Mapping known and novel genetic variation in the human genome:

bioinformatic tool development and applications

Mircea Cretu Stancu

Cover design: Alessio Marcozzi

Interior design: Mircea Cretu Stancu

Printed by: Proefschriftmaken.nl ll Uitgeverij BOXPress

ISBN: 978-94-6295-957-6

Mapping known and novel genetic variation in the human genome:

bioinformatic tool development and applications

Het in kaart brengen van bekende en nieuwe genetische varianten in het humane genoom:
ontwikkeling van bioinformatica software en applicaties
(met een samenvatting in het Nederlands)

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op dinsdag 22 mei 2018 des middags te 12.45 uur

door

Mircea Cretu Stancu
geboren op 23 juli 1989 te Calarasi, Roemenië

Promotor: Prof.dr. E.P.J.G. Cuppen

Copromotor: Dr. W.P. Kloosterman

Table of Contents

CHAPTER 1

INTRODUCTION

Genetic code inheritance and variation

The genetic code of an organism is encoded in the deoxyribonucleic acid (DNA) which is represented as an arbitrary long (linear) sequence of four admissible nucleotide bases: adenine (A), cytosine (C) , guanine (G) and thymine (T). Each consecutive nucleotide in the sequence is hydrogen bound, to its complementary base (A to T, C to G and vice-versa), forming two complementary strands (i.e.: arbitrarily designated as the forward and the reverse strand respectively) that form the double helix shape of a DNA molecule[1]. We note that one of the two strands is sufficient to completely describe a particular DNA molecule.

The diploid human genome consists of approximately 6.4 billion bases, organized into 22 pairs of (nearly identical) homologous chromosomes, denoted 1 to 22 (where chromosome 1 is the longest), and a pair of sex chromosomes, XX for females and XY for males. The two copies of an homologous pair of chromosomes (i.e.: autosomes) encode for the same genetic information, potentially with various content differences, whereas the two sex chromosomes (X and Y) encode for different genetic information, with the exception of three genomic regions, that have been observed to match in an autosomal fashion, denoted pseudo-autosomal regions (PAR) 1 through 3.

The genetic content of an individual, at some genetic locus, whether diploid (i.e.: autosomes) or haploid (i.e.: non-homologous sex chromosomes) is denoted as the genotype of the locus, and the genetic content of one/each individual chromosomes at that locus is denoted as an allele (Table 1).

Organismal development and mating

Any organism grows and develops through repeated cell divisions, from an original cell called a zygote. The genetic code of the zygote, the germline DNA, is thus present (and ideally identical) in all cells of the body. The human zygote is the result of fertilization, where two cells called gametes, produced through meiosis in each parent respectively, merge. Each gamete is haploid and contains one (random) allele of each chromosomal pair of the originating parent (including the sex chromosomes) that is transmitted to the offspring. Sometimes, when the two parental chromosomes separate during meiosis, homologous (or not) genetic material is swapped between two parental chromosomes, through the process of recombination, and the haploid chromosome that is transmitted is a mosaic of the two original parental chromosomes. After fertilization, the zygote contains the genetic code resulting from pairing the inherited paternal and maternal chromosomes respectively, as postulated by Mendel's laws of inheritance.

De novo mutations

While it was known that the genetic may vary between individuals, in the 1940s it was inferred that the DNA molecule is altered by discrete "jump like" events (i.e.: sequence changes)[2]. Thus, while inherited from its parents, an offspring's genetic code may contain variation not present in the parents, at a relatively small number of positions across the genome. *De novo* germline mutations (DNMs) are introduced as faulty repairs produced by the cell's DNA repair mechanisms and appear during the DNA replication that is performed at cell divisions[3,4], prior or during to the generation of gametes[5]. *De novo* mutations may be introduced during the DNA replication needed for any cell division, and such mutations that arise within an organism (i.e.: as opposed to transmitted) are called somatic (*de novo*) mutations. All genetic variation observed between individuals today was introduced, at some point, *de novo* and passed on through the generations.

Selection

This mutational process is agnostic to the effect that it may produce to the individual and indeed *de novo* mutations with a detrimental effect are relatively more frequent than beneficial mutations[6,7]. Nonetheless, DNMs are the raw material that natural selection operates on. Thus, if a specific mutation produces a beneficial effect on an individual's fitness to survive and reproduce, then this advantage will propagate successfully over generations and the respective mutation will be found in more individuals over time (i.e.: rise in frequency in the population - Table 1).

De novo mutation rate

In humans, it is estimated that the probability that a *de novo* mutation arises is $1.8 * 10^{-8}$ per base, per generation[8,9], amounting to an average number of ~54 DNMs in any individual, although it was found that this mutation rate is not uniform across the genome[10]. Specifically, mutation rates vary with respect to sequence content, with transitions (A <-> G and C <-> T) being more frequent than transversions[11] (all other combinations) and CpG sites are 18 fold more mutable than non CpG sites[9]. Ocasionally, de *novo* mutations cluster[10,12,13] or co-segregate with larger de *novo* events such as large copy number variations (CNVs)[14,15]. Furthermore, mutation rates vary with respect to genomic properties such as recombination levels, transcription and replication timing, although these effects were found to vary with father's age at conception[10]. Thus, early replicating DNA regions, correlated with higher gene density and gene expression, were found to be relatively depleted of DNMs in offspring of younger fathers, reducing the chances of passing on deleterious DNMs. Due to continuous cell divisions and generation of gametes throughout the father's life, through spermatogenesis[5], DNMs were found to accumulate in the paternal germline, at a rate of ~2 DNMs per year.

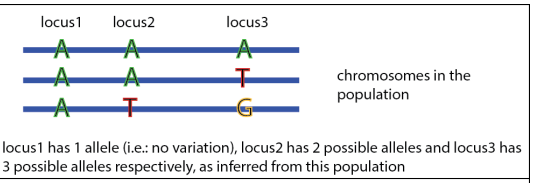
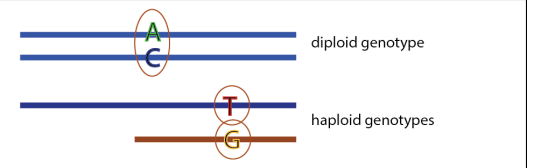
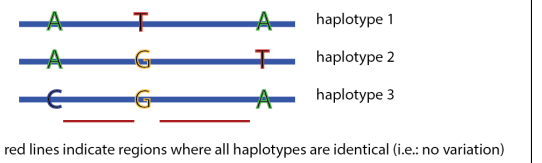
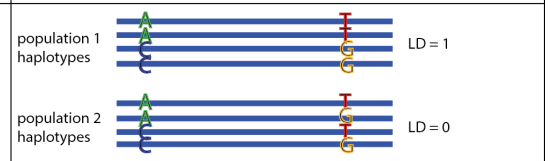
Allele: The genetic content of one chromosome, at a specific locus. Typically, only alleles that have been observed in at least one individual are considered. Alleles may be one nucleobase long or more and genetic locus may refer to one nucleobase or whole genes, etc.	
Genotype: Specific combination of alleles that an individual possesses at a genetic locus. Genotypes are diploid for the autosomes and X chromosome in females and haploid for the X an Y chromosome in males	
Haplotype: Contiguous stretch of DNA, usually spanning more variant sites, that lies on the same chromosome; may refer to anything from a few consecutive basses to whole chromosomes	
Linkage Disequilibrium (LD): metric that quantifies how often the alleles at different loci co-segregate on the same haplotype, in a population; e.g.: for 2 bi-allelic loci, LD = 1 implies that a specific allele at locus1 always occurs on the same haplotype with a specific allele at locus2 (and vice-versa), while LD = 0 implies that the allele combinations at the two loci are random	

Table1: Illustration of genomic terms used throughout this introduction

De novo mutations and human disease

De novo mutations are an important determinant of human disease in two ways. First, they may severely disrupt the normal function of a specific gene, drastically reducing the survival chances of the offspring and their possibility to reproduce. This typically results in early onset, rare and severe syndromes[16–18]. The phenotype of the individual is typically so severe that he or she will not have offspring, so these syndromes arise mostly through DNMs (i.e.: rather than transmission) and hence remain rare. More than 500 gene associations emerged, that explain sporadic disease by DNMs[19,20]. Secondly, DNMs have also been reported to contribute to common, typically neuro-developmental and psychiatric diseases such as autism-spectrum disorders(ASD)[21,22], epilepsy[23] and intellectual disability(ID)[24]. In such life-long phenotypes, the contribution of DNMs is typically smaller and more diffuse: *de novo* mutations in any of many genes involved with these common diseases are found to increase risk of affection[25]. A method for increasing power to detect mutations and genes that affect disease is to group a set of phenotypes that is heterogeneous and hard to diagnose accurately because of overlapping phenotypes under a common umbrella phenotype. Such is the case for neurodevelopmental disorders (NDDs), which includes ASD, ID, developmental delay (DD) and epilepsy. Furthermore, due to the rarity and spread of DNMs, methods to aggregate evidence have been also been devised. Typically, DNMs are first classified by their predicted effect on the target gene and/or by position (i.e.: intronic, exonic, enhancer regions)[26,27]. Specifically, patients with NDDs are enriched for "likely gene damaging" (LGD) mutations (stop codon, frameshift, splice donor, and acceptor) and missense mutations[22,28,2930]. Hundreds of genes have thus emerged as potentially causal for NDDs[25]. Clinical practice has also recognized the diagnostic relevance of DNMs and whole genome sequencing is increasingly suggested as a means of identifying the exact causal variants[24 31]. In **chapter 2** of this thesis I present an algorithm and a pipeline that can be used to accurately detect these rare *de novo* events using current state of the art whole genome sequencing technologies.

Genetic variation in humans

The first (nearly) complete human genome was released in 2001, as a result of the Human Genome Project[32], providing unprecedented insight into the content and structure of our genetic material. This was later adopted as the Human Reference Genome, which is continuously maintained and (still) used as an unique reference point and coordinate system when identifying and representing genetic variation in any individual. Since then, many international collaborations and large sequencing projects have drastically increased the number of genomes and genetic variation analyzed[33–35]. Genetic variation falls into three broad categories: 1) single nucleotide variants (SNVs), 2) short insertions or deletions that affect less than 20 base-pairs (indels) and 3) structural variants (SVs) that affect more than 50 base-pairs. The 1000Genomes Phase3 release[34], describing the genomes of 2,504 individuals sampled from 26 populations across the globe, found that the average human genome differs from the reference genome at ~4.1 million variant sites, although this estimate varies between different populations, corresponding to approximately one variant site every 1000 bases.

Of these variants, ~3.6 million are SNVs, ~556,000 are indels and ~2,500 are SVs. While the majority of the variants within an individual are common in the population (frequency > 0.5%), a total of ~80 million variants are reported, most of which are rare, specific to a population or found in only one individual. The genetic variation described in such large population samples (and others[33]) have been used to characterize the co-segregation of variant alleles onto haplotypes (i.e.: linkage-disequillibrium - LD - Table 1) and to produce large reference panels of such haplotypes. Using statistical imputation methods and such resources has enabled researchers to accurately reconstruct the genetic variation of an individual, by interrogating only a subset of specific regions of the genome, thus making the analysis of hundreds of thousands of individuals' genetic data affordable[36]. Further, the "Exome Aggregation Consortium" (EXAC)[37] and its follow-up the "Genome Aggregation Database" (gnomAD) are international collaborations that aim at centralizing and integrating whole-exome and whole-genome generated across the globe (currently containing ~123,000 whole-exomes and ~15,000 whole genomes). Comprising mostly of rare variation, they offer accurate estimations of the frequency of any detected variant and aggregate measures such as each gene's tolerance to variation, thus providing a useful resource for the interpretation of genetic variation found in a clinical setting and for resolving the genetic causes of rare diseases[38].

