. Single nucleotide changes affect only one base pair, whereas multi-nucleotide changes are changes in a stretch of nucleotides. Variation in sequence include substitutions, insertions, deletions, duplications, and inversions. Indels are more complex changes and can be regarded as a combination of single nucleotide changes and multi-nucleotide changes. Sequence variations include "mutations" and "polymorphisms" and are usually referred to as single nucleotide polymorphisms (SNPs) or single nucleotide variants (SNVs). In some disciplines the term **"**mutation**"** is used to indicate "a change", while in other disciplines it is used to indicate "a disease-causing change". Similarly, the term **"**polymorphism**"** is used both to indicate "a non-disease-causing change", a change involved in human traits or a change found at a frequency of 1% or higher in the population [3]. To prevent this confusion, neutral terms such as **"** variant**"** or "alteration" are preferred as they are less ambiguous [4].

Structural variation is the genetic variation in structure of an organism's chromosome. It is generally defined as a region of DNA 1 kb and larger in size. As the resolving power of genetic analysis has increased, the focus of structural variation has shifted from entire chromosomes (in the prebanding chromosome era), parts of chromosomes (in the G banding era) to kilobases (using restriction enzymes, DNA probes and the Southern blotting). Sequencing techniques have shown that structural variants also include much smaller events, and overlap with the spectrum of sequence variation. Structural variation includes cytogenetically detectable and submicroscopic types of variation, such as deletions, duplications, insertions, inversions, translocations, indels and transpositions. Deletions and duplications, collectively referred to as copy-number variation (CNV), are a subset of structural variation and result in variable copy numbers of copies of specific DNA sequences [5-11]. CNVs are a major source of human genetic variation. Over 10,000 distinct CNVs have been described, ranging in size from kilobases to megabases. Most CNVs in humans are <50 Kb in size [12]. CNVs can be defined as recurrent or nonrecurrent, depending on their mechanism of formation [13]. Many recurrent CNVs are flanked by segmental duplications (also called low copy repeats; regions of DNA >1 kb present more than once in the genome with copies which are >90% identical) and are of a fixed size [9, 14]. Because these repeated sequences tend to misalign during meiosis, the resultant rearrangements tend to recur, creating clusters of variants with common endpoints. CNVs at these loci arise, by a mechanism named nonallelic homologous recombination (NAHR). Nonrecurrent CNVs, in contrast, which are not flanked by low copy repeats but other DNA elements (ALU elements or other repetitive elements), are of variable size and are thought to arise via mechanisms like nonhomologous end joining (NHEJ) and replication-based mechanisms such as fork stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication models (MMRDR) [15]. It is becoming clear that most disease-causing CNVs are nonrecurrent, and generally arise via replication-based mechanisms.

Structural variants are associated with repetitive DNA, making accurate characterization more difficult [16]. Systematic assessment of structural variation in our genome has been difficult, primarily due to lack of appropriate methods for analysis. As such, the nucleotide resolution architecture of most structural variants remains unknown.

Table 1a: Description of variation types based on HGVS definitions

Types of variation	Description
Substitution	One nucleotide is replaced by another nucleotide
Deletion	One or more nucleotides are removed
Duplication	A copy of one or more nucleotides is inserted elsewhere in the genome
Tandem duplication	A copy of one or more nucleotides, directly following the original sequence
Insertion	One or more nucleotides are inserted between two original nucleotides but the insertion is not a tandem duplication
Insertion-deletion *(Indel)*	One nucleotide is replaced by more than one other nucleotide or More than one nucleotide is replaced by one or more other nucleotides
Inversion	More than one nucleotide is the reverse complement of the original sequence and replaces the original sequence
Translocation	The sequence of one chromosome interchanges with the sequence of another chromosome
Transposition *(interchromosomal insertion)*	The sequence of one chromosome inserts into another chromosome
Copy Number Variation	Submicroscopic deletions and duplications, which are losses or gains of DNA segments
Conversion	A range of nucleotides replacing the original sequence and are a copy of a sequence elsewhere in the genome

Table 1b: The spectrum of variation in the human genome

Variation	Type	Size Range
Single base pair change	Substitution, deletion, insertion, duplication	1 bp
Multiple base pair changes	Insertions, deletions, duplications, inversions, indels	Up to 1 kb
Full chromosome changes	aneuploidy	Entire chromosome/genome
Fine and intermediate scale structural variation	Insertions, deletions, duplications, inversions, indels, transpositions	1 kb-50 kb
Large scale structural variation	Insertions, deletions, duplications, inversions, indels, transpositions	50 kb-5 Mb
Chromosomal variation	translocations	~ ≥5 Mb

Light grey = sequence variation, black= structural variation. The operational spectrum partially overlaps with respect to size range.

Figure 1a: Types of variation in the human genome (sequence view). These can be single base changes (substitution, deletions, and insertions) or involve larger segments of DNA (deletions, insertions, inversions, indels, and duplications). Adapted from Frazer et al [17]

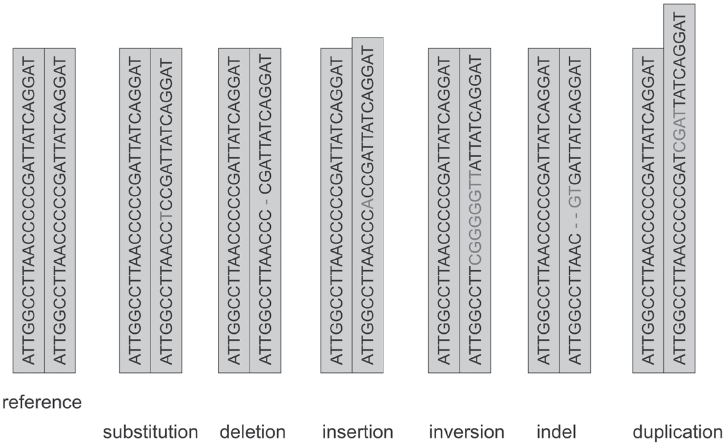
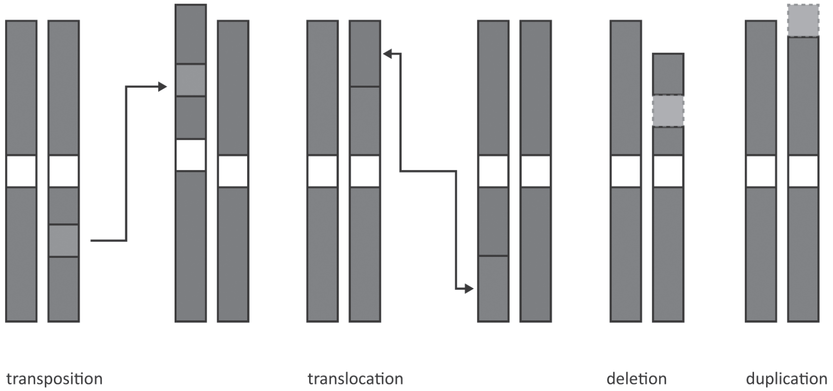

ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCTCCGATTATCAGGAT
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCC - CGATTATCAGGAT
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCCACCGATTATCAGGAT
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTCGGGGTTATTATCAGGAT
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAAC - - GTGATTATCAGGAT
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCCCCGATCGATTATCAGGAT
reference
substitution
deletion
insertion
inversion
indel
duplication

Figure 1b: Types of variation in the human genome (chromosome view). Examples of structural variation. 1) transposition-transfer of segment of DNA to a new position. 2) translocation-balanced event where two DNA segments are interchanged. 3) Copy number variation: deletion and duplication-loss or gain of DNA segments.



Genomic variation in the general population

Genomic variation refers to alterations at the DNA-level. Variation in DNA occurs due to malfunction of DNA replication during cell division or if DNA repair mechanisms fail after DNA damage induced by chemicals or radiation. These are random and spontaneous changes. Recent parent to offspring studies determined the human mutation rate (the rate at which variation occurs) at 1 x 10^{-8} per base per generation [18].