Structural variation

Structural variation (SV), despite consisting of relatively fewer events per genome (i.e.: than SNVs and indels as described above), collectively affects a larger proportion of the genome than all other types of variation (1% of the human, compared to 0.1% for SNVs)[39,40,41]. There are many types of SVs, depending on the biological mechanisms by which they are thought to arise and definitions are not mutually exclusive[42–44]. Furthermore, complex events involving more than one simple SV are often detected[45,46].Transposable elements are genomic elements that are able to copy themselves in our genome either independently or by recruiting the copy machinery of a cell. They are estimated to account for the origin of ~45% of the sequence of human genome [32,47], are still found to occur de novo at a rate of ~1/43 genomes[48] and explain ~23% of structural variation within a genome[49]. SVs may disrupt a large genomic locus, with consequences on gene content and/or gene expression levels[50]. Large chromosomal abnormalities such as trisomies, monosomies, large copy-number variations (CNVs), that are readily identifiable through karyotyping, are known to cause rare and severe syndromes such as the Smith-Magenis syndrome, Williams-Beuren syndrome or Potocki-Lupski syndrome[50]. Common and smaller SVs (1-200kb) play a role in the onset of common diseases such as Crohn's disease[51], attention deficit hyperactivity disorder[52], rheumatoid arthritis and type 1 diabetes[53]. Furthermore, genomic instability, that in turn leads to an accumulation of SVs is a hallmark of cancer genomics[54] and has been observed across different types of cancers[55,56].

A particularly complex and dramatic molecular event is chromothripsis, where, during a single event, the genome of a cell (i.e.: potentially germline cell) acquires tens to hundreds of structural variants[57]. During chromothripsis large chunks of one or more chromosomes are

shattered to pieces and subsequently re-arranged to form new, derivative chromosomes. These rearrangements may contain translocation of genetic code within or between chromosomes, large deletions and insertions, duplications and inversions and many times they are visible on a microscopic scale through karyotyping. Chromothripsis is a common event in some types of cancers such as colorectal cancer[58,59], multiple myelomas[60] or prostate cancers[61,62]. Germline chromothriptic events that result in a reshuffling of the transmitted chromosomes typically cause severe congenital abnormalities[63–65]. However, when the cell's DNA repair mechanisms reconstruct derivative chromosomes with minimal overall loss of DNA and the chromothripsis breakpoints do not affect major functional genes/regions of the genome, rearranged chromosomes may be found in the germline of healthy individuals[65].

The Major Histocompatibility Complex

The human Major Histocompatibility Complex (MHC) is a 3.6 mega-base (MB) long genomic region, on the short arm of chromosome 6, which harbours a cluster of genes involved in both the acquired and the innate immune response[66]. It was discovered more than 60 years ago for its antigen histocompatibility role in the transplantation of organs, hence the alternative name of "Human Leukocyte Antigen" (HLA) region. Soon after its discovery, the link was made, between the genetic content at the MHC locus and immune response by evaluating the response of rabbit and mouse strains to synthetic antigens[67]. Much later it was estimated that 40% of the genes across the MHC locus are involved in the immune response[68]. Later on it was established that the MHC contains 8 "classical" (i.e.: for historical reasons) HLA genes, three class I genes (HLA- A/B/C) and 5 class II genes (HLA- DQA/DQB/DRB/DPA/DPB) respectively[69], that encode for proteins expressed on the surface of cells. These proteins in turn bind peptides (i.e.: small molecular protein residues, potentially from pathogens) either from within the cell (class I) or from outside the cell (class II) and may trigger an immune response by further binding to T-lymphocyte cell receptors.

The role of the MHC in human disease

The importance of the MHC locus was demonstrated, beyond its modulation of host-graft interactions in transplantations[70], by association to a range of diseases, predominantly autoimmune and inflammatory diseases, and infectious diseases[71], but also to other phenotypes, such as neurological disorders[72–74]. The first disease association to the MHC dates back to 1973[75], pre-dating any sequencing of the human genome, where different genetic content at the HLA locus in individuals was inferred from serological typing. Furthermore, HLA class I expression levels correlate with tumor progression in breast cancer[76] and accumulation of somatic mutations in HLA class I genes was observed in various types of cancer[77,78]. An enrichment for functional mutations and putatively loss-of-function mutations[79], led to the hypothesis that HLA mutations offer tumorous cells a selective advantage by suppression and/or evasion of immune response, thus enabling tumour progression.

Due to complex and population specific LD patterns across the MHC[80], identifying the exact variants, from commonly co-occurring allele combinations on haplotypes, that are associated to disease may be cumbersome and require special analyses[81]. The ancestral haplotype 8.1, spanning the whole MHC, found at a common frequency within European populations, contains risk conferring alleles for many diseases and was found to be associated as a whole to susceptibility for human immunodeficiency virus (HIV), systemic lupus erythematosus (SLE), insulin dependent diabetes (IDDM), and others[82]. By characterizing genetic variation and LD patterns more extensively across the MHC, all these associations were much more precisely fine-mapped to individual gene alleles and even amino acid alleles and independent signals from different genes were disentangled[83,84,85]. Another example is hereditary haemocromatosis (HH), where an increased iron deposition may result into multi-organ dysfunction. Initially, the disease was found to be associated to an HLA-A allele[86] and only after several more studies were performed, the association was narrowed down to the HLA-H gene, positioned three megabases telomeric of HLA-A and sequence-wise very similar to the initially associated HLA-A allele[87].

Finally, disease associated MHC loci contribute to disease susceptibility in conjunction with other MHC loci[85], but also in conjunction with distal, non-MHC genomic loci. Interactions with the innate immune system, through the killer cell immunoglobulin-like receptors (KIR), located on chromosome 19, may affect pregnancy[88], or malaria progression[89]. Another genomic locus suggested to modulate HLA associations to disease is the chromosome 5 endoplasmic reticulum aminopeptidase (ERAP 1&2) locus, encoding for enzymes that degrade proteins into peptides that are subsequently bound by HLA molecules. Interactions between HLA and ERAP, in association to (auto-immune) disease include ankylosing spondylitis (AS)[90,91], inflammatory bowel disease (IBD) [92], and Behcet disease[90].

Genetic variation across the MHC

With the advent and progress of genome sequencing technologies, it became clear that the MHC is one of the most variable regions of the genome[93]. The extended-MHC was further defined as a 7MB region, that includes the canonical MHC and an additional ~3MB block telomeric to the MHC that shows extensive LD patterns with the MHC. It and encodes more than 400 genetic elements including genes, pseudo-genes and transcripts[93]. Haplotypes of the whole region were sequenced and assembled, using modern high throughput sequencing technologies, in order to characterize and understand population variation[94,95]. Some of these haplotypes are well conserved and reach common frequency in certain populations, and were introduced as alternative scaffolds for the whole MHC locus in the reference genome, to better map genetic variation when analyzing the genome a previously unsequenced individual (https://www.ncbi.nlm.nih.gov/grc/human).

The alternative MHC scaffolds contain more than 44,000 SNVs and indel variants, compared to the human reference haplotype for the MHC[68] (sequenced from a consanguineous lymphoblastoid cell line, denoted by the Human Genome Organisation Gene Nomenclature

Committee -- HGNC -- as the PGF cell line) and comprise of different alleles across the classical HLA loci[96]. The variant density around the classical HLA genes is an order of magnitude larger than across the rest of the genome[97]. Furthermore, structural variation was observed both in coding and non-coding regions, including long indels, retrotransposon mediated indels and recombination, as well as different gene copy numbers[97]. The class II HLA-DR gene complex for example, contains different combinations of any of 9 versions (i.e.: paralogues) of the HLA-DRB gene. These different paralogues of the HLA-DRB gene were introduced through gene duplication and some are actively expressed, while others are pseudogenes, both contributing to the observed gene redundancy across the MHC locus[98].

Most of the genetic variation across the MHC is concentrated within and around the classical HLA genes[97]. The International ImMunoGeneTics (IMGT/HLA) database holds and maintains sequence information about classical and non-classical HLA genes, as well as online tools for summarizing and interrogating this information[99]. Alleles across the HLA loci are described as individual variant alleles (i.e.: single nucleotide or indel variants), but also as alternative sequences for the whole gene, containing many SNV and indel differences between each other. IMGT currently contains the sequences of more than 17,000 alleles of more than 40 genes, although the vast majority of alleles describe the classical class I genes (12,461 alleles) and classical class II genes (4,364 alleles). This unprecedented amount of variation led to the development of a nomenclature that is used to denote each allele, which hierarchically takes into account serological differences between allelic groups, functional aminoacid variation, down to synonymous and intronic variation where the consequence of a variant is harder to interpret (**Figure 1**). Exons that code for the binding pockets of the resulting HLA proteins (exons 2 & 3 for class I and exon 2 for class II) are found to be most variable, with virtually every position being polymorphic[100]. While most of these alleles are rare, 1,122 alleles across the classical loci (plus HLA-DRB3/4/5) were found to reach common frequencies (frequency > 1%) amounting to 14% of the IMGT/HLA database[101]. Furthermore most of the observed classical HLA class I alleles have evolved incrementally through mutation, recombination and genetic material exchange between genes, from a few ancestral alleles[100].

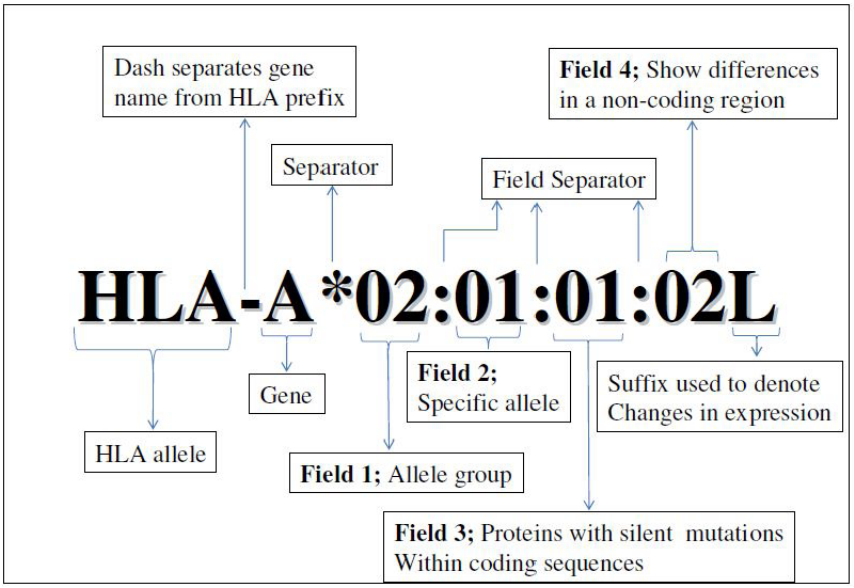


Figure 1: Nomenclature that is used to represent each allele across any HLA gene (adapted from figure2 of[184]).

Maintaining MHC variation

F then arises, by what mechanism is such extensive variation maintained across the MHC and two main (related) theories are generally pursued: (i) the MHC is under balancing selection by exposure of cells (i.e.: or individuals) to pathogens, which in turn favours a pool of alleles for any locus, rather a specific, universally most advantageous allele and (ii) individuals that are heterozygous across the MHC have a selective advantage over homozygous individuals, for being able to respond to a larger number of pathogens. Both these ideas of balancing selection by interaction with pathogens were first suggested and illustrated decades ago when it was observed that individuals with sickle-cell trait manifest protection against malaria and thalassemia [102,103] (although, as it turned out, neither sickle-cell disease nor thalassemia are HLA mediated diseases). More recent analyses of malaria susceptibility have brought further evidence and quantification of how natural selection influences MHC allele frequencies. Thus, the HLA-B*53 allele, that conveys a reduced susceptibility for severe malaria[104] is found in up to 30% of the individuals from Sub-Saharan regions (allelefrequencies.org), which account for ~70% of yearly malaria deaths, whereas the same allele is found in ~4% of the individuals from European populations, where malaria is now extinct. Other evidence for balancing selection include an excess of non-synonymous (i.e.: protein sequence changing) variation compared to synonymous variation across the HLA genes[100] and the large number of common alleles present in any given population. While abundant evidence exists for a heterozygote advantage in other animals such as mice[105–107], this signal has been directly observed more sparsely in humans[108], and it may be hard to disentagle a the advantageous effect of heterozygosity from the effect of one (or more) individual alleles[109].

An additional theory that would serve to explain the extensive variation levels across the HLA loci is preferential mating. Hypotheses about sexual selection have mostly speculative roots and initially concerned behavioural and visible traits[110], rather than the immune system. Haldane[103] in turn, recognized that individuals that carry rare alleles associated to pathogen resistance may acquire a fitness advantage with respect to mating and this will result in a balancing selection of the locus. More recently it has been shown, unequivocally, that several species show preferential mating behaviour, such that the MHC loci of the two mates are divergent[111–113,114] (i.e.: MHC disassortative mating). In humans, testing this hypothesis, repeatedly, did not offer a conclusive answer yet. Studies investigating human MHC disassortative mating fall under two categories. There are studies, such as the "sweaty t-shirts" experiments and others[115,116], that show odour based preference for MHC dissimilar mates. Furthermore, MHC dissimilarity was connected to couple satisfaction and the intention to have children[117]. The limitation of such conclusions is that MHC dissimilarity is associated to questionnaire answers expressing preferences and intentions, and the relation of such preferences to biological outcome cannot be evaluated.

Other studies have analyzed couples with children, where their potential genetic dissimilarity across the MHC would undeniably have an effect on population diversity, but a signal for MHC disassortative mating was far less obvious in these cases and even divergent conclusions were reached[118,119–121]. In **chapter 3**, I show how we address the question of MHC disassortative mating in humans. Using a dataset that is an order of magnitude larger than previous analyses on this topic, and more sensitive characterization of genetic variation across the MHC, we offer a more conclusive result.

Advances in Sequencing Technology

Genome sequencing is the biochemical and computational process through which the DNA sequence of an organism is read. For sequencing the germline genome, typically, the nucleic DNA is extracted from a small number of cells and is considered representative of the organism's germline. Because no sequencing technology can yet sequence entire chromosomes, from one end to the other, the DNA molecules are broken down randomly, into (short) fragments and the sequence of (part of) these fragments is in turn deciphered. The first DNA sequencing technologies were developed in the 1970s. The technology developed by Fred Sanger[122] has been used for several decades as the state-of-the-art sequencing method. Although the initial versions of Sanger sequencing were laborious and slow, and it was virtually impossible to sequence the whole 3.2 billion bases of the (haploid) human genome, the technology was further optimized. The introduction of capillary sequencing led to a boost in throughput and lay the foundation for the Human Genome Project. As test cases, several smaller genomes were first sequenced in the 1990s before the project commenced with the human genome. Developments advanced during the Human Genome Project (HGP)[123] and Sanger/capillary sequencing instruments were able to decipher a DNA molecule with a very high accuracy of > 99.999%, from sequencing reads that spanned up to 1000bp.

A major breakthrough in sequencing technology development came in 2005 with the introduction of massively parallelized sequencing technology[124]. This technology, marketed by 454 Life Sciences, enabled the sequencing of an entire human genome much faster and cheaper and was used for one of the first personal human genome sequences - the genome of J. Craig Venter[125]. . Folowing this milestone, a host of new technologies emerged, under the umbrella term of Next-Generation Sequencing (NGS). These technologies used different implementations of very similar processes, and were able to obtain read-outs of the sequenced genome of varying length and qualities [126–128], but here we will briefly describe the technology that in the meantime emerged as market leader, commercialized by Illumina.

Template preparation for Illumina sequencing

Short read Illumina sequencing falls under a family of technologies termed cyclic reversible termination (CRT)[129]. A prepared library of DNA fragments is primed with pre-designed universal adaptors (short single-stranded artificial DNA sequences) and loaded onto a solid glass slide platform. This platform is in turn populated by densely spaced, bound primers (with sequence complementary to that of the library fragment adaptors). Specific media conditions are used such that the hybridization of fragments at spaced primer sites is favoured. A series of amplification (PCR) cycles follows, where the two strands of a DNA fragment are separated and a polymerase (primed by the primer attached to the glass slide) builds the complementary strand of each resulting single strand template, effectively copying the original fragment. This eventually results in clusters of densely populated copies of the exact same original DNA fragment. Each of these fragments are finally separated into single stranded DNA and primed, at one or both ends, for the sequencing reaction.

Illumina sequencing reaction

The sequencing reaction is also cyclical and resembles an amplification step, in that the single stranded DNA molecules are again copied by the polymerase incorporating bases complementary to the primed template. However, within each cycle, the polymerase incorporates only a single fluorophore dyed, modified nucleobase. Each nucleobase is dyed with a different colour, and microscopic imaging distinguishes what base was added exactly. After this readout, the dye is cleaved and the reaction continues with incorporation of the next nucleobase. DNA sequences within one amplified cluster are copies of the same DNA fragment by construction so, ideally, at each step, the same base is incorporated for all molecules within a cluster. The cluster of clones then serves to boost the signal detected by microscopic imaging and each base readout is a consensus of all the clones within one cluster. This results in a very high per base sequencing accuracy > 99.5%[130]. The sequence of consecutive readouts (signals) from each cluster is outputted as one “read”, to be used in downstream analyses, and each consensus basecall is reported with along with a quality score that is proportional to the certainty of the measurement. The read length is determined by the number of cycles allowed during the sequencing reaction (ranging from 90 to 300 base-pairs, usually much smaller than the initial fragment length of ~800 base-pairs) and the high throughput is determined by the high number of clusters that can be sequenced within

the same reaction (up to 200 million per plate).

There are many sequencing instruments commercialized, that vary in throughput and read length, but also, slightly, in the underlying chemistries used[131]. Furthermore, the fragment clones may be primed and sequenced from one end, or from both ends simultaneously and sequenced in opposite directions. In the latter case, the two equal length reads that are produced from each cluster of clones are two mate-pairs originating from the same fragment, and the unsequenced (i.e.: and thus unknown) DNA sequence that separates them is the insert size. Typically, paired-end sequencing is employed, as being more informative, and the insert size controlled by selecting initial DNA fragments of a specific length. In combination with different, appropriate library preparation protocols, NGS sequencing platforms may also be used to sequence and quantify gene transcripts (RNA-seq)[132] or DNA containing modified basses (such as methylated cytosines)[133].

Short read sequencing errors and biases

The sequencing errors are determined by the key components of the process. First, the polymerases that create the clonal templates, during amplification, are susceptible to errors and incorporating a wrong base can propagate throughout the clones of a cluster. Indeed, base-substitution errors are the predominant error mode for short read sequencing[134]. Secondly, the process by which DNA fragments attach to clonal clusters is a stochastic one, and one of the (potentially different) alleles of a locus may be not captured or under-represented. Furthermore, many times, polymerase chain reaction (PCR), is performed during the library preparation of a DNA sample, in order to increase the raw amount of DNA available for sequencing. For this pre-processing PCR reaction, the DNA fragments are not separated into clusters and any allele representation imbalance that occurs may propagate exponentially. Also, short read sequencing shows systematic under-representation of AT-rich and GC-rich regions of the genome altogether[135,136]. During the sequencing cycles, it is important that all molecules of a cluster are elongated at the same rate; if the polymerase of one molecule fails to incorporate a base or incorporates more bases during a cycle, the molecules progress out of phase and the signal becomes noisy. Lastly, the imaging process becomes increasingly impaired by noise, as the number of cycles progresses and the quality of the base calls drops, as the read elongates.

Single molecule long read sequencing

The strongest limitation of short reads is perhaps, as the name suggests, that each read describes only a very short snippet (i.e.: 90 to 300 base-pairs) of a 3 billion human genome[131]. This fact generates significant limitations in characterising entire human genomes, which are overcome with various degrees of success by complex downstream bioinformatic analysis, as I briefly introduce in the next section.

Alternative sequencing technologies have emerged, that can produce reads in the order of

kilobases (kb) and hundreds of kilobases long, from a genome under investigation. These technologies are generally termed third generation sequencing technologies, or single molecule real-time sequencing techniques. The first such technology was commercialized by Pacific Biosciences (PacBio)[137] in 2011, followed in 2014 by Oxford Nanopore Technologies' (ONT) release of the MinION platform. Although in very different ways, both 3rd generation sequencing technologies differ from NGS approaches in two fundamental ways, which represent the source of their advantages as well as their current limitations: First of all, each measurement of a nucleobase is not the consensus of an amplified population of clones anymore, but the sole measurement of a single nucleobase, as it traverses a detection apparatus (i.e.: single polymerase molecule or protein pore respectively). Secondly, the sequencing process is not discrete anymore (i.e.: NGS elongation is paused after each modified base is incorporated, for a measurement and for cleavage of the nucleobase dye to take place), but measurements are continuously made as the DNA molecule is traversed.

Pacific Biosciences

PacBio uses a similar concept of incorporating fluorophore modified nucleobases into a template DNA sequence as is used for Illumina sequencing. However, instead of the having the DNA template fixed, the polymerase is fixed in a small well with a transparent bottom, through which the just added, labeled nucleobase is imaged, without interference from adjacent bases. Furthermore, the single stranded DNA fragments that are used as template are circularized, such that the same template may be traversed by the polymerase multiple times, in which case a more accurate, consensus sequence is produced. The ability of the polymerase to read the same template multiple times is a function of the length of the template and, typically, longer templates (> 3kb) may only be read once[137]. The single pass per-base accuracy of the latest PacBio Sequel system is reported to be 85-87%[138]; average read-lengths of 10kb are reported along with a read N50 of 20kb (i.e.: minimum read length such that half of the sequencing data is contained within reads longer or equal than this value) and a longest read of 92.7kb[139].

Nanopore third generation sequencing

Oxford Nanopore Technologies (ONT) implemented a completely different sequencing concept[140]. They determine the content of a sequence of nucleotides, from its ionic current signature trace, as it traverses a nano-meter sized protein pore, embedded in a polar membrane. The membrane does not permit the flow of ions (electrically charged anorganic molecules) from one side to the other and it is placed in an ionic solution (i.e.: containing K^+ and Cl^- ions). Protein pores then form physical channels through this membrane and a constant voltage (i.e.: in the order of milli Volts) is maintained across the membrane, which drives the translocation of electrically charged ions from one side to the other. As a result, an ionic current (i.e.: in the order of pico Amperes) is measured through the pore by embedded field effect transistors. Nucleic acid molecules are electrically (negatively) charged (i.e.: due to individual nucleic bases as well as the DNA backbone structure) so the potential difference across the membrane drives them through the pore as well. As a DNA molecule traverses the pore, it

physically obstructs the passage of (the much smaller) ions in the solution, and a lower ionic current amplitude is measured through the pore (**Figure 2**). This concept of sequencing was marketed my Oxford Nanopore Technologies in the form of the portable MinION sequencer and subsequent sequencers termed GridION and PromethION.

The idea for this original sequencing method was sparked in the 1970s by the observation that the membrane of biological cells contains protein pores that facilitate the flow of nutrients and ions to the cell (for a historical review[140]). During the 1990s, systematic efforts and prototyping of nanopore sequencing used simple homopolymer sequences (poly-As, poly-Cs, etc.) and showed that different DNA molecules (in terms of length and base composition) modulate the current amplitude measured through the pore differently[141,142]. This finding served as first empirical evidence that discrimination of DNA content passing through the pore is possible and it was later followed by proof that the induced current changes are sensitive enough to measure a one-base difference between two DNA molecules[143,144].

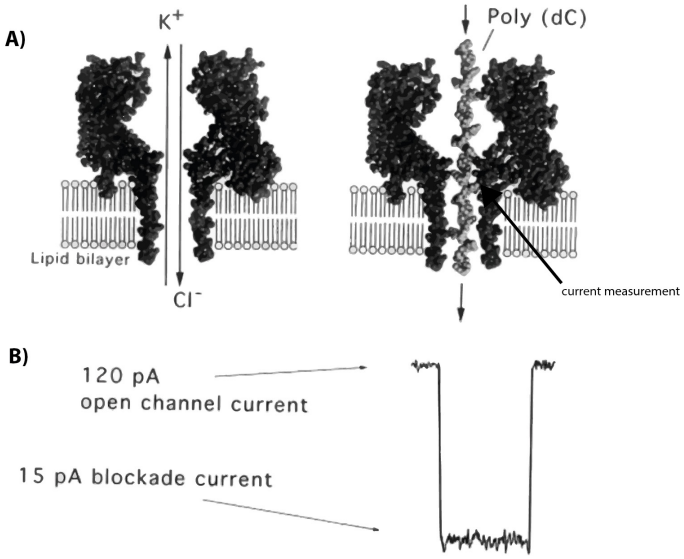


Figure 2: Illustration of a nanopore sequencing media (A) and output measurement (B).
A) shows a non-obstructed membrane embedded protein nanopore, through which K- and Cl+ ions can freely flow when a voltage is applied across the membrane is applied (left) and a nanopore obstructed by a poly-C homopolymer traversing it. B) shows the amplitude of the ionic current flowing through the pore. While the DNA homopolymer traverses the pore, it blocks the normal flow of ions and a lower ionic current amplitude is measured. Figure adapted from figure1 of[140]

A limiting factor towards single base discrimination was the very high speed at which a DNA molecule traverses the nanopore. On a free pass, it is estimated that each DNA base spends ~1 micro-second in the nanopore, which is too short for accurate read-outs to be measured[140]. Adding a motor protein to the DNA molecule to be sequenced, allowed reducing the speed at which DNA is translocated through the pore by three orders of magnitude, increasing the time that a base spends in the pore to miliseconds[145]. Furthermore, while the width of a nanopore is small enough to ensure that nucleotides pass sequentially, one at a time, the length of the narrow channel forbids the measurement of each individual base's modulation of the current. Rather, the influence of a stretch of consecutive bases (i.e.: a k-mer, where k denotes the length of the sequence) is measured. Initially, the current measurement through the pore was estimated to be determined by 15 consecutive bases (15-mers), using an alpha-hemolysin (α-HL) pore.