Population genetics is based on the study of genomic variation in natural selection. If a variant increases fitness it will undergo positive selection, and eventually be conserved in the genome. In contrast, a variant that has a negative effect will be selected against, and eventually lost. Our genome mutates spontaneously and randomly; mostly neutral, sometimes detrimental and, very rarely, beneficial. Ultimately, a gene pool arises which is best adapted to the environment. Most common genetic variants arose once in human history, and are shared by many individuals today through descent from common ancestors. Most analysis estimate that SNVs occur 1 in 1000 base pairs, although they do not occur at a uniform density. On average, every human being has 3 million nucleotide differences (SNVs) with any other human. Everyone is genetically unique. Even monozygotic twins, derived from the same fertilization event, have infrequent genetic differences due to early somatic changes (somatic variants) [19, 20]. There is also significant genetic difference between individuals from different ethnic backgrounds and numbers may increase if indels and structural variation is taken into account [21]. It has become apparent that human genomes differ more as a consequence of structural variation than single base-pair differences [9]. In 2010 the human genome contains an estimated 15 million SNVs, 1 million short insertions and deletions and 20000 structural variants [1].

CNVs have been recognized as a common form of genomic structural variation. High resolution microarrays and sequencing approaches are able to identify 600–900 CNVs in a single individual [22, 23]. Approximately 65% to 80% of individuals carry a CNV that is at least 100 kbp in size, 5 to 10 % of individuals harbour a CNV at least 500 kbp, and 1% of individuals carry a large CNV of at least 1 Mbp in size [24]. This means that larger CNVs are skewed toward rare variants. As the full extent of structural variation in our genome has been revealed, it has been estimated that CNVs account for ~ 13% of the human genome [7, 25]. The *de novo* rate of large (> 100 kb) CNV formation in humans was estimated at 1.2×10^{-2} CNVs per genome per transmission against a high selection pressure, suggesting that each of these *de novo* CNVs persists in the population for only a few generations [26, 27]

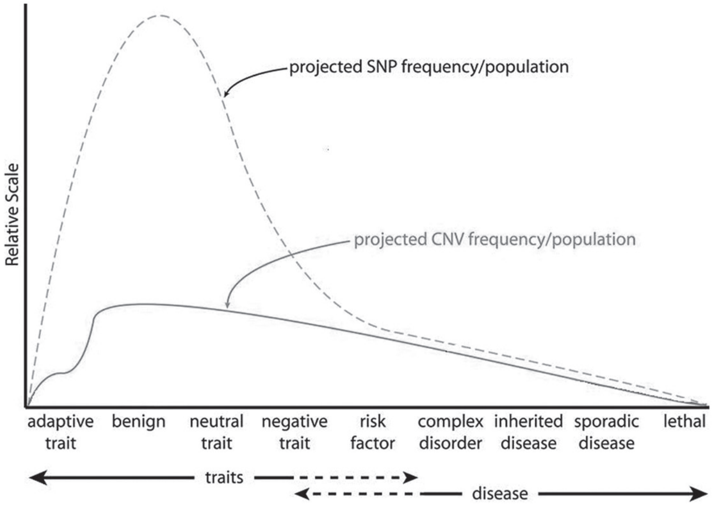
The full extent to which CNVs are likely to contribute to the diversity of human phenotypes is still under assessment. It is clear that the phenotypic impact of CNVs occurs as a continuum from 'neutral' to pathogenic, and can act in more complex and sometimes unexpected ways. Figure 2 shows conceptual curves of projected frequencies of SNVs and CNVs in the population associated with their phenotypic impact. In general, SNVs and CNVs with a high population frequency are annotated as benign or neutral in their effect, while rare variants are more likely to be pathogenic. In between, CNVs and SNVs can be associated with degrees of function described as 'traits', 'risk factors' and, beyond a certain threshold, 'disease'. After extensive phenotype-genotype studies, some of the SNVs and CNVs previously annotated as benign or neutral in their effect will be reclassified as predisposing risk factors.

Uncovering the genetic basis of human phenotypic differences requires a comprehensive understanding of all forms of genetic variation, both at fine scale (sequence variants) and large scale (structural variants). Different genomic technologies are required to detect structural variation at different levels. Next generation sequencing technologies will be used to generate comprehensive maps of human genetic variation.

Effects of genomic variation

The decoding of information from DNA to protein begins with transcription, as messenger RNA (mRNA) is created from a DNA template, followed by translation. Translation is the process of protein synthesis from mRNA, with specific amino acids encoded by three nucleotide combinations (codons). Synthesis of the protein takes place in a specific reading frame according to these codons. As such, each type of variant may have consequences at different levels (DNA, RNA, protein), also depending on the surrounding genomic context

Figure 2: Conceptual curves of SNV and CNV characteristics. Projected frequencies in the population for SNVs (dashed grey) and CNVs (blue) show the relation between genomic variation and a spectrum of phenotypic impact. Adapted from Buchanan et al [28].



(coding or non-coding). Changes in coding or regulatory sequences are most likely to affect gene expression or affect the function of protein products [29]. Any phenotypic effect is highly dependent on the location (type of tissue) and the developmental stage in which the sequence variant is expressed.

Classification of genomic variation can also be based on the effect at different levels (DNA, RNA, protein). In diploid organisms (such as humans), changes in the DNA may occur on one (heterozygous) or both (homozygous) alleles. Single nucleotide changes that lead to amino acid changes in the protein coding region of a gene may be classified into three types (silent, missense and nonsense), depending upon what the erroneous codon codes for. Synonymous variants (also known as silent variants) do not lead to a different amino acid being encoded (due to redundancy in the codon code). Non synonymous variants do alter the amino acid sequence of a protein, and can be sub-classified such as missense variants (encoding a different amino acid) and nonsense variants (creating a stop codon, leading to premature protein truncation). Frameshift variants (variants disrupting the reading frame such as deletions and duplications) lead to an altered, usually non-functional, protein product.

On RNA level any step of gene expression, post-transcriptional modification (capping, polyadenylation, splicing) may be modulated. Most human genes can be transcribed into different mRNAs composed of different exons, leading to alternative splicing and the expression of functionally diverse protein isoforms [30]. A significant fraction of exonic variants (including silent and missense variants) and intronic variants cause disease by disrupting normal splicing [31]. Silent changes may affect translational efficiency, resulting in different protein levels or affect protein function via folding [32, 33].

Variation can also be classified by effect on function. Variants can have no effect (neutral), or lead to a change in function. Loss-of-function changes result in non-functional proteins. When the dose of a gene product is affected and not enough for a normal function this is called haploinsufficiency or dosage effect. Gain-of-function changes yield a protein that has a new function. Dominant negative changes have an altered gene product that has a negative impact on normal function. Variants that prevent viability are termed lethal.

Changes not to the DNA sequence itself but modifications or addition of chemical groups to individual nucleotides (e.g. methyl groups or proteins introduce another type of variation called epigenetics. This type of variation may have an effect on phenotype by altering gene expression. Comparable to genomic variation, epigenetic differences between individuals and monozygotic twins have been described [34, 35]. However, the epigenome is not the subject of this thesis and will not be discussed further.

Genomic variation in relation to disease: intellectual disability and/or congenital anomalies

The number of known pathogenic variants in human genes that underlie or are associated with human inherited disease is > 110000, in more than 4000 genes [36]. Currently, causal variants for ~ 3000 Mendelian disorders have been reported in the online database for Mendelian disorders in man (OMIM May 2012, # entries for phenotype description, molecular basis known).

Mental retardation (MR) or intellectual disability (ID) is defined as a significant impairment of cognitive and adaptive functions [37] and affects around 1-3% of individuals [38, 39]. The proportion of intellectual disability that can be at least partially accounted for by genetic factors (chromosomal, monogenic or multifactorial) is difficult to estimate, but may account for ~ 30% of all cases. In approximately 50% of cases, the cause remains unknown [40]. G-banded karyotyping supplemented by (sub)telomere screening or fluorescence in-situ hybridization (FISH) detects significant abnormalities in up to 10% of intellectual disability cases [41]. ~ 5% of patients have a monogenic disorder, where a causative variant in a single gene can be identified [42]. CNVs (> 500 kb) detected by array-based technologies (aCGH, SNP array) explain another ~ 15% of cases [43]. *De novo* CNVs and point pathogenic variants of large effect may explain the majority of all intellectual disability cases in the population, and could thereby explain why such a disorder with reduced fecundity remains present in the population [44].