Many pores have been tested, in their naturally occurring structure or mutated versions. The nanopore used in current releases is a modified *Escherichia coli* (*E. coli*) *CsgG* pore, which is able to produce readouts corresponding to individual 5-mers. As a DNA molecule traverses the pore, measurements of the ionic current across the pore cross-section are recorded. These measurements are recorded at a very high frequency, such that many (ideally similar) measurements are recorded in the (milli-seconds) time that it takes one base to traverse the pore. This trace of ionic current measurements is the raw output of the sequencer and

ONT calls this electric wave representation “squiggle space”. The translation to symbolic, genetic bases (i.e.: base calling) that offers the final read is subsequently performed through computational discrimination methods, but it is a non-trivial task and still subject to research.

ONT commercializes its sequencing technology through a number of different platforms. The most commonly used one is the MinION, that is able to sequence using a flow cell with 512 pores. Each flow cell can be run for up to 48 hours and is not reusable. The chemistry version released is R9 (R9.4 or R9.5), which uses a CsgG pore that is able to read DNA molecules at a translocation speed of 450 bases per second. The throughput from an R9 MinION flow-cell thus has a theoretical upper bound of ~40Gb (i.e.: 512 pores continuously sequencing for 48 hours). In practice, the average throughput was found to be ~2.3 Gb although it may vary considerably between different runs[146]. Other platforms commercialized by ONT are parallelizations of flow cells similar to the MinION flow cell, sometimes with different number of pores, conceptualized in order to increase throughput even further. Thus, the GridION is a device that simultaneously runs 5 MinION flow cells. The PromethION was designed to be suitable for population sequencing, as it may run up to 48 flow cells in parallel, each containing 3,000 pores. It can generate tera-bases of sequence in one run and is currently released only under an early access programme to various research groups.

Nanopore library preparation

Pre-sequencing library preparation is minimal for nanopore sequencing. Once extracted from the cell, the DNA molecules may be sheared into fragments of desired size (similar to NGS practices), followed by an end-repair reaction. A PCR can be done, prior to sequencing, to increase the raw amount of DNA. If enough DNA is collected from cells, then PCR is not required and the native (non-amplified) DNA fragments may be sequenced directly. An adaptor that initiates the pore traversal is ligated to each end of each fragment, along with a motor-protein that controls the traversal speed and separates the double stranded DNA, allowing only one strand at a time to pass through the pore. The ONT platforms can sequence one strand of double-stranded DNA molecules, producing 1-dimensional reads (1D read) or both strands of double-stranded DNA molecules, producing 2-dimensional reads (2D reads, now superseded by so-called $1D^2$ reads). When $1D^2$ reads are produced, a consensus call is made for each base, based on sequence information from both strands, further increasing accuracy. Library preparation for $1D^2$ sequencing runs uses a modified adaptor, that keeps the complementary strand close to the pore, as the template strand is being sequenced. The complementary strand will thus most likely be the next molecule to be sequenced, after the template strand was processed. After the first strand has been sequenced, its motor protein detaches from the pore and the complementary strand or a different molecule follows through.

Long reads

There are three main advantages that nanopore sequencing delivers, in comparison to NGS methods, that have also sparked a lot of enthusiasm in research. First and foremost, the

technology offers no conceptual limit for the read length that may be produced (i.e.: other than the chromosome length of the genome under analysis). By design, the nanopores sequence DNA fragments supplied until one of two things happens: the pore gets entirely blocked by the DNA fragment and cannot sequence further, or the fragment is finished (broken, e.g. as a result of a nick in the DNA). A major determinant of the read-length distribution outputted by nanopore sequencing seems to be the DNA library preparation step, which is rapidly evolving in part by improved methods for DNA extraction that are aimed at maintaining high-molecular weight DNA. Indeed, while standard protocols report a median read length of 7-9kb[147,148] and a read N50 (i.e.: minimum read length such that half the sequencing data is contained within reads longer or equal than it) of 17kb[148], we and others are able to reproducibly obtain sequencing runs with a read N50s of 50 kb or even 99.7kb[146]. The longest reported nanopore read so far is > 1Mb.

The long nanopore reads have proved very useful in accurately reconstructing the sequence and structure of whole genomes, starting with smaller genomes such as the bacterial *E. coli* genome (4.6Mb)[149], a yeast genome (~12Mb)[150] and recently, larger genomes such as the European eel (~860Mb)[151] and the human genome (3.2Gb) [146]**.** Furthermore, long reads may span the entire RNA transcripts of a gene, readily offering information about gene expression levels, about different gene isoforms that an individual may express, as well as detecting complex events such as fusions between two different genes[152].

Everything is in the current wave

A DNA molecule may contain additional, biologically relevant chemical modifications, beyond its mere base composition. For example cytosines (Cs) may have a methyl group attached, which modulates the binding of DNA transcription machinery, at that site. When DNA is copied synthetically, through PCR or linear amplification, these additional modifications are not preserved. Therefore, by being able to sequence DNA that is not cloned, nanopore technologies can be used to read these base modifications as well. Furthermore, no modification to the sequencing process is required; if such base modifications are present on a particular base, they modify both its physical shape and volume as well as its electrical charge, and these changes are directly reflected in the ionic current measurements, as the molecule traverses the pore. The information may then be extracted from the outputted electric trace. This concept was successfully applied to detection of methylated C bases and an efficiency similar to other, state of the art methods was achieved[153]. The signature produced by these base modifications is present in the raw electric trace produced by a pore, regardless of whether their detection is intended or not, and this fact produces a trade-off: if one wishes to detect these base modification, the discrimination task becomes harder, as more outcomes than the mere four bases (A, C, G and T) need to be discriminated (i.e.: methylated C) and if such modifications are to be ignored, then their current signature generates noise.

Portability and entry costs

The third reason, for which nanopore sequencing has sparked enthusiasm, is the miniatur-

ization of the technology, in sharp contrast to the average sequencing machine size, and the (very) low initial investment. With the MinION being the largest USB stick that you have ever seen and ONT's release of rapid library preparation protocols, it has become feasible to bring sequencing to remote places or to areas from where extracting biological samples might pose high risks of viral spread. Nanopore sequencing was recently used to successfully sequence fast spreading viruses such as Ebola[154] and Zika[155] and the migration of different strains across regions was illustrated.

High throughput sequencing machines (i.e.: that can sequence an entire human genome), either NGS or 3rd generation sequencers, typically cost more than 200,000 US dollars. Subsequent use of the machines is much cheaper, but the high initial cost drives the market towards specialized sequencing centers. Although the MinION is advertised to be free, an initial investment of 1,000 US dollars is required, by the mandatory purchase of the first two flow cells. This extreme difference in initial investment may stimulate more decentralized sequencing practices.

Complex translation of raw nanopore data to a base sequence

A current limitation of nanopore sequencing is the (still) relatively high error rate (i..e: compared to NGS and/or capillary sequencing) and this results from the interplay of two factors: the sensitivity and stability of the ionic current measurements during sequencing and the downstream translation of the timeseries current trace into a discrete sequence of nucleobases. Small pore biases and fluctuating speed of the DNA molecule traversing the pore introduce noise in the measurements, that have to be overcome upon basecalling. Early implementations of nanopore basecalling, such as Metrichor and Nanocall[156] relied on Hidden Markov Models (HMMs) to produce the most likely sequence of DNA corresponding to a sequence of measurements. Because the ionic current sampling frequency across the pore is much higher than the frequency at which the content inside the pore changes (i.e.: the DNA translocation speed), more consecutive measurements correspond to each k-mer in the pore (i.e.: currently, approximately 9). These early algorithms first perform a segmentation of the raw signal, and determine when the content of the pore has changed by comparing the measurement at some point in time to the measurements immediately preceding it, to identify sudden changes in current level. These sudden changes are termed events and the average current value between two events is then used by the HMM to estimate the most likely k-mer traversing the pore, for every interval, and thus reconstruct the original DNA sequence. Subsequent implementations such as Albacore, Scrappie or DeepNano[157] relied on recurrent neural networks (RNNs) to discriminate between k-mers, but used the same (or slightly modified) segmentation process, that requires to specifically assign the measurements corresponding to each k-mer, prior to the discrimination step. Because the employed segmentation is based on differences in the measured ionic current, and highly repetitive DNA regions (i.e.: homopolymers) are characterized by the lack of sequence diversity, base-calling typically performs very poorly in these regions, and the data is not reliable for downstream analyses. Scrappie predicts k-mer identity and duration simultaneously, and

is therefore increases sequencing accuracy across highly repeated regions as well. The first implementation that is able to bypass segmentation and use the full, unprocessed raw signal to produce basecalls is Chiron[158]. It uses a combination of convolutional neural networks (CNNs) and RNNs and is able to achieve results comparable to the current standard algorithm Albacore, although still producing an excess of deletion errors.

Biases of nanopore sequencing errors

As nanopore sequencing machines output raw ionic current measurements, sequencing accuracy is usually computed after base calling, by comparing the read bases to the expected sequence of bases (i.e.: read percent identity [PID] to a reference sequence). The current average PID (i.e.: accuracy) of nanopore reads, for R9 chemistries, varies between 84% and 90%[146,159], regardless of whether 1D or $1D^2$ sequencing is considered. The error rate is driven by short indel errors (predominantly deletions), that account for approximately two thirds of the errors (**chapter 4**). In low complexity regions of the genome, containing homopolymer stretches or short tandem repeats, indel errors are 2.6 times and 1.4 times more frequent respectively. Indel errors are thought to be caused by 2 (related) factors: 1) irregularities in the speed of the DNA molecule traversing the pore and 2) incorrect estimation, at base calling, of the number of bases observed; in sequencing a homopolymer for example, the ion current level does not change for the duration of the homopolymer traversal, which makes it harder to estimate how many bases exactly were sequenced. The error rate does not increase as the sequencing of each read progresses, which is crucial for long read sequencing. Lastly, nanopore sequencing, displays much less sequencing bias, against GC and AT rich regions of the genome (i.e.: than NGS technologies) offering the opportunity to sequence and characterize previously inaccessible regions of the genome (**chapter 4**).

Bioinformatic analysis of Whole Genome Sequencing data

Sequencing an entire human genome, whether using short read or long read technologies, produces vast amounts of data. A central idea to whole genome sequencing (WGS) is that since any sequencing technology is, to its own extent, prone to errors, capturing the same genomic region in multiple sequenced reads will offer more evidence of the true underlying sequence, and the sequencing errors that may (randomly) occur at any locus, will thus be in the minority. Tens of billions of bases are sequenced within all the reads produced for one genome, and computer storage requirements are in the order of ~100-200GB per genome. The number of reads that offer information about a specific genomic position is referred to as the coverage across that position.

The main challenge in leveraging the sequencing data to decipher an individual's genome is that, while each read contains information regarding a small locus of the genome, the information about which locus exactly it describes, is lost. There are two main approaches in reconstructing a genome from sequencing data, and various hybrid approaches: 1) a ref-

erence based approach of identifying variation with respect to the human reference genome and 2) a reference free assembly approach, where an individual's genome is first reconstructed as accurately as possible from read data and only afterwards it is compared to the reference genome.

The human reference genome is currently an essential resource that resulted from the Human Genome Project. In its minimal form, it contains the DNA sequence of one copy of each human chromosome, including X and Y chromosomes and the mitochondrial DNA. While it is built such that its structure is representative of a general human genome, the actual sequence of the human reference genome is not necessarily any individual's genome, as DNA from five individuals with European ancestry was used to build it. Beyond enabling computationally feasible analyses of whole genome sequencing data, the most important function of the human reference genome is to serve as a "reference" (i.e.: coordinate system) for representing genetic variation between individuals. The human reference genome is continuously maintained and updated by an international collaboration reunited under the Genome Reference Consortium (GRC). For the remainder of this introduction I will detail reference based variant calling as most relevant for the work presented here.

Reference based variant calling

In a typical workflow for the analysis of short of long read human sequencing data, each sequencing read is first mapped to the reference genome, to find the genome region that it describes. Mapping is performed based solely on sequence similarity, between the read and the reference genome. Reads may differ from the reference genome due to genuine genetic variation of the individual sequenced or due to sequencing errors (**Figure 3B**). The interplay of underlying sequencing variation and sequencing errors determine how well reads can be aligned to different regions of the genome. Regardless of the underlying reason, mapping algorithms need to allow for such (small) differences when comparing the two sequences. The first implementations of DNA sequence alignment algorithms were FASTA[160] and BLAST[161]. Most alignment algorithms build on the idea of finding short substrings of the read to be aligned (i.e.: anchors/words/seeds) that match some region of the reference genome (3.2 billion bases for the human reference) very well, and then locally align the rest of the read to such identified target regions. The best match is then reported, along with a quantitative score, that is proportional to the quality of the alignment (i.e.: usually proportional to the confidence that the respective read aligns to the reported position and not somewhere else in the genome). When a read can be matched to different regions of the genome equally well (or bad), it is reported with a mapping quality of zero and is usually discarded from subsequent analyses. Many alignment algorithms exist to date, that are optimized for speed (i.e.: they index the human reference genome for faster search of matches) and for the properties of the read-data being aligned. Thus, due to the lower error rate, fewer differences with respect to the reference genome are expected and allowed for short read NGS data, and indel differences are penalized higher. For long read (nanopore) data indel errors are penalized less (i.e.: as being the predominant error mode), more errors are allowed overall and

the length of the seeds can be increased to obtain better initial hits. While Burrow-Wheeler Aligner (BWA)[162] has become a standard for fast and accurate alignment of short read NGS data, the unique properties of long read data have sparked the development of new algorithms and there is no clear best practice yet, as alignment may influence the accuracy of downstream analysis results considerably[159,163]. For paired-end short read sequencing, the additional information that the approximate genomic distance between the two mates of a pair is known, can be leveraged to obtain a good alignment for both reads, when only one of them aligns well independently. Lastly, alignment around indel variation is prone to errors, due to sequence homology around the indel or in repeated regions of the genome. Typically, an additional local realignment is performed, around a set of known indel sites (**Figure 3A**).

Once the read data is aligned, statistical algorithms are used to evaluate the evidence for genetic variation at each genomic position, and, subsequently, genotype an individual. Under the assumption that sequencing errors are (generally) randomly distributed, a site is considered variant if there is systematic evidence that an allele different than the reference genome allele is present (**Figure 3B**). Because a variant allele might not be properly captured within an individual (due to PCR or sequencing bias), in order to increase the power of detecting variant sites, the sequencing data of multiple individuals can be evaluated jointly, for each position (i.e.: GATK UnifiedGenotyper). Once a genomic locus is identified as variant and the possible alleles are determined, the likelihood of each possible genotype (i.e.: any diploid combination of alleles) is computed, by taking into account the base composition and sequencing quality of each read spanning that locus. The most likely genotype is outputted, along with a measure of confidence in the call made (i.e.: genotype likelihood -- GL). The genotype likelihood is proportional to the amount and quality of the read evidence that supports the respective genotype (i.e.: contrasted to evidence supporting other genotype assignments); thus, read coverage and proportional capturing of alleles are crucial determinants of genotyping accuracy. When sequencing quality or coverage is low, additional population level information may be leveraged to increase genotyping accuracy, although the individual's sequencing data remains the main determinant. Specifically, population allele frequencies can be used to compute a prior expectation for each allele combination, for an individual (i.e.: GATK UnifiedGenotyper). Further, machine learning filters may be applied that discriminate between true variant sites and false positive variant sites based on features related to coverage and properties of the reads spanning the variant site (GATK VariantQualityScoreRecalibration -- VQSR).

De novo mutations

By definition, in order to detect *de novo* germline mutations within a genome, the data of an individual needs to be considered jointly with the genetic data of his or her parents. In its simplest form, one may identify DNMs by evaluating the genotype calls of both parents and the offspring for mendelian inheritance inconsistencies (i.e.: the offspring has an allele combination that cannot have been inherited from the parents). Because the per-base estimated

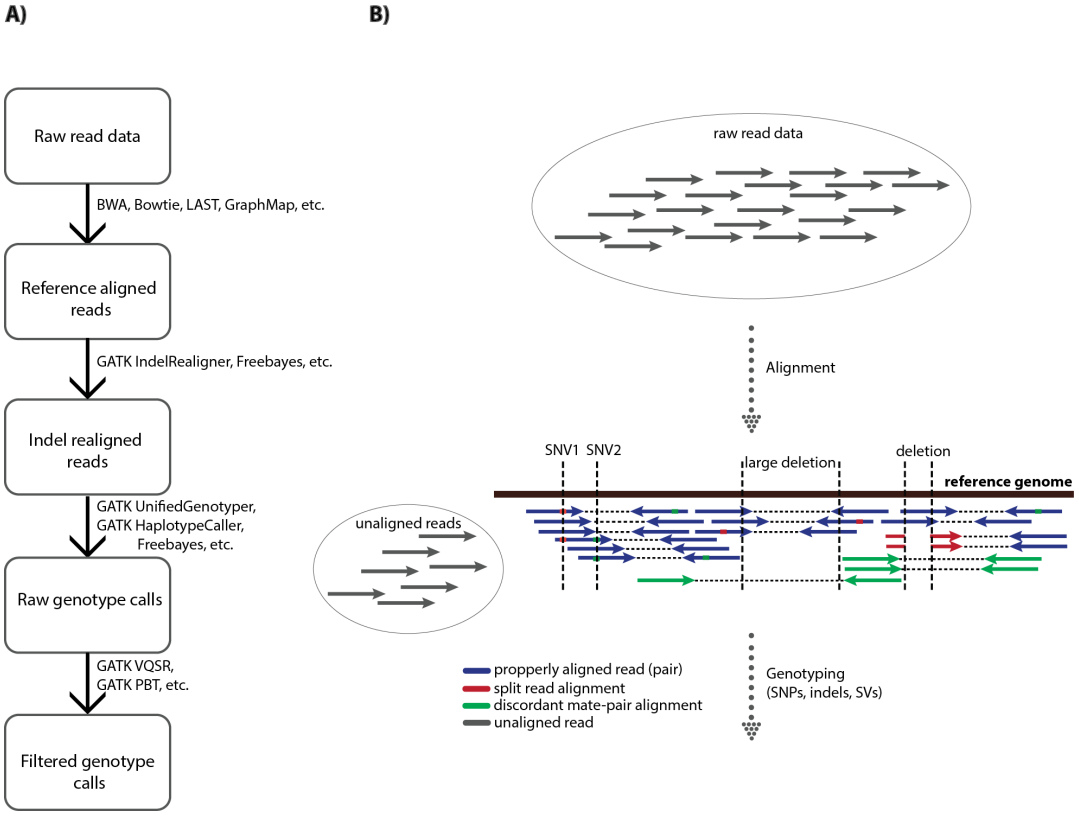


Figure 3: Bioinformatic analysis pipeline
A) Typical steps in a whole genome sequencing analysis pipeline, from raw read data to genotype data. Boxes mark intermediate, sequential results and the arrows indicate commonly used tools for each step. Conceptually, the same pipeline can be used in the analysis of long read data, but the tools used for each step may differ. B) Illustration of short read, paired-end sequencing data alignment and the information that it holds towards subsequent genotyping. Uniform read colour marks no sequence divergence from the reference genome; and single base substitutions are marked on the reads.

probability for a *de novo* event ($1.8 * 10^{-8}$) is much lower than the sequencing error rate of NGS technologies (10^{-1}) the detection of DNMs is polluted by many false positives. As many as 6 times more false positive DNMs may be found in the raw genotype calls[35]. DNM calling algorithms[164,165] further leverage the genotype quality scores of the trio individuals and evaluate the probability of a mis-genotype in one of the individuals, against the (low) probability of a true *de novo* event.

Structural variant calling

Individual base differences and short indels are genetic variation that affect a few bases only, and are typically captured within a (short) read alignment, without compromising the alignment. Structural variation by contrast, affects the local composition of the genome to an extent that can make local alignment of short read data impossible. There are three sources of information that can be exploited from short read data alignment, to identify SV variant

sites and to genotype them: split-read alignments, discordant mate-pair alignment and average depth of coverage across a region (**Figure 3B**). SVs disrupt the expected linearity of the genome, as described by the human reference genome. In NGS data, they are detectable as breakpoint junctions (BPs), where two, non-adjacent (i.e.: as described in the reference genome) sequences of DNA are now connected, or joined, in an individual's genome (**Figure 4**).

Split read alignments are the most direct form of information that can be observed in this case, allowing for direct reconstruction of the investigated genome. In these cases, alignment algorithms find that two adjacent segments of a read align optimally at different places in the genome, and therefore split the read into two (or more) subsequences and output the best alignment for each. Depending on the operational definition for an SV that is used (i.e.: events larger than 20 or 50 base-pairs respectively), SV evidence may also be captured within gapped read alignments (typically for deletions < ~40 base-pairs, when most aligners do not split a read into but mark the deletion in one alignment instead). When a read is split during alignment, the reference genome placement of the two alignments, and their orientation (i.e.: aligning to the forward strand of the reference genome or the reverse strand respectively), determine the location and type of the breakpoint junction (**Figure 4**). Discordant mate-pair alignments offer similar information but in an indirect manner. If, for example, there is a large deletion or insertion in the part of a DNA fragment that is not sequenced (i.e.: the insert of a paired-end read), this variation will generate a significant change (decrease for deletions and increase for insertions respectively) in the insert size observed between the two mate-pairs, after alignment. It can thus be inferred, that somewhere between the two mate-pair alignments a BP occured (**Figure 3B**). Finally, if a particular region of the reference genome is duplicated within the investigated genome, upon alignment, the reads originating from any of the copies, will align to the same (single) region of the reference genome, producing a proportional excess of coverage across that region. Conversely, if a genome contains less copies of a particular region (i.e.: than the reference genome), then coverage will be proportionally lower across the region. Read depth of coverage is thus informative for copy number variation (CNV) across the genome.

A myriad of tools have emerged for generating high confidence sets of SVs from NGS data, but, due to the above described data particularities, no universal tool exists, to date, that may detect all types of structural variation[166]. In particular CNVs and insertion typically require dedicated attention. There are numerous bioinformatic tools dedicated to the accurate detection of CNVs genome wide [167–170]. They use fluctuations in depth of coverage across the genome to identify CNV sites, estimate their exact copy number and accurately define their borders. Furthermore, some CNV detection algorithms model and correct for technology specific sources of bias (i.e.: such as GC content and/or repeat content of each genomic region) that influence read coverage independently of the underlying copy number, and hence constitute noise[170].

The presence and genomic location of an insertion can be estimated either from a split-aligned read (**Figure 4**) or from a discordant mate-pair alignment, where one of the read mates aligns to the reference genome, but the second mate-pair does not align (i.e.: because it contains the inserted sequence, that is not present in the reference genome) or indicates a larger insert size. The sequence of a detected insertion however, in particular for insertions larger than the read length, such as retrotransposon-mediated insertions, that are estimated to occur very often in the population[171], is not easily retrieved. The reads that do not align to the human reference genome, including unaligned parts of split reads and unaligned mates of discordant mate pairs are assembled to determine the sequence and length of large insertions; their position in the reference genome is then determined by the aligned segment of a split read, or the aligned mate of mate-pair, that serve as an "anchor". Long read sequencing simplifies the reconstruction of large insertions, because a much larger proportion (i.e.: or all) of the inserted sequence may be covered by each read that spans the insertion. A comparison between short-read and long-read data revealed that ~58% of large insertions (i..e: > 2kb) are only called in the long read data[172].