In recent years, several new monogenic disease genes and numerous submicroscopic deletion and duplication syndromes ('genomic disorders') have been identified. It has been estimated that about 0.7-1 per 1000 live births has a genomic disorder [45] . In general, CNVs implicated in genomic disorders are *de novo* and large in size (> 50 kb). However, it seems likely that additional genomic disorders due to much smaller *de novo* CNVs remain to be discovered [46].

In addition to microdeletion/microduplication syndromes, CNV's are involved in many common complex traits, including autism and schizophrenia [47-51]. Several genomic disorders involving inherited CNVs have been described (1q11.2, 1q21.1, 15q11, 16p13.11, 16p11.2, 22q11). Microdeletions and microduplications at these loci show variable penetrance and expression, and are thus not necessarily recognised as clinically distinctive syndromes. Inherited CNVs may be involved in disease, by acting as modifiers in milder phenotypes, either in combination with other CNVs or with sequence variants. This is illustrated by the fact that 25% of ID children carry a 2nd CNV, in addition to an inherited CNV [52]. Some authors have coined the term 'second site' or 'two hit' model for this phenomenon, which is confusing since those terms are widely used in the cancer field to indicate somatic mutations disturbing the regulation of cell growth. We prefer to stick to 'multifactorial' a term formerly used to indicate any number of unknown factors above one, but can be used as well for known factors. Recently, compound inheritance of a null allele (deletion 1q21.1) together with the presence of a SNV was proven to be associated with TAR syndrome [53]. The effect of different combinations of (multiple) inherited CNVs and/or sequence variants on phenotype urgently requires further study, and identifying additional genetic and/or environmental factors is the challenge of our time.

Genetic inheritance of genomic variation

Mendelian inheritance patterns can demonstrate a greater level of complexity than simple dominant, recessive or X-linked inheritance, through a number of non-Mendelian processes including imprinting, X-inactivation, mosaicism and variation in penetrance and variable expressivity. When searching for the exact genetic cause of a disease, and calculating risk of transmission to offspring, it is important to take this level of complexity into account. Examples of complex inheritance includes disorders caused by more than one gene (di/polygenic disorders such as in Axenfeld-Rieger syndrome [54]), trinucleotide repeat expansion disorders (such as Huntington's disease [55]) where the number of CAG repeats correlate with the severity of disease and the age of onset in combination with the characteristic of anticipation (the tendency for progressively earlier or more severe expression of the disease in successive generations), genes whose expression is governed by parent of origin (Angelman/Prader Willi syndrome [56], Beckwith-Wiedemann/Silver Russell syndrome [57]), triallelic inheritance (Bardet-Biedl syndrome [58]). Complex inheritance involving co-inheritance of CNVs [52] or a CNV in combination with sequence variants [59, 53] have also been demonstrated.

Sequence variants and CNVs can be inherited from a parent, or occur *de novo*. From the analysis of complete genome sequences of two parent-offspring ('trios') studies, it has been estimated that each child inherits about 30 to 50 new variants [9, 18]. These new variants can either be germline variants that have arisen during the production of gametes in the parental generation, or be present in a subset of cells from either parent, and represent a germ line mosaic with recurrence risk to subsequent offspring [60]. A significantly greater proportion of new variants is of paternal origin, as a result of the larger number of divisions during spermatogenesis [61-63], a phenomenon already described by population geneticists more than half a century ago [61]. A similar finding was reported for *de novo* CNVs [64].

Techniques to detect genomic variation

History

Within the past 40 years, a variety of experimental methods have emerged; typically each focuses on a particular class of genomic variation limited by the size range of the events. Early cytogeneticists studying chromosome variation can be legitimately regarded as the first genome pioneers [65]. The establishment of the human chromosome count [66] and the discovery of trisomy 21 in Down syndrome [67] have been the groundwork for studying disease-related genomic variation. Since the 1970's, G-banding and later high resolution chromosome banding techniques, enabled detection of microscopic variation (> 5 Mb). In the 1980's, FISH (Fluorescent *in situ* hybridisation) was developed for microscopic detection of specific structural chromosome abnormalities [68, 69]. At the same time, with the development of molecular markers and application of recombinant DNA techniques, genomic variation at DNA level (sequence variation) was discovered [70]. The relative easy and reliable chain-termination method developed by Sanger soon became the method of choice for sequencing [71]. In the early 90's, Comparative Genome Hybridization (CGH) allowed the detection of submicroscopic structural variation [72]. In the last decade, array-based methods were developed (array-CGH, SNP-arrays) and contributed greatly to our knowledge of structural variation.

Overview of techniques

Quantitative and qualitative changes require different detection methods. At this moment it is not possible to detect all types of genomic variation with just one technique. Each technique has advantages and disadvantages concerning not only specificity, sensitivity, throughput and resolution, but also costs and feasibility in the laboratory. In the field of clinical genetics, many techniques have been used to study our genome. The techniques discussed here are those used in the research comprising this thesis. Table 4 gives an overview of the techniques used for detection of genomic variation. In general, the resolution of a technique depends on its design and is rarely a fixed number. Improvement in resolution mainly depends on probe size and distances between probes.

Cytogenetic and molecular techniques

Karyotyping, using the light microscope to visualize G-banding patterns, enables rapid identification of all chromosomes in a metaphase spread in one view. This enables clonal analysis, but the resolution of a light microscope is limited. Large (> 5 Mb) chromosomal rearrangements such as deletions, duplications, translocations, inversions and insertions can usually be detected. FISH using probes covering the regions affected enables microscopic detection of structural chromosomal abnormalities directly on metaphase chromosomes and interphase nuclei using fluorescently labelled DNA probes. As with karyotyping, routine FISH analysis has a limited resolution (50 kb-2 Mb). Both techniques do not detect single nucleotide variants or small structural variations. The ultimate resolution of the microscopic approach is obtained by Fiber FISH [75] allowing the detection of small deletions and duplications (5 kb-500 kb).

In addition to FISH, array-based methods have been developed. Array comparative genome hybridisation (Array-CGH) platforms are based on the principle of combined hybridisation of two differentially labelled samples to hybridisation targets. As targets, DNA isolated from Bacterial Artificial Chromosomes (BAC clones), or PCR products there of, or synthetically produced long oligonucleotides are spotted onto a glass slide (array). Subsequently, labelled genomic DNA from a test and a reference sample are hybridised to the array and fluorescence intensities are measured. Relative intensity ratios are calculated, with imbalances indicating the presence of copy number variation. The resolution of array-CGH is unlimited but dependent on the spacing of the clones and their insert size. Arrays with 3500 BAC clones (resolution ~ 1Mb) or tiling arrays (33000 BAC clones) with a 10 fold higher resolution are usually used. This resolution does not allow the accurate determination of breakpoints. Custom high-density oligonucleotide arrays are available on demand, allowing the discovery of CNVs down to 500 bp and more precise breakpoint mapping [9]. For high resolution, genome wide analysis of copy number changes,

Table 4: Techniques to detect genomic variation. Table adapted from Schoumans and Ruivenkamp [73], Gijsbers [74].
-= not possible, ±= less suitable, + = sufficient, ++ = appropriate

Genomic variation	Conventional karyotyping (> 5Mb)	locus specific metaphase FISH (50 Kb- 2 Mb)	locus specific interphase FISH (50 Kb- 2 Mb)	array CGH (~ 0,1-1 Mb)	snp array (3-5 snp probes)	MLPA (45-70 bp)	QF-PCR	HRMA (≥ 1 bp)	WES (≥ 1 bp)
balanced translocation	+	++	-	-	-	-	-	-	±
unbalanced translocation	+	++	-	++	++	±	-	-	±
inversion	+	++	-	-	-	-	-	-	±
insertion	-	++	-	-	-	-	-	-	+
complex rearrangement	+	++	-	±	±	±	-	-	-
deletion	±	+	+	++	++	++	++	-	±
duplication	±	-	±	++	++	++	++	-	±
triplication	±	-	±	++	++	+	++	-	±
trisomy	++	++	++	+	+	++	++	-	±
triploidy	++	++	++	-	±	±	++	-	-
monosomy	++	++	++	+	+	++	++	-	±
uniparental disomy	-	-	-	-	+	-	-	-	±
methylation defect	-	-	-	-	-	++	++	+	+
copy neutral LOH	-	-	-	-	+	-	-	-	±
single base pair changes	-	-	-	-	-	±	±	++	++

SNP-based arrays can be used. There are differences in experimental design by different manufacturers, which will not be discussed here. In general, they are single colour assays, designed to be used for SNP genotyping and copy number analysis in one experiment. One DNA sample is hybridised per array, and copy number ratios are determined by clustering the intensities of each probe across many samples. A major advantage is the possibility to detect low level mosaicism (~>15%), uniparental disomies (UPD) and identity-by-descent (IBD) [76-78]. An important limitation of microarrays are the difficulties of probe design in repeat-rich and duplicated regions, with structural variants being strongly correlated to these regions. Smaller CNVs (< 10kb) are also more difficult to detect routinely, which also leads to under-representation in CNV databases. All array-based techniques cannot detect very small structural deletions or duplications and balanced rearrangements (e.g. translocations). Although it is technically possible, SNP arrays are not routinely used to detect single nucleotide variants.