There is a myriad of general purpose SV detection tools using NGS data, that combine and aggregate information from split-read alignments and discordant mate-pair alignments to produce candidate SV sites and genotypes[55,173–177]. A 2016 review[166] indexes 50 structural variant tools to date. Despite all the invested effort, accurate SV detection remains cumbersome, with Pindel[178], for example, reporting up to two million false positive SV candidates per genome[172]. Structural variant analysis groups of large sequencing projects have converged on building complex pipelines for SV detection. They use multiple independent SV calling algorithms that together capture the full spectrum of structural variation and they report high confidence consensus sets[39,48,179,49]. While they offer very valuable guidelines, obtaining such high quality SV consensus sets are subject to engineering the optimal rules for the dataset under analysis. Efforts to make accurate SV calling more tractable resulted in post-genotyping frameworks[180] and classification algorithms[181] that are able to integrate SV calls from different algorithms. Lastly, in order to detect structural variants that affect the long range structure of the genome, such as complex chromothripsis events, special sequencing library preparation protocols are used, that generate DNA fragments up to 10kb long. Sequencing such long fragments offers the necessary long range information needed to capture complex events[182].

Long read genome sequencing offers the potential to greatly enable and simplify the genotyping of SVs. First of all, by sequencing the whole genomic fragment, one does not need to make genetic inferences, based on not-sequenced regions (i.e.: the insert of paired-end sequencing). Extracting all information from the (long) split-read alignments, enables a unified approach to SV detection and an increased accuracy in determining the exact BP position (**Figure 4**). The lower per-base sequencing quality is not particularly relevant for SV detection, as the reads need only be accurate enough for accurate mapping. Furthermore, highly repetitive genomic regions that are hard to resolve from short read sequencing, be-

cause the read-length is shorter than the repeat span, may now be bridged by long reads that enable accurate evaluation of the underlying sequence. Indeed, long read sequencing has increased the number of SVs that are detected in the human genome[172];[146,172,183] and an enrichment of SVs within highly repeated regions of the genome was reported every time. A particular class of structural variation, where long-read data may substantially improve genotyping is SVs that affect 50-200 bases. These short SVs are hard to map with short read data because the length of the SV event is close enough to the length of standard short read data (150 base-pairs) to compromise alignment and are indeed underrepresented in SV datasets[39]. Furthermore, complex structural variation events, and copy-number neutral SVs are hard to detect with short read data, ~58% of such events are only captured with long read sequencing.[172].

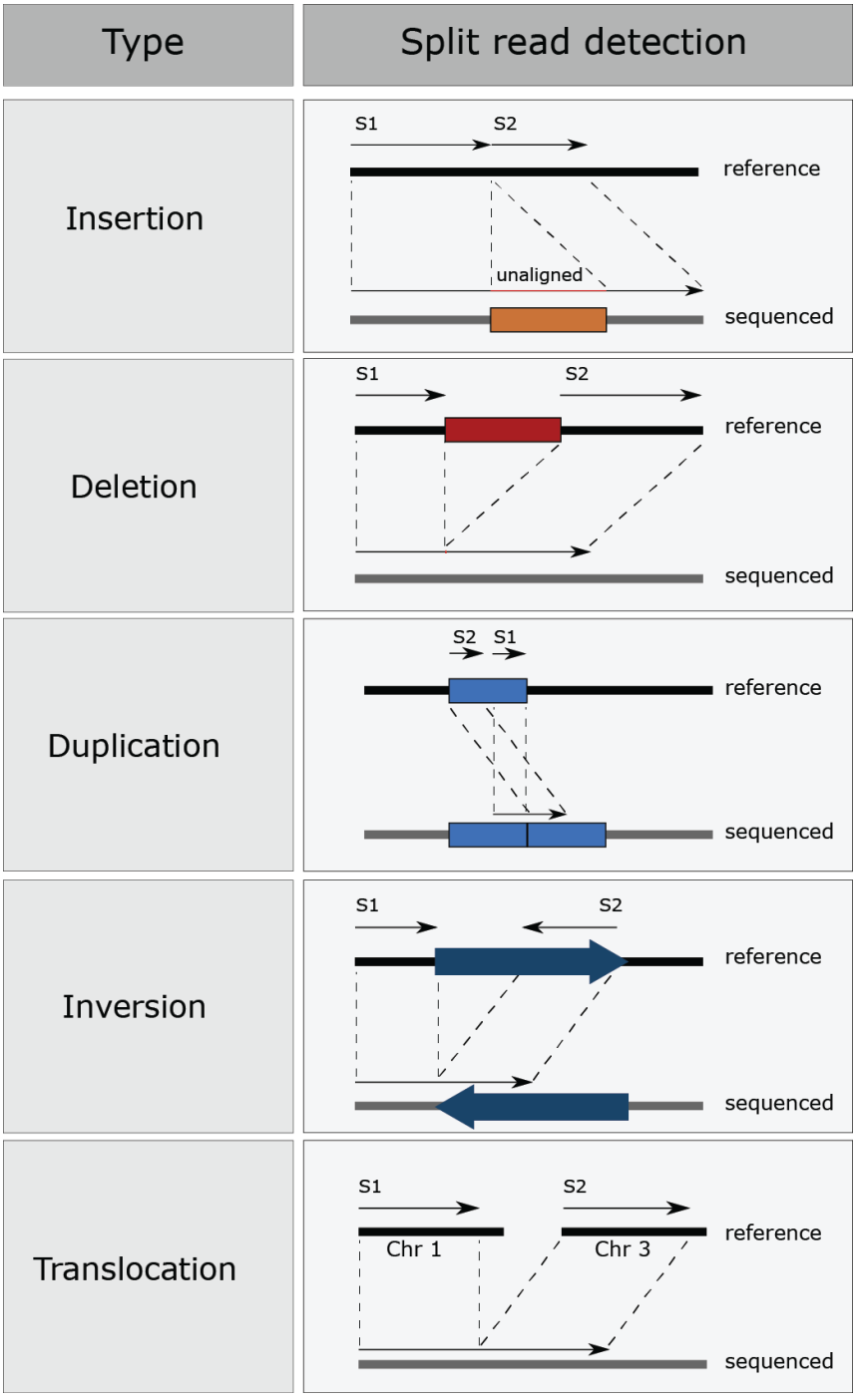


Figure 4: Split read detection for all basic types of structural variants.
Black and grey lines are the long range forward strand of the reference genome and of the genome under investigation respectively. Black arrows on the sequenced genome are sequenced reads. Black arrows on the reference genome are read segment alignments (denotes S1 and S2) to the reference genome, across the SV breakpoint. Arrow direction indicates strand that it aligns to (right direction for forward strand and left direction for reverse strand). Figure adapted from supplementary figure 11 of **chapter 4**.

In this thesis

In **chapter 2** I present PhaseByTransmission, an algorithm that enables accurate detection of de novo events, from whole genome next generation sequencing data. I show how we are able to obtain higher sensitivity than other DNM algorithms in the literature, while maintaining a very good specificity, especially in lower coverage data and on the X chromosome. I show how we successfully apply it to the Genome of the Netherlands dataset to obtain accurate de novo SNV calls across the X chromosome of 246 offspring validating the known increase of DNM mutations with increasing paternal age, for the X chromosome. In **chapter 3** I show how we use the whole genome sequencing data of the 248 couples of the Genome of The Netherlands dataset to answer the recurrent question of preferential MHC mediated mating in human populations. Specifically, I show how we use an order of magnitude larger dataset than previous studies did, to refute the claim that preferential mating serves to increase MHC diversity in our population. We use the whole allele frequency spectrum that WGS captures, including SNVs and indels and show that effects such as subtle population stratification in our sample do not seclude an underlying signal. In **chapter 4** I turn my attention to the emerging long read nanopore sequencing technology and show, for the first time, that nanopore sequencing quality and throughput are mature enough to enable accurate investigation of structural variation in patient genomes. I illustrate how existing pipelines and software tools can be adapted to overcome the particular limitations of the technology and exploit its advantages. We are able to accurately resolve the complex chromothripsis events in both our patients with increased sensitivity over conventional NGS approaches, from a lower coverage of 11-16x (i.e.: compared to an average NGS coverage of 30x). Furthermore, I show how nanopore long read sequencing can be used agnostically, to obtain genome-wide sets of SVs that match any state-of-the-art quality threshold, with precision > 95% and sensitivity > 72%, and how the long reads may be used to obtain genome-wide SNV and indel phasing that matches current statistical phasing in terms of accuracy and connectivity. I conclude with **chapter 5**, where I discuss the implications of these results and perspectives that they open.

References

1. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171,** 737–738 (1953).

2. Schrödinger, E. & Penrose, R. *What is Life?: With Mind and Matter and Autobiographical Sketches*. (Cambridge University Press, 1992).

3. Lieber, M. R. The Mechanism of Human Nonhomologous DNA End Joining. *J. Biol. Chem.* **283,** 1–5 (2007).

4. Lieber, M. R. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. *Annu. Rev. Biochem.* **79,** 181–211 (2010).

5. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1,** 40–47 (2000).

6. Shendure, J. & Akey, J. M. The origins, determinants, and consequences of human mutations. *Science* **349,** 1478–1483 (2015).

7. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8,** 610–618 (2007).

8. Michaelson, J. J. *et al.* Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. *Cell* **151,** 1431–1442 (2012).

9. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).

10. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47,** 822–826 (2015).

11. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. *Trends Genet.* **29,** 575–584 (2013).

12. Chan, K. & Gordenin, D. A. Clusters of Multiple Mutations: Incidence and Molecular Mechanisms. *Annu. Rev. Genet.* **49,** 243–267 (2015).

13. Goldmann, J. M. *et al.* Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48,** 935–939 (2016).

14. Carvalho, C. M. B. *et al.* Replicative mechanisms for CNV formation are error prone. *Nat. Genet.* **45,** 1319–1326 (2013).

15. Neumann, R., Lawson, V. E. & Jeffreys, A. J. Dynamics and processes of copy number instability in human -globin genes. *Proceedings of the National Academy of Sciences* **107,** 8304–8309 (2010).

16. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.* **42,** 483–485 (2010).

17. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42,** 790–793 (2010).

18. Hoischen, A. *et al.* De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat. Genet.* **43,** 729–731 (2011).

19. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97,** 199–215 (2015).

20. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* **14,** 681–691 (2013).

21. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485,** 242–245 (2012).

22. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515,** 216–221 (2014).

23. Epi4K Consortium & Project, E. P. De novo mutations in epileptic encephalopathies. *Nature* **501,** 217–221 (2013).

24. de Ligt, J. *et al.* Diagnostic Exome Sequencing in Persons With Severe Intellectual Disability. *Obstet. Gynecol. Surv.* **68,** 191–193 (2013).

25. Vissers, L. E. L. M., Lisenka E L, Gilissen, C. & Veltman, J. A. Genetic studies in intellectual disability and related disorders. *Nat. Rev. Genet.* **17,** 9–18 (2015).

26. Wilfert, A. B., Sulovari, A., Turner, T. N., Coe, B. P. & Eichler, E. E. Recurrent de novo mutations in neurodevelopmental disorders: properties and clinical implications. *Genome Med.* **9,** 101 (2017).

27. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46,** 944–950 (2014).

28. Turner, T. N. *et al.* Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171,** 710–722.e12 (2017).

29. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).

30. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485,** 246–250 (2012).

31. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380,** 1674–1682 (2012).

32. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

33. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).

34. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

35. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–825 (2014).

36. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet.* **5,** e1000529 (2009).

37. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

38. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12,** 745–755 (2011).

39. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).

40. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).

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

42. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7,** 85–97 (2006).

43. Brand, H. *et al.* Paired-Duplication Signatures Mark Cryptic Inversions and Other Complex Structural Variation. *Am. J. Hum. Genet.* **97,** 170–176 (2015).

44. Brand, H. *et al.* Cryptic and complex chromosomal aberrations in early-onset neuropsychiatric disorders. *Am. J. Hum. Genet.* **95,** 454–461 (2014).

45. Carvalho, C. M. B. *et al.* Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat. Genet.* **43,** 1074–1081 (2011).

46. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517,** 608–611 (2015).

47. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–467 (2005).