Several PCR-based techniques have been developed to investigate targeted regions for the presence of CNVs. Multiplex ligation-dependent probe amplification (MLPA) [79] is based upon the ligation of two adjacent-annealing oligonucleotides, followed by a quantitative PCR of the ligated products. This technique detects copy number variations of test DNA, where the relative amount of each product is correlated to the copy number of the locus being tested. Normalization using control probes, and normalization against control samples, determines the relative copy number in the test sample. Detection of deletions (ratio threshold under 0.75) and simple duplications (ratio threshold above 1.25) is relatively straightforward, but this technique is less accurate in detecting higher numbers of duplicates or mosaic changes. In addition, using specific probe design at the ligation site it is possible to detect point single nucleotide variants. At the same time, this also means detection of a deletion may be false positive if a SNV is present at, or close to, the ligation site. The advantage of this technique is the flexibility in experimental design, high resolution of CNV detection (45-70 bp) and the possibility to screen large numbers of patients for different loci in a single experiment.

QF-PCR (quantitative PCR) or QMPSF (Quantitative Multiplex PCR of Short Fluorescent Fragments), like MLPA, is a quantitative assay based on PCR amplification of genomic DNA using fluorescently labelled primers. It monitors the amount of product generated during the amplification process compared to a reference. Quantitative differences are used to estimate a relative copy number.

In addition to the previously described techniques that allow detection of quantitative differences, several methods have been developed to detect qualitative differences.

High resolution melting curve analysis (HRMA) is a method used for genotyping, single nucleotide variant scanning and sequence matching [80, 81] and is also suitable for detecting mosaicism. It is based on the dissociation-characteristics of double-stranded DNA during heating. Following PCR amplification of a target region with a DNA binding fluorescent dye (LC-green), the amplicon is melted out by increasing the temperature in the solution. Double stranded DNA will become single stranded and the dye will be released. The specific sequence of the amplicon (primarily GC content and length) determines the melting behaviour. Each amplicon has a unique melting pattern, based on its sequence. DNA with a higher G-C content, whether because of its source or because of SNVs, will have a higher melting temperature than DNA with a higher A-T content. Amplicons can be compared by plotting the change in fluorescent signal against the melting temperature (Tm).

DNA sequencing technology has developed at an unprecedented rate in recent years. Sanger sequencing is regarded as the gold standard technique to study alterations at the single nucleotide level. It allows easy detection of SNVs, small insertions, deletions and a moderate level of mosaicism (15-50%) [82, 83]. There is a risk of false negative results when there is allelic dropout due to variants in the primer binding sites, or when the target (exon/gene) sequence is deleted [84]. When the amplified alleles have significant size differences (e.g. due to insertions or deletions) preferential amplification of the shorter allele may mask the second allele. Current sequencing techniques ('Next generation sequencing' or NGS) allow screening of the whole genome in one experiment. Many different sequencing technologies and systems have emerged the past three years, utilising different methodologies. Ideally, complete genome sequencing followed by *de novo* assembly and comparison to a high-quality reference could identify thousands of sequence variants. In practice, NGS still faces technical, but more importantly, substantial computational and bioinformatic challenges. Read lengths are <100 bp for most approaches, significantly less than the 500-1000 bp routinely obtainable from Sanger sequencing. Base calling error rates are dependent on the platform that is used, but ranges from 0.01% to 16% [85]. Increasing the coverage (depth) can minimize false calls, allowing for accurate detection of SNVs [86].

Targeted sequencing approaches have the advantage of isolating multiple genomic regions of interest in a single experiment in an efficient and cost-effective manner. Whole Exome Sequencing (WES) captures only coding regions (~ 180000 exons) of the genome, and has been successfully applied in finding causative variants in many Mendelian disorders [87-91]. The total size of the human exome is approximately 30 Mb and comprises ~ 1% of the human genome. Therefore, when the region of interest is only protein coding sequences, exome sequencing is an efficient approach for obtaining the desired coverage for variant detection [92]. For WES, publicly available programs can be used for variant calling and annotation. Sequenced individuals have typically had 5000-10000 variant calls, representing non-synonymous substitutions in exonic regions, splicing alterations or small indels [88, 89, 93, 94]. Advanced filtering based on predicted effect (silent, nonsense, missense, effect on splicing, frameshift) or position (intronic, UTR, exonic, intergenic) can be used to find pathogenic variants. Two important assumptions underlie filtering strategies. The first is that causal variants for Mendelian disorders are rare, and therefore unlikely to be present in public databases or control sequencing data. The second is that synonymous variants are unlikely to be causative. Both assumptions are not always true [32, 95] . Mendelian disorders with milder phenotypes may have been overlooked in the general population, and may thus be present in control databases. Also, variants involved in recessive disorders with a high carrier frequency can be reported in control databases. For example, the most common variant in cystic fibrosis has an allele frequency of about 3% in Western European populations. Filtering such variants might erroneously exclude those pathogenic variants from consideration. As synonymous variants may have functional effects, and can be targeted by natural selection [96, 97] it is not always appropriate to filter these.

There are a several strategies that can be applied for the detection of structural variants. Various computational algorithms for identifying and characterizing variants have been developed. Read-pair methods assess the span and orientation of paired-end reads, and can be used to detect all types of structural variation (indels, inversions, tandem duplications and translocations) [22, 98]. Read-depth methods are simply based on the theory that CNV regions show differences in the number of reads, and therefore assess mapping depth in the sequenced sample. This will detect CNVs (including aneuploidy) and large insertions, but not inversions and translocations. Split-read approaches are able to detect all breakpoints (deletions, tandem duplications, translocations and inversions) [98, 99]. However, none of these strategies is comprehensive, and each will require substantial further development. Recently, an integrated computational pipeline for whole genome sequencing was published, enabling detection of all types of genetic variations (single nucleotide variants, short insertions or deletions (indels) and larger structural variations (SVs) [100] .

Interpretation of genomic variation

The major challenge in the analysis of genomic variation is to distinguish benign variants from variants that have clinical consequences. Increasingly, clinical molecular laboratories are detecting novel CNVs and sequence variations of unclear significance (UVs) in the course of testing patients. Guidelines on the interpretation of such variants have been developed [76, 101-105]. Sequence variants can consequently be classified on their phenotypic consequences within a spectrum of interpretations, ranging from those in which the variation is almost certainly of clinical significance, to those in which it is almost certainly not [106, 107].

When the impact of a sequence variant is undetermined, follow-up activities may be useful to clarify this relationship, and assist with risk assessment. There are several lines of evidence suggested for designating a variant as either phenotype-modifying or neutral. (1) Occurrence in patients or in a control population as listed in a database, (2) Variant type (missense, nonsense, silent), (3) *In silico* prediction of effect on protein structure or RNA splicing, (4) Conservation among species, and or presence in a known functional domain, (5) Co-segregation with a phenotype in the family.

The presence or absence of a variant in established databases, ideally containing all the lines of evidence listed above, is of great importance. It is helpful to establish whether a variant has been reported before, and if there is an associated phenotype. Distinction should be made between variants directly linked to a phenotype or variants found by genome-wide association studies (GWAS) which have identified many variants with a very small contribution to an associated trait, in which additional genes and/or environmental factors also influence phenotypic outcome (multifactorial traits) [108].

Currently, the numbers of submicroscopic imbalances that have been reported are increasing, but the delineation of associated clinical features remains difficult and incomplete. When large stretches of DNA encompassing several genes are deleted, complex phenotypes may emerge (so called 'contiguous gene syndromes'). It may often be unclear which gene in the interval is responsible for a given part of the phenotype (e.g. heart malformations, limb malformations), which is another reason why CNVs are collected and compared in patient databases.

As described previously, the capacity of a given variant (silent, missense, nonsense, frameshift, splicing) to affect gene expression or the function of its protein products must be determined. *In silico* prediction of pathogenic effects using information on interspecies sequence conservation can be helpful. The observation that pathogenic variants are more likely to occur at positions that are conserved through evolution suggested that prediction could be based on sequence homology [109]. Variants that alter conserved residues by replacing them with amino acids with different physical characteristics are likely to affect polypeptide structure and function. It was also observed that disease-causing amino acid substitutions have a common structural feature (for instance sites with low solvent accessibility, sites involved in disulphide bonds, sites involved in folding) that distinguishes them from neutral substitutions, suggesting that structure can also be used for predicting functional consequences [110, 111].

Databases collecting information on prevalence and frequency of sequence and structural variants in the human genome have been established, and provide information on the frequency of previously reported variants in disease and control cohorts. For rare recessive disorders it is generally assumed that a variant is unique, or at least has a low carrier frequency in the general population, whereas a variant that is found at higher frequency may be neutral. However, some recessive diseases are relatively frequent in subpopulations due to survival benefit for heterozygotes or genetic drift (for example sickle cell trait gives resistance to malarial infection), stressing the importance of registration of ethnicity in databases.

In silico prediction of an effect of a DNA variant on mRNA splicing should be verified by RNA studies. Such studies are not always feasible since not all genes are transcribed in easily accessible tissues such as blood.

Identification of a variant in a patient should be followed by testing of the parents. *De novo* variants are more likely to be pathogenic, but since we know each child may have between 30 to 50 new variants [9, 18], this is by no means certain. In rare cases co-segregation of the variant with the disease in a family may corroborate the suspicion of pathogenicity. Finally, testing the gene product in a functional assay or constructing animal models (e.g. mouse, rat or fruit fly) carrying the same variant may provide definitive assessment of the phenotypic effect of the variant *in vivo*. A reliable functional assay is generally regarded as one of the best means of confirming pathogenicity, however this is rarely available as part of a routine diagnostic service.

For structural variation such as CNV, it is generally considered that if a phenotypically normal family member carries the same chromosomal anomaly, the anomaly is of no phenotypic relevance [112]. However, caution remains necessary since the identification of a CNV in a patient that is also present in an apparently healthy parent does not rule out pathogenicity as there could be a pathogenic variant on the other allele. Also, some phenotypes may be mild and therefore overlooked in a parent. As described earlier for inherited CNVs, the CNV may in fact be a risk factor that only reveals a phenotype when combined with one or more other variants in the genome. In addition, this also means that even if a CNV is relatively frequent in a population this does not rule it out as a risk factor as long as it has not been reported as homozygously deleted in a healthy population.

Tools and databases

Databases allow researchers to share knowledge and retrieve information about genomic sites of variation under study. However, at present there is no database summarizing all that is known about a certain variant.

There are several human cytogenetic databases that link submicroscopic chromosomal imbalances (microdeletions/duplications/insertions, translocations and inversions) with clinical phenotype. These include DECIPHER (DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources)[113, 114] and ECARUCA (European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations) [115]. Since these databases contain patient information, complete access is limited to clinicians or cytogeneticists. Some information is publicly available through specified tracks in genome browsers. The Database of Genomic Variants (DGV) is a comprehensive database for the deposition, retrieval and visualization of human structural variation [116]. The database currently contains 179450 CNVs (April 2012) and reports CNVs, inversions and Indels (100 bp-1 Kb) in apparently healthy human cases. However, interpretation of the variants should be performed carefully, as different platforms have been used to detect CNVs, and population and medical data of the cohorts are poorly defined or absent. DbVar is a public repository that accepts direct submissions and provides archiving, accessioning and distribution of publicly available genomic structural variants [117]. It accepts data from all species, and includes clinical data. It marks and allows searching for variants that are known to be pathogenic.

The dbSNP database serves as a central repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms [118]. More than 17 million SNPs have been documented in this database, with a false-positive rate estimated at 15-17% [92]. However, the database does contain validated SNPs (allele frequencies provided) and can be queried accordingly, increasing the utility of the database. Allele and genotype frequencies from the HapMap project [119] have also been submitted to dbSNP. In the HapMap project, four large populations of African, Asian, and European ancestry have been studied extensively to catalogue population specific variation.

The 1000 Genomes Project [1] generated the most comprehensive map of human genetic variation yet, using next-generation sequencing technologies. It contains three pilot studies with a range of coverage. The exome variant server (EVS) [120] contains exome data of 6500 samples. A recent initiative is the generation of the genome of the Netherlands (GoNL), where whole genome sequences of 250 healthy Dutch parent-child trios are collected and stored in a biobank [121].

Some databases only contain information on genetic variation causing genetic disorders or traits. The Human Gene Mutation Database (HGMD) includes the first example of all exonic and +1,+2, -1,-2 splice-site variants causing or associated with human inherited disease, plus disease-associated polymorphisms reported in the literature [36]. Locus Specific DataBases (LSDB) are databases recording all variation within a gene. The databases contain accurate (curated), clearly referenced data naming variants at the DNA, RNA and protein level, and include all relevant comments relating to the clinical interpretation of the variant. Existing LSDBs can be checked by the URL "GeneSymbol.LOVD.nl" (e.g.MBTPS2.LOVD.nl) [122]. The Human Genome Variation Society (HGVS) keeps a list of locus specific variant databases [123]. Several tools have been developed to predict the effect of non-synonymous variants on a protein. The Grantham score (GMS) [124] represents one of the first attempts to assess the effect of amino acid substitutions on protein structure based on chemical properties, including the side-chain composition, polarity and molecular volume. The GMS is a measure of dissimilarity between a human amino acid and the residues seen at the same site in homologs. Several studies used a GMS score less than 60 to define neutral variants, whereas a GMS score significantly larger than 60 indicates that the amino-acid change is evolutionarily intolerant [125, 126].

Polyphen [29, 127, 128] utilizes a combination of 3D structural parameters and sequence homology to make a prediction about a functional effect. This prediction is based on a number of features comprising the sequence, phylogenetic and structural information characterizing the variant. It returns predictions of "probably damaging, "possibly damaging," benign and "unknown." 'Sorting Intolerant From Tolerant' (SIFT) [129] uses sequence homology and the physical properties of amino acids to predict effect on proteins. Next to SNVs, it can classify coding indels. It returns predictions of "affect protein function" and "tolerated" for each SNV. Due to differences in algorithms used for the predictions, Polyphen2 and SIFT may present contradictory results.

Tools that calculate conservation scores can aid in variant interpretation. A variant that leads to a nonconservative substitution of an evolutionarily conserved amino acid is more likely to be causative of the disorder than a variant that leads to a conservative substitution or alters an amino acid that is not evolutionarily conserved. Both phastCons [130] and phyloP [131] calculate conservation scores for three groups of organisms (primates, placental mammals and vertebrates). The two conservation scores are informative in different ways. Phastcons estimates the probability that a nucleotide belongs to a conserved element, taking neighbouring bases into account, while PhyloP predicts conservation purely at the base level.

UMD predictor [132] provides a combinatorial approach that associates several data such as localization within the protein, conservation, biochemical properties of the variant and wild-type residues, splice-site predictions and the potential impact of the variant on mRNA. Alamut [133] is a commercial package designed to help interpret variants quickly and uses different splice site prediction algorithms, PolyPhen, SIFT and calculates theoretical consequences of substitutions, insertions and deletions (effects on protein sequence, frameshifts, splicing effects, miRNA targets, nonsense-mediated mRNA decay).

Computational prediction of pathogenicity may also give false-negative results [134]. Most prediction methods for protein alterations do not take DNA sequence context into account. As a result, they can miss changes that alter splice sites [135]. Experimental verification by functional analysis of possible pathogenicity remains the golden standard.

Outline and scope of this thesis

Intellectual disability (ID) with or without multiple congenital malformations (MCA) is one of the main reasons for referral to a clinical geneticist. Causes of this ID/MCA are extremely heterogeneous and range from point mutation in one single specific gene to loss or gain of an entire chromosome. Despite enormous progress in diagnostic techniques in the past 50 years the cause for ID remains unknown in approximately 50% of the cases.

Establishing a diagnosis and understanding the cause of a genetic disorder is of great benefit for the patient and his/her family. This may provide information on prognosis, clinical management options, and anticipation on associated health issues for the patient, and may even be of therapeutic relevance in the future. Family members can be informed about recurrence risk and may be provided with options for prenatal (PND) and pre-implantation genetic diagnosis (PGD).

The main objective of the research in this thesis was to develop and apply novel molecular techniques to study the genetic basis of patients with intellectual disability (ID), multiple congenital anomalies (MCA), or other inherited disorders, and ultimately to gain insight into genetic disease mechanisms.

This thesis is a mirror image of the rapid evolution of techniques for analysis of DNA between 2006 and 2011. Many techniques described in this thesis (karyotyping, fluorescence in situ hybridization (FISH), array-comparative genome hybridization (CGH), single nucleotide polymorphism (SNP)-arrays, multiplex ligation dependent probe amplification (MLPA), high resolution melting curve analysis (HRMA), Sanger sequencing, and whole exome sequencing (WES) have been applied to find causative genome variants in our patient cohort. All of these techniques are now routinely used in clinical diagnostic laboratories, except whole exome and whole genome sequencing, implementation of which is being worked on.

More detailed data on human genetic variation are now rapidly accumulating, as new techniques for determining the primary structure of genomes have been developed at an unprecedented rate.

Although new genomic variants can arise in both the germline cells (whose DNA will be passed to offspring) and in somatic cells (the majority of cells in the human body), this thesis has a focus on the genetic characterization of germline variation.

At the outset of this work five years ago, genomic microarrays were introduced and have been applied in this thesis to identify structural variation in genetic disorders. **Chapter 2** describes methods for detecting CNV in the human genome. The focus lies on the development of a targeted array used to study patients with intellectual disability of unknown aetiology. **Chapter 3** shows the application of microarrays and MLPA in the elucidation and delineation of a new microdeletion syndrome. Because microarrays increase the resolution of chromosome analysis by 100-1000 fold and because small deletions and duplications are a major cause of ID/MCA, conventional karyotyping was replaced by microarray analysis in routine diagnostics.

Because of its speed, simplicity, and low cost, HRMA has become a popular technique for detection of sequence variants. The two major applications are targeted genotyping and gene scanning. **Chapter 4** describes the versatility of HRMA as a molecular technique and illustrates several applications. In **Chapter 5,** SNP arrays were used to define the breakpoints of a previously defined locus for Keratosis Follicularis Spinulosa Decalvans (KFSD) and show the application of HRMA as a presequencing tool in the identification of the genetic basis of this disorder.

The introduction of NGS techniques allowed genome wide detection of sequence variants in disorders of unknown aetiology. **Chapter 6 and 7** describe the success of next generation sequencing in the identification of the genetic basis in two disorders (TOD and Aarskog-Scott syndrome), and shows the value of exome sequencing in detection of variants outside coding regions. Besides identification of the genetic cause of Coffin Siris syndrome**. Chapter 8** outlines recent findings on the mode of inheritance and frequency of this monogenic disorder using exome sequencing.

Finally, in **Chapter 9** the rapid evolution of techniques for genome analysis is discussed and alterations in the diagnostic approach to ID/MCA patients are proposed.

1. The 1000 Genomes Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467 (7319):1061-1073

2. Rahim NG, Harismendy O, Topol EJ, and Frazer KA (2008) Genetic determinants of phenotypic diversity in humans. Genome Biol 9 (4):215

3. Cummings M (1999) Human Heredity: Principles and Issues . 5th ed.

4. Cotton RG (2002) Communicating "mutation:" Modern meanings and connotations. Hum Mutat 19 (1):2-3

5. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N et al (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453 (7191):56-64

6. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, and Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36 (9):949-951

7. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H et al (2006) Global variation in copy number in the human genome. Nature 444 (7118):444-454

8. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, and Eichler EE (2005) Fine-scale structural variation of the human genome. Nat Genet 37 (7):727-732

9. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J et al (2010) Origins and functional impact of copy number variation in the human genome. Nature 464 (7289):704-712

10. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, and Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305 (5683):525-528

11. Feuk L, Carson AR, and Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7 (2):85-97

12. Sharp AJ (2009) Emerging themes and new challenges in defining the role of structural variation in human disease. Hum Mutat 30 (2):135-144

13. van Binsbergen (2011) Origins and breakpoint analyses of copy number variations: up close and personal. Cytogenet Genome Res 135 (3-4):271-276

14. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, and Eichler EE (2002) Recent segmental duplications in the human genome. Science 297 (5583):1003-1007

15. Lee JA, Carvalho CM, and Lupski JR (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131 (7):1235-1247

16. Alkan C, Coe BP, and Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12 (5):363-376

17. Frazer KA, Murray SS, Schork NJ, and Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10 (4):241-251

18. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, and Galas DJ (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328 (5978):636-639

19. Reumers J, De RP, Zhao H, Liekens A, Smeets D, Cleary J, Van LP, Van Den Bossche M, Catthoor K, Sabbe B, Despierre E, Vergote I, Hilbush B, Lambrechts D, and Del-Favero J (2012) Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nat Biotechnol 30 (1):61-68

20. Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, Diaz de ST, Menzel U, Sandgren J, von TD, Poplawski A, Crowley M, Crasto C, Partridge EC, Tiwari H, Allison DB, Komorowski J, van Ommen GJ, Boomsma DI, Pedersen NL, den Dunnen JT, Wirdefeldt K, and Dumanski JP (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 82 (3):763-771

21. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, Kim BC, Kim SY, Kim WY, Kim C, Park D, Lee YS, Kim S, Reja R, Jho S, Kim CG, Cha JY, Kim KH, Lee B, Bhak J, and Kim SJ (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res 19 (9):1622-1629

22. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, and Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318 (5849):420-426

23. Hehir-Kwa JY, Wieskamp N, Webber C, Pfundt R, Brunner HG, Gilissen C, de Vries BB, Ponting CP, and Veltman JA (2010) Accurate distinction of pathogenic from benign CNVs in mental retardation. PLoS Comput Biol 6 (4):e1000752

24. Girirajan S, Campbell CD, and Eichler EE (2011) Human copy number variation and complex genetic disease. Annu Rev Genet 45:203-226

25. Li J, Yang T, Wang L, Yan H, Zhang Y, Guo Y, Pan F, Zhang Z, Peng Y, Zhou Q, He L, Zhu X, Deng H, Levy S, Papasian CJ, Drees BM, Hamilton JJ, Recker RR, Cheng J, and Deng HW (2009) Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations. PLoS One 4 (11):e7958

26. Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, and Eichler EE (2010) *De novo* rates and selection of large copy number variation. Genome Res 20 (11):1469-1481

27. Rees E, Moskvina V, Owen MJ, O'Donovan MC, and Kirov G (2011) *De novo* rates and selection of schizophrenia-associated copy number variants. Biol Psychiatry 70 (12):1109-1114

28. Buchanan JA and Scherer SW (2008) Contemplating effects of genomic structural variation. Genet Med 10 (9):639-647

29. Ramensky V, Bork P, and Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30 (17):3894-3900

30. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, and Shoemaker DD (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302 (5653):2141-2144

31. Wang GS and Cooper TA (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet 8 (10):749-761

32. Waldman YY, Tuller T, Keinan A, and Ruppin E (2011) Selection for translation efficiency on synonymous polymorphisms in recent human evolution. Genome Biol Evol 3:749-761

33. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, and Gottesman MM (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315 (5811):525-528

34. Gringras P and Chen W (2001) Mechanisms for differences in monozygous twins. Early Hum Dev 64 (2):105-117

35. Petronis A (2006) Epigenetics and twins: three variations on the theme. Trends Genet 22 (7):347-350

36. Stenson PD, Ball EV, Howells K, Phillips AD, Mort M, and Cooper DN (2009) The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. Hum Genomics 4 (2):69-72

37. Battaglia A and Carey JC (2003) Diagnostic evaluation of developmental delay/mental retardation: An overview. Am J Med Genet C Semin Med Genet 117C (1):3-14

38. World Health Organization. Assessment of People with Mental Retardation. 1992.

39. Maulik PK, Mascarenhas MN, Mathers CD, Dua T, and Saxena S (2011) Prevalence of intellectual disability: a meta-analysis of population-based studies. Res Dev Disabil 32 (2):419-436

40. Stevenson RE, Procopio-Allen AM, Schroer RJ, and Collins JS (2003) Genetic syndromes among individuals with mental retardation. Am J Med Genet A 123A (1):29-32

41. Baker K, Raymond FL, and Bass N (2012) Genetic investigation for adults with intellectual disability: opportunities and challenges. Curr Opin Neurol 25 (2):150-158

42. Rauch A, Hoyer J, Guth S, Zweier C, Kraus C, Becker C, Zenker M, Huffmeier U, Thiel C, Ruschendorf F, Nurnberg P, Reis A, and Trautmann U (2006) Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. Am J Med Genet A 140 (19):2063-2074

43. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C et al (2011) A copy number variation morbidity map of developmental delay. Nat Genet 43 (9):838-846

44. Vissers LE, de LJ, Gilissen C, Janssen I, Steehouwer M, de VP, van LB, Arts P, Wieskamp N, del RM, van Bon BW, Hoischen A, de Vries BB, Brunner HG, and Veltman JA (2010) A *de novo* paradigm for mental retardation. Nat Genet 42 (12):1109-1112

45. Ji Y, Eichler EE, Schwartz S, and Nicholls RD (2000) Structure of chromosomal duplicons and their role in mediating human genomic disorders. Genome Res 10 (5):597-610

46. McCarroll SA (2008) Extending genome-wide association studies to copy-number variation. Hum Mol Genet 17 (R2):R135-R142

47. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, Chu SH et al (2011) Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70 (5):863-885

48. Stankiewicz P and Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med 61:437-455

49. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B et al (2007) Strong association of *de novo* copy number mutations with autism. Science 316 (5823):445-449

50. Cook EH, Jr. and Scherer SW (2008) Copy-number variations associated with neuropsychiatric conditions. Nature 455 (7215):919-923

51. Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, Richmond TA, Guiver C, Albertson DG, Pinkel D, Eis PS, Schwartz S, Knight SJ, and Eichler EE (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet 38 (9):1038-1042

52. Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, Itsara A, Vives L et al (2010) A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nat Genet 42 (3):203-209

53. Albers CA, Paul DS, Schulze H, Freson K, Stephens JC, Smethurst PA, Jolley JD et al (2012) Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nat Genet 44 (4):435-2

54. Kelberman D, Islam L, Holder SE, Jacques TS, Calvas P, Hennekam RC, Nischal KK, and Sowden JC (2011) Digenic inheritance of mutations in FOXC1 and PITX2 : correlating transcription factor function and Axenfeld-Rieger disease severity. Hum Mutat 32 (10):1144-1152

55. The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72 (6):971-983

56. Knoll JH, Nicholls RD, Magenis RE, Graham JM, Jr., Lalande M, and Latt SA (1989) Angelman and Prader-Willi syndromes share a common chromosome 15 deletion but differ in parental origin of the deletion. Am J Med Genet 32 (2):285-290

57. Eggermann T (2009) Silver-Russell and Beckwith-Wiedemann syndromes: opposite (epi) mutations in 11p15 result in opposite clinical pictures. Horm Res 71 Suppl 2:30-35

58. Katsanis N, Ansley SJ, Badano JL, Eichers ER, Lewis RA, Hoskins BE, Scambler PJ, Davidson WS, Beales PL, and Lupski JR (2001) Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. Science 293 (5538):2256-2259

59. Lerer I, Sagi M, Ben-Neriah Z, Wang T, Levi H, and Abeliovich D (2001) A deletion mutation in GJB6 cooperating with a GJB2 mutation in trans in non-syndromic deafness: A novel founder mutation in Ashkenazi Jews. Hum Mutat 18 (5):460

60. Helderman-van den Enden AT, de JR, den Dunnen JT, Houwing-Duistermaat JJ, Kneppers AL, Ginjaar HB, Breuning MH, and Bakker E (2009) Recurrence risk due to germ line mosaicism: Duchenne and Becker muscular dystrophy. Clin Genet 75 (5):465-472

61. Haldane J.B (1947) The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. Ann Eugen 13 (4):262-271

62. Hurst LD and Ellegren H (1998) Sex biases in the mutation rate. Trends Genet 14 (11):446-452

63. Makova KD and Li WH (2002) Strong male-driven evolution of DNA sequences in humans and apes. Nature 416 (6881):624-626

64. Hehir-Kwa JY, Rodriguez-Santiago B, Vissers LE, de LN, Pfundt R, Buitelaar JK, Perez-Jurado LA, and Veltman JA (2011) *De novo* copy number variants associated with intellectual disability have a paternal origin and age bias. J Med Genet 48 (11):776-778

65. Pearson PL (2006) Historical development of analysing large-scale changes in the human genome. Cytogenet Genome Res 115 (3-4):198-204

66. Tjio JH and Levan a (1956) The chromosome number in man. Hereditas 42:1-6

67. Lejeune J, Gautier M, and Turpin MR (1959) Etude des chromosomes somatiques de neuf enfants mongoliens. CR Acad Sci III (248):1721-1722

68. Gerhard DS, Kawasaki ES, Bancroft FC, and Szabo P (1981) Localization of a unique gene by direct hybridization in situ. Proc Natl Acad Sci U S A 78 (6):3755-3759

69. Van Prooijen-Knegt AC, Van Hoek JF, Bauman JG, Van DP, Wool IG, and Van der Ploeg M (1982) In situ hybridization of DNA sequences in human metaphase chromosomes visualized by an indirect fluorescent immunocytochemical procedure. Exp Cell Res 141 (2):397-407

70. Francomano CA and Kazazian HH, Jr. (1986) DNA analysis in genetic disorders. Annu Rev Med 37:377-395

71. Sanger F, Nicklen S, and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74 (12):5463-5467

72. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, and Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258 (5083):818-821

73. Schoumans J and Ruivenkamp C (2010) Laboratory methods for the detection of chromosomal abnormalities. Methods Mol Biol 628:53-73

74. Gijsbers AC (2010) high-resolution karyotyping by oligonucleotide microarrays: the next revolution in cytogenetics.

75. Florijn RJ, Bonden LA, Vrolijk H, Wiegant J, Vaandrager JW, Baas F, den Dunnen JT, Tanke HJ, van Ommen GJ, and Raap AK (1995) High-resolution DNA Fiber-FISH for genomic DNA mapping and colour bar-coding of large genes. Hum Mol Genet 4 (5):831-836

76. Gijsbers AC, Lew JY, Bosch CA, Schuurs-Hoeijmakers JH, van HA, den Hollander NS, Kant SG, Bijlsma EK, Breuning MH, Bakker E, and Ruivenkamp CA (2009) A new diagnostic workflow for patients with mental retardation and/or multiple congenital abnormalities: test arrays first. Eur J Hum Genet 17 (11):1394-1402

77. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, and Gunderson KL (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16 (9):1136-1148

78. Rodriguez-Santiago B, Malats N, Rothman N, Armengol L, Garcia-Closas M, Kogevinas M, Villa O, Hutchinson A, Earl J, Marenne G, Jacobs K, Rico D, Tardon A, Carrato A, Thomas G, Valencia A, Silverman D, Real FX, Chanock SJ, and Perez-Jurado LA (2010) Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. Am J Hum Genet 87 (1):129-138

79. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, and Pals G (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res 30 (12):e57

80. Reed GH and Wittwer CT (2004) Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. Clin Chem 50 (10):1748-1754

81. Reed GH, Kent JO, and Wittwer CT (2007) High-resolution DNA melting analysis for simple and efficient molecular diagnostics. Pharmacogenomics 8 (6):597-608

82. Rohlin A, Wernersson J, Engwall Y, Wiklund L, Bjork J, and Nordling M (2009) Parallel sequencing used in detection of mosaic mutations: comparison with four diagnostic DNA screening techniques. Hum Mutat 30 (6):1012-1020

83. Necker J, Kovac M, Attenhofer M, Reichlin B, and Heinimann K (2011) Detection of APC germ line mosaicism in patients with *de novo* familial adenomatous polyposis: a plea for the protein truncation test. J Med Genet 48 (8):526-529

84. Sian Ellard, Ruth Charlton, Michael Yau, David Gokhale, Graham R Taylor, and AndrewWallace6 and Simon C Ramsden. CMGS: Practice guidelines for Sanger Sequencing Analysis and Interpretation. 2009.

85. Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour 11 (5):759-769

86. Koboldt DC, Ding L, Mardis ER, and Wilson RK (2010) Challenges of sequencing human genomes. Brief Bioinform 11 (5):484-498

87. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, and Shendure J (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 42 (9):790-793

88. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, and Bamshad MJ (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42 (1):30-35

89. Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, van LB, Steehouwer M, van RJ, Kant SG, Roepman R, Knoers NV, Veltman JA, and Brunner HG (2010) Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am J Hum Genet 87 (3):418-423

90. Hoischen A, van Bon BW, Gilissen C, Arts P, van LB, Steehouwer M, de VP, de RR, Wieskamp N, Mortier G, Devriendt K, Amorim MZ, Revencu N, Kidd A, Barbosa M, Turner A, Smith J, Oley C, Henderson A, Hayes IM, Thompson EM, Brunner HG, de Vries BB, and Veltman JA (2010) *De novo* mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet 42 (6):483-485

91. Wang JL, Yang X, Xia K, Hu ZM, Weng L, Jin X, Jiang H, Zhang P, Shen L, Guo JF, Li N, Li YR, Lei LF, Zhou J, Du J, Zhou YF, Pan Q, Wang J, Wang J, Li RQ, and Tang BS (2010) TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain 133 (Pt 12):3510-3518

92. Ku CS, Naidoo N, and Pawitan Y (2011) Revisiting Mendelian disorders through exome sequencing. Hum Genet 129 (4):351-370

93. Rodelsperger C, Krawitz P, Bauer S, Hecht J, Bigham AW, Bamshad M, de Condor BJ, Schweiger MR, and Robinson PN (2011) Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. Bioinformatics 27 (6):829-836

94. Rios J, Stein E, Shendure J, Hobbs HH, and Cohen JC (2010) Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. Hum Mol Genet 19 (22):4313-4318

95. Gilissen C, Hoischen A, Brunner HG, and Veltman JA (2012) Disease gene identification strategies for exome sequencing. Eur J Hum Genet 20 (5):490-497

96. Chamary JV, Parmley JL, and Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7 (2):98-108

97. Sun Y, Almomani R, Aten E, Celli J, van der HJ, Venselaar H, Robertson SP, Baroncini A, Franco B, Basel-Vanagaite L, Horii E, Drut R, Ariyurek Y, den Dunnen JT, and Breuning MH (2010) Terminal osseous dysplasia is caused by a single recurrent mutation in the FLNA gene. Am J Hum Genet 87 (1):146-153

98. Albers CA, Lunter G, Macarthur DG, McVean G, Ouwehand WH, and Durbin R (2011) Dindel: accurate indel calls from short-read data. Genome Res 21 (6):961-973

99. Ye K, Schulz MH, Long Q, Apweiler R, and Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25 (21):2865-2871

100. Lam HY, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, Bustamante CD, and Snyder M (2012) Detecting and annotating genetic variations using the HugeSeq pipeline. Nat Biotechnol 30 (3):226-229

101. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, Lyon E, and Ward BE (2008) ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med 10 (4):294-300

102. Cotton RG and Scriver CR (1998) Proof of “disease causing” mutation. Hum Mutat 12 (1):1-3

103. Bell J, Danielle Bodmer Erik Sistermans and Simon C Ramsden. Practice guidelines for the Interpretation and Reporting of Unclassified Variants (UVs) in Clinical Molecular Genetics. 2007.

104. Lee C, Iafrate AJ, and Brothman AR (2007) Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. Nat Genet 39 (7 Suppl):S48-S54

105. Koolen DA, Pfundt R, de LN, Hehir-Kwa JY, Nillesen WM, Neefs I, Scheltinga I, Sistermans E, Smeets D, Brunner HG, van Kessel AG, Veltman JA, and de Vries BB (2009) Genomic microarrays in mental retardation: a practical workflow for diagnostic applications. Hum Mutat 30 (3):283-292

106. American college of medical genetics. Practice guidelines for Sanger Sequencing Analysis and Interpretation . 2006. The Standards and Guidelines for Clinical Genetics Laboratories.

107. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, and Tavtigian SV (2008) Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat 29 (11):1282-1291

108. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, and Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9 (5):356-369

109. Miller MP and Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10 (21):2319-2328

110. Wang Z and Moult J (2001) SNPs, protein structure, and disease. Hum Mutat 17 (4):263-270

111. Sunyaev S, Ramensky V, and Bork P (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 16 (5):198-200

112. Barber JC, Maloney V, Hollox EJ, Stuke-Sontheimer A, du BG, Daumiller E, Klein-Vogler U, Dufke A, Armour JA, and Liehr T (2005) Duplications and copy number variants of 8p23.1 are cytogenetically indistinguishable but distinct at the molecular level. Eur J Hum Genet 13 (10):1131-1136

113. DECIPHER. http://decipher.sanger.uk.

114. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van VS, Moreau Y, Pettett RM, and Carter NP (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet 84 (4):524-533

115. ECARUCA. http://www.ecaruca.net.

116. Database of Genomic Variants. http://projects.tcag.ca/variation/.

117. Dbvar. http://www.ncbi.nlm.nih.gov/dbvar.

118. dbSNP. http://www.ncbi.nlm.nih.gov/projects/SNP.

119. HapMap. http://hapmap.ncbi.nlm.nih.gov.

120. Exome Variant Server. http://evs.gs.washington.edu/EVS/.

121. GoNL. http://www.nlgenome.nl.

122. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, and den Dunnen JT (2011) LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 32 (5):557-563

123. Human Genome Variation Society. http://www.hgvs.org/dblist/glsdb.html.

124. Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185 (4154):862-864

125. Abkevich V, Zharkikh A, Deffenbaugh AM, Frank D, Chen Y, Shattuck D, Skolnick MH, Gutin A, and Tavtigian SV (2004) Analysis of missense variation in human BRCA1 in the context of interspecific sequence variation. J Med Genet 41 (7):492-507

126. Lee E, McKean-Cowdin R, Ma H, Chen Z, Van Den Berg D, Henderson BE, Bernstein L, and Ursin G (2008) Evaluation of unclassified variants in the breast cancer susceptibility genes BRCA1 and BRCA2 using five methods: results from a population-based study of young breast cancer patients. Breast Cancer Res 10 (1):R19

127. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, and Kuznetsov EN (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 12 (5):387-394

128. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7 (4):248-249

129. Ng PC and Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11 (5):863-874

130. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, and Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15 (8):1034-1050

131. Siepel A, Pollard KS and Haussler D. New methods for detecting lineage-specific selection. 190-205. 2006. In Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006).

132. Frederic MY, Lalande M, Boileau C, Hamroun D, Claustres M, Beroud C, and Collod-Beroud G (2009) UMD-predictor, a new prediction tool for nucleotide substitution pathogenicity -- application to four genes: FBN1, FBN2, TGFBR1, and TGFBR2. Hum Mutat 30 (6):952-959

133. Alamut. http://www.interactive-biosoftware.com/alamut.html

134. Raymond FL, Whibley A, Stratton MR, and Gecz J (2009) Lessons learnt from large-scale exon re-sequencing of the X chromosome. Hum Mol Genet 18 (R1):R60-R64

135. Ng PC and Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61-80