48. Kloosterman, W. P. *et al.* Characteristics of de novo structural changes in the human genome. *Genome Res.* **25,** 792–801 (2015).

49. Hehir-Kwa, J. Y. *et al.* A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* **7,** 12989 (2016).

50. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14,** 125–138 (2013).

51. McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered

IRGM expression and Crohn's disease. *Nat. Genet.* **40,** 1107–1112 (2008).

52. Williams, N. M. *et al.* Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *Lancet* **376,** 1401–1408 (2010).

53. The Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464,** 713–720 (2010).

54. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011).

55. Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153,** 919–929 (2013).

56. Moncunill, V. *et al.* Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* **32,** 1106–1112 (2014).

57. Maher, C. A. & Wilson, R. K. Chromothripsis and human disease: piecing together the shattering process. *Cell* **148,** 29–32 (2012).

58. Kloosterman, W. P. *et al.* Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. *Genome Biol.* **12,** R103 (2011).

59. Stephens, P. J. *et al.* Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* **144,** 27–40 (2011).

60. Magrangeas, F., Avet-Loiseau, H., Munshi, N. C. & Minvielle, S. Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients. *Blood* **118,** 675–678 (2011).

61. Teles Alves, I. *et al.* Gene fusions by chromothripsis of chromosome 5q in the VCaP prostate cancer cell line. *Hum. Genet.* **132,** 709–713 (2013).

62. Böttcher, R. *et al.* Cribriform and intraductal prostate cancer are associated with increased genomic instability and distinct genomic alterations. *BMC Cancer* **18,** 8 (2018).

63. Kloosterman, W. P. *et al.* Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum. Mol. Genet.* **20,** 1916–1924 (2011).

64. Collins, R. L. *et al.* Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.* **18,** 36 (2017).

65. de Pagter, M. S. *et al.* Chromothripsis in healthy individuals affects multiple protein-coding genes and can result in severe congenital abnormalities in offspring. *Am. J. Hum. Genet.* **96,** 651–656 (2015).

66. Trowsdale, J. & Knight, J. C. Major histocompatibility complex genomics and human disease. *Annu. Rev. Genomics Hum. Genet.* **14,** 301–323 (2013).

67. McDevitt, H. The discovery of linkage between the MHC and genetic control of the immune response. *Immunol. Rev.* **185,** 78–85 (2002).

68. The MHC sequencing consortium. Complete sequence and gene map of a human major his-

tocompatibility complex. *Nature* **401,** 921–923 (1999).

69. Harding, C. V. & Unanue, E. R. Cellular mechanisms of antigen processing and the function of class I and II major histocompatibility complex molecules. *Mol. Biol. Cell* **1,** 499–509 (1990).

70. Morishima, Y. *et al.* Biological significance of HLA locus matching in unrelated donor bone marrow transplantation. *Blood* **125,** 1189–1197 (2015).

71. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18,** 76 (2017).

72. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).

73. Song, S. *et al.* Major histocompatibility complex class I molecules protect motor neurons from astrocyte-induced toxicity in amyotrophic lateral sclerosis. *Nat. Med.* **22,** 397–403 (2016).

74. Hamza, T. H. *et al.* Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat. Genet.* **42,** 781–785 (2010).

75. Caffrey, M. F. & James, D. C. Human lymphocyte antigen association in ankylosing spondylitis. *Nature* **242,** 121 (1973).

76. Kaneko, K. *et al.* Clinical implication of HLA class I expression in breast cancer. *BMC Cancer* **11,** (2011).

77. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513,** 202–209 (2014).

78. Lohr, J. G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 3879–3884 (2012).

79. Shukla, S. A. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33,** 1152–1158 (2015).

80. Weiss, K. M. & Clark, A. G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18,** 19–24 (2002).

81. Cucca, F. *et al.* The HLA-DPB1--associated component of the IDDM1 and its relationship to the major loci HLA-DQB1, -DQA1, and -DRB1. *Diabetes* **50,** 1200–1205 (2001).

82. Price, P. *et al.* The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. *Immunol. Rev.* **167,** 257–274 (1999).

83. The International HIV Controllers Study. The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. *Science* **330,** 1551–1557 (2010).

84. Morris, D. L. *et al.* Unraveling multiple MHC gene associations with systemic lupus erythematosus: model choice indicates a role for HLA alleles and non-HLA genes in Europeans. *Am. J. Hum. Genet.* **91,** 778–793 (2012).

85. Hu, X. *et al.* Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-

DR molecules drive type 1 diabetes risk. *Nat. Genet.* **47,** 898–905 (2015).

86. Simon, M., Bourel, M., Fauchet, R. & Genetet, B. Association of HLA-A3 and HLA-B14 antigens with idiopathic haemochromatosis. *Gut* **17,** 332–334 (1976).

87. Feder, J. N. *et al.* A novel MHC class I–like gene is mutated in patients with hereditary haemochromatosis. *Nat. Genet.* **13,** 399–408 (1996).

88. Hiby, S. E. *et al.* Maternal activating KIRs protect against human reproductive failure mediated by fetal HLA-C2. *J. Clin. Invest.* **120,** 4102–4110 (2010).

89. Hirayasu, K. *et al.* Significant association of KIR2DL3-HLA-C1 combination with cerebral malaria and implications for co-evolution of KIR and HLA. *PLoS Pathog.* **8,** e1002565 (2012).

90. Kirino, Y. *et al.* Genome-wide association analysis identifies new susceptibility loci for Behçet's disease and epistasis between HLA-B*51 and ERAP1. *Nat. Genet.* **45,** 202–207 (2013).

91. Cortes, A. *et al.* Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with ERAP1. *Nat. Commun.* **6,** 7146 (2015).

92. Castro-Santos, P., Moro-García, M. A., Marcos-Fernández, R., Alonso-Arias, R. & Díaz-Peña, R. ERAP1 and HLA-C interaction in inflammatory bowel disease in the Spanish population. *Innate Immun.* **23,** 476–481 (2017).

93. Horton, R. *et al.* Gene map of the extended human MHC. *Nat. Rev. Genet.* **5,** 889–899 (2004).

94. Horton, R. *et al.* Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project. *Immunogenetics* **60,** 1–18 (2008).

95. Jensen, J. M. *et al.* Assembly and analysis of 100 full MHC haplotypes from the Danish population. *Genome Res.* **27,** 1597–1607 (2017).

96. Stewart, C. A. Complete MHC Haplotype Sequencing for Common Disease Gene Mapping. *Genome Res.* **14,** 1176–1187 (2004).

97. Traherne, J. A. *et al.* Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. *PLoS Genet.* **2,** e9 (2006).

98. Sospedra, M. *et al.* Redundancy in antigen-presenting function of the HLA-DR and -DQ molecules in the multiple sclerosis-associated HLA-DR2 haplotype. *J. Immunol.* **176,** 1951–1961 (2006).

99. Robinson, J., Waller, M. J., Parham, P., Bodmer, J. G. & Marsh, S. G. IMGT/HLA Database--a sequence database for the human major histocompatibility complex. *Nucleic Acids Res.* **29,** 210–213 (2001).

100. Robinson, J. *et al.* Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLoS Genet.* **13,** e1006862 (2017).

101. Mack, S. J. *et al.* Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue Antigens* **81,** 194–203 (2013).

102. Allison, A. C. Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection.

BMJ **1,** 290–294 (1954).

103. Haldane, J. B. S. Disease and Evolution. in *Emerging Infectious Diseases of the 21st Century* 175–187

104. Hill, A. V. *et al.* Common west African HLA antigens are associated with protection from severe malaria. *Nature* **352,** 595–600 (1991).

105. McClelland, E. E., Penn, D. J. & Potts, W. K. Major histocompatibility complex heterozygote superiority during coinfection. *Infect. Immun.* **71,** 2079–2086 (2003).

106. Penn, D. J., Damjanovich, K. & Potts, W. K. MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 11260–11264 (2002).

107. Meyer-Lucht, Y. & Sommer, S. MHC diversity and the association to nematode parasitism in the yellow-necked mouse (Apodemus flavicollis). *Mol. Ecol.* **14,** 2233–2243 (2005).

108. Thursz, M. R., Thomas, H. C., Greenwood, B. M. & Hill, A. V. Heterozygote advantage for HLA class-II type in hepatitis B virus infection. *Nat. Genet.* **17,** 11–12 (1997).

109. Carrington, M. *et al.* HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science* **283,** 1748–1752 (1999).

110. Jones, A. G. & Ratterman, N. L. Mate choice and sexual selection: what have we learned since Darwin? *Proc. Natl. Acad. Sci. U. S. A.* **106 Suppl 1,** 10001–10008 (2009).

111. Penn, D. J. & Potts, W. K. The Evolution of Mating Preferences and Major Histocompatibility Complex Genes. *Am. Nat.* **153,** 145–164 (1999).

112. Potts, W. K., Manning, C. J. & Wakeland, E. K. Mating patterns in seminatural populations of mice influenced by MHC genotype. *Nature* **352,** 619–621 (1991).

113. OlsÉn, K. H., Grahn, M., Lohm, J. & Langefors, Å. MHC and kin discrimination in juvenile Arctic charr, Salvelinus alpinus (L.). *Anim. Behav.* **56,** 319–327 (1998).

114. Reusch, T. B. H., Häberli, M. A., Aeschlimann, P. B. & Milinski, M. Female sticklebacks count alleles in a strategy of sexual selection explaining MHC polymorphism. *Nature* **414,** 300–302 (2001).

115. Wedekind, C., Seebeck, T., Bettens, F. & Paepke, A. J. MHC-dependent mate preferences in humans. *Proc. Biol. Sci.* **260,** 245–249 (1995).

116. Wedekind, C. & Furi, S. Body odour preferences in men and women: do they aim for specific MHC combinations or simply heterozygosity? *Proceedings of the Royal Society B: Biological Sciences* **264,** 1471–1479 (1997).

117. Kromer, J. *et al.* Influence of HLA on human partnership and sexual satisfaction. *Sci. Rep.* **6,** 32550 (2016).

118. Ober, C. *et al.* HLA and mate choice in humans. *Am. J. Hum. Genet.* **61,** 497–504 (1997).

119. Ihara, Y., Aoki, K., Tokunaga, K., Takahashi, K. & Juji, T. HLA and Human Mate Choice. Tests on Japanese Couples. *Anthropol. Sci.* **108,** 199–214 (2000).

120. Chaix, R., Cao, C. & Donnelly, P. Is Mate Choice in Humans MHC-Dependent? *PLoS Genet.* **4,** e1000184 (2008).

121. Derti, A., Cenik, C., Kraft, P. & Roth, F. P. Absence of evidence for MHC-dependent mate selection within HapMap populations. *PLoS Genet.* **6,** e1000925 (2010).

122. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology* **24,** 104–108 (1992).

123. Venter, J. C. GENOMICS: Shotgun Sequencing of the Human Genome. *Science* **280,** 1540–1542 (1998).

124. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).

125. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).

126. Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Res.* **15,** 1767–1776 (2005).

127. Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5,** 335–344 (2004).

128. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–1145 (2008).

129. Ju, J. *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 19635–19640 (2006).

130. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).

131. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–351 (2016).

132. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

133. Brunner, A. L. *et al.* Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res.* **19,** 1044–1056 (2009).

134. Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol.* **12,** R112 (2011).

135. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10,** R32 (2009).

136. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39,** e90 (2011).

137. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).

138. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res.* **6,** 100 (2017).

139. Nakano, K. *et al.* Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum. Cell* **30,** 149–161 (2017).

140. Deamer, D. W. & Branton, D. Characterization of Nucleic Acids by Nanopore Analysis. *Acc. Chem. Res.* **35,** 817–825 (2002).

141. Akeson, M., Branton, D., Kasianowicz, J. J., Brandin, E. & Deamer, D. W. Microsecond Time-Scale Discrimination Among Polycytidylic Acid, Polyadenylic Acid, and Polyuridylic Acid as Homopolymers or as Segments Within Single RNA Molecules. *Biophys. J.* **77,** 3227–3233 (1999).

142. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 13770–13773 (1996).

143. Ashkenasy, N., Sánchez-Quesada, J., Bayley, H. & Ghadiri, M. R. Recognizing a single base in an individual DNA strand: a step toward DNA sequencing in nanopores. *Angew. Chem. Int. Ed Engl.* **44,** 1401–1404 (2005).

144. Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G. & Bayley, H. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 7702–7707 (2009).

145. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.* **31,** 247–250 (2013).

146. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *bioRxiv* 128835 (2017). doi:10.1101/128835

147. Istace, B. *et al.* de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* **6,** 1–13 (2017).

148. Carter, J.-M. & Hussain, S. Robust long-read native DNA sequencing using the ONT CsgG Nanopore system. *Wellcome Open Res* **2,** 23 (2017).

149. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12,** 733–735 (2015).

150. Fournier, T. *et al.* High-quality de novo genome assembly of the Dekkera bruxellensis UMY321 yeast isolate using Nanopore MinION sequencing. (2017). doi:10.1101/151167

151. Jansen, H. J. *et al.* Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *Sci. Rep.* **7,** 7213 (2017).

152. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8,** 16027 (2017).

153. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14,** 407–410 (2017).

154. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530,** 228–232 (2016).

155. Faria, N. R. *et al.* Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* **8,** 97 (2016).

156. David, M., Dursi, L. J., Yao, D., Boutros, P. C. & Simpson, J. T. Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics* **33,** 49–55 (2017).

157. Boža, V., Brejová, B. & Vinař, T. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PLoS One* **12,** e0178751 (2017).

158. Teng, H. *et al.* Chiron: Translating nanopore raw signal directly into nucleotide sequence using deep learning. *bioRxiv* 179531 (2017). doi:10.1101/179531

159. Cretu Stancu, M. *et al.* Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.* **8,** 1326 (2017).

160. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences* **85,** 2444–2448 (1988).

161. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

162. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

163. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single molecule sequencing. *bioRxiv* 169557 (2017). doi:10.1101/169557

164. Wei, Q. *et al.* A Bayesian framework for de novo mutation calling in parents-offspring trios. *Bioinformatics* **31,** 1375–1381 (2015).

165. Ramu, A. *et al.* DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat. Methods* **10,** 985–987 (2013).

166. Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods* **102,** 36–49 (2016).

167. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21,** 974–984 (2011).

168. Xie, C. & Tammi, M. T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10,** 80 (2009).

169. Wang, Z., Hormozdiari, F., Yang, W.-Y., Halperin, E. & Eskin, E. CNVeM: copy number variation detection using uncertainty of read mapping. *J. Comput. Biol.* **20,** 224–236 (2013).

170. Xi, R., Lee, S., Xia, Y., Kim, T.-M. & Park, P. J. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res.* **44,** 6274–6286 (2016).

171. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10,** 691–703 (2009).

172. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv* 193144 (2017). doi:10.1101/193144

173. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15,** R84 (2014).

174. Mohiyuddin, M. *et al.* MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* **31,** 2741–2744 (2015).

175. Schröder, J. *et al.* Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics* **30,** 1064–1072 (2014).

176. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).

177. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6,** 677–681 (2009).

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

179. Noll, A. C. *et al.* Clinical detection of deletion structural variants in whole-genome sequences. *NPJ Genom Med* **1,** 16026 (2016).

180. English, A. C. *et al.* Assessing structural variation in a personal genome—towards a human reference diploid genome. *BMC Genomics* **16,** 286 (2015).

181. Parikh, H. *et al.* svclassify: a method to establish benchmark structural variant calls. *BMC Genomics* **17,** 64 (2016).

182. Redin, C. *et al.* The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet.* **49,** 36–45 (2017).

183. Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27,** 677–685 (2017).

184. Bose, B., W., D. & B., S. Transplantation Antigens and Histocompatibility Matching. in *Current Issues and Future Direction in Kidney Transplantation* (ed. Rath, T.) (InTech, 2013).

CHAPTER 2

A FRAMEWORK FOR THE DETECTION OF DE NOVO MUTATIONS IN FAMILY-BASED SEQUENCING DATA

Laurent C Francioli[1,2,3,*,43], Mircea Cretu-Stancu[1,*], Kiran V Garimella[4], Menachem Fromer[2,3,5,6], Wigard P Kloosterman[1], Genome of the Netherlands Consortium[44], Kaitlin E Samocha[2,3], Benjamin M Neale[2,3], Mark J Daly[2,3], Eric Banks[3], Mark A DePristo[3], Paul IW de Bakker[1,7]

* | These authors contributed equally to this work.
1 | Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands;
2 | Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA;
3 | Program in Medical and Population Genetics, The Broad Institute of Harvard and MIT, Cambridge, MA, USA;
4 | Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, UK;
5 | Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA;
6 | Department of Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA;
7 | Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, CG, The Netherlands;
43 | Correspondence: Dr L Francioli, Department of Medical Genetics, Utrecht University, Heidelberglaan 100, Utrecht, The Netherlands. Tel: (617) 953-0400; Fax: (617) 643-3293;
E-mail: lfran@broadinstitute.org
44 | Genome of the Netherlands Consortium members are listed before the references.

Manuscript adapted from the European Journal of Human Genetics

Abstract

Germline mutation detection from human DNA sequence data is challenging due to the rarity of such events relative to the intrinsic error rates of sequencing technologies and the uneven coverage across the genome. We developed PhaseByTransmission (PBT) to identify de novo single nucleotide variants and short insertions and deletions (indels) from sequence data collected in parent-offspring trios. We compute the joint probability of the data given the genotype likelihoods in the individual family members, the known familial relationships and a prior probability for the mutation rate. Candidate de novo mutations (DNMs) are reported along with their posterior probability, providing a systematic way to prioritize them for validation. Our tool is integrated in the Genome Analysis Toolkit and can be used together with the ReadBackedPhasing module to infer the parental origin of DNMs based on phase-informative reads. Using simulated data, we show that PBT outperforms existing tools, especially in low coverage data and on the X chromosome. We further show that PBT displays high validation rates on empirical parent-offspring sequencing data for whole-exome data from 104 trios and X-chromosome data from 249 parent-offspring families. Finally, we demonstrate an association between father's age at conception and the number of DNMs in female offspring's X chromosome, consistent with previous literature reports.

INTRODUCTION

De novo mutation (DNM) between generations is a key mechanism in evolution. In humans, the mutation rate is estimated between $1x10^{-8}$ and $3x10^{-8}$ per base per generation from direct observations[1–4] and from species comparisons,[5] although mutation rates have been shown to vary locally,[2,6] across families[2–4] and to depend on paternal age.[3] While most DNMs are thought to be selectively neutral, the phenotypic consequences can be severe when functional elements in the genome are mutated,[7] and such cases are therefore of critical

interest for medical genetics.[8] Next generation sequencing (NGS) technologies applied to whole genomes in pedigrees enable systematic discovery and analysis of DNMs. Because the error rates from NGS data are currently much greater than the underlying DNM rate, detecting DNMs from NGS data requires accurate, quantitative calibration of the evidence supporting a novel allele in the offspring and the evidence against Mendelian transmission of this allele from (one of) the parents. A miscalled genotype in the parents or the offspring may lead to a false positive or false negative result. Consequently, variant callers[9,10] emit genotype likelihoods for each possible genotype to incorporate the uncertainty from the raw data. We developed an algorithm called PhaseByTransmission (PBT) to compute the posterior probability for each genotype combination within a trio at each site given the genotype likelihoods in the individual family members, the known familial relationships and (optionally) the allele frequency in the population. PBT considers biallelic single nucleotide variants (SNVs) and short insertions and deletions (indels) within the autosomes and the X chromosome, and generates a list of all candidate DNMs ranked by their posterior probability. A key advantage is the integration of PBT within the widely used Genome Analysis Toolkit (GATK)[9] and its ability to leverage phase information from the GATK ReadBackedPhasing module to identify the parental origin of DNMs.

Materials and Methods

PhaseByTransmission takes individual genotype likelihoods as input, defined as the likelihood L of the bases D observed at a site given each bi-allelic genotype G: L(DIG). These likelihoods can be computed from the sequence data using different genotype calling algorithms, such as the GATK UnifiedGenotyper (UG), GATK HaplotypeCaller or Samtools.[11] For a given parent–parent–offspring trio, we enumerate all possible genotype combinations at a unique site in the genome. For bi-allelic autosomal sites, there are 27 possible genotype combinations within a trio: 15 are consistent with Mendelian inheritance, 10 correspond to a single DNM and 2 correspond to a pair of DNMs (involving a mutation from both parents). For bi-allelic sites on the X chromosome of a female offspring, only 18 genotype combinations exist because the father is haploid: 8 are consistent with Mendelian inheritance, 8 correspond to a single DNM and 2 correspond to a pair of DNMs. Because male offspring are haploid on the X chromosome and inherited their X chromosome from their mothers, there

are only 6 mother-offspring genotype combinations: 4 are consistent with Mendelian inheritance and 2 correspond to a single DNM. Given a mutation rate μ, n genotype combinations consistent with a single DNM (from 1 parent) and *m* genotype combinations consistent with two DNMs (from both parents), we define the following genotype combination prior:

$$P_C = \begin{cases} 1 - n\mu - m\mu^2, \text{ if the combination follows Mendel's laws} \\ \mu, \text{ if the combination implies 1 mutation} \\ \mu^2, \text{ if the combination implies 2 mutations} \end{cases} \quad (1)$$

By using these genotype combination priors, we can compute the posterior probability of observing the sequencing data D given each of these possible underlying genotype combinations:

$$P(D|G_M, G_F, G_C) = P_C \cdot P(D|G_M) \cdot P(D|G_F) \cdot P(D|G_C) \quad (2)$$

where G_M, G_F and G_C are the genotypes of the mother, father and child, and P_C the genotype combination prior.

Following the posterior calculation for each of the *N* possible genotype combinations in the trio, we assign the most likely one to the trio, at each site, and compute its normalized posterior probability. All sites and trios assigned a genotype combination violating Mendel's laws are reported as putative DNMs and the posterior probability assigned to each of them reflects the confidence of the call. In addition to the familial relationships among samples, population allele frequencies can be incorporated as a prior into our model. Because one of the most common sources of false positive DNM calls is lack of sequence coverage in (one of) the parents, informing the model about allele frequencies in the population can help to reduce false positive rates. When adding allele frequency priors, Equation(2) becomes:

$$P(D|G_M, G_F, G_C) = P_C \cdot P_{AF}^{G_M} \cdot P(D|G_M) \cdot P_{AF}^{G_F} \cdot P(D|G_F) \cdot P(D|G_C) \quad (3)$$

where G_M, G_F and G_C are the genotypes of the mother, father and child, P^{GM}_{AF} and P^{GF}_{AF} the allele frequency priors for the mother's and father's genotypes, and P_C the genotype combination prior.

The allele frequencies for the sites can be provided either as a separate VCF file or computed from the genotypes of the samples in the input VCF file when multiple samples from

a single population are studied. In this case, the allele frequencies are estimated as P^{G}_{AF} for each genotype G following Hardy-Weinberg equilibrium expectation:

$$P^{G}_{AF} = \begin{cases} p^2, \text{ if the genotype } G \text{ is homozygous reference} \\ 2pq, \text{ if the genotype } G \text{ is heterozygous} \\ q^2, \text{ if the genotype } G \text{ is homozygous alternative} \end{cases} \quad (4)$$

where p and q are the estimated allele dosage for the reference and alternate alleles, respectively, in the parents (founders). In addition to calling DNMs, PBT also phases the inherited variants based on the segregation of alleles within a trio. By considering all possible genotype combinations and following Mendelian inheritance, we can infer phase deterministically for all trio individuals in all but two situations: when all trio individuals are heterozygous for the same two alleles, or when there is a DNM in the offspring. Except for these two cases, the phasing quality is bounded by the joint probability of the trio genotype combination.

Results

Simulated data

In order to evaluate the performance of PBT we simulated sequencing data for 10 parent-offspring trios, 5 with a male offspring and 5 with a female offspring (Figure 1). We randomly selected 10 families from the Genome of the Netherlands (GoNL) Project[4] and used previously reconstructed haplotypes for the parents for our simulations. We created haplotypes for the children by randomly selecting one haplotype from each of the parents and introduced on average 11,435 DNMs across the autosomes and 1,821 on the X chromosome per offspring (all single base changes). In order to obtain a realistic genome-wide distribution of DNMs, we applied substitution-specific local mutation probabilities, which we empirically derived from the GoNL mutation rate map.[12] This mutation map covers 75% of the human genome and provides mutation rate estimates at the megabase scale for all substitution types, as well as for C4T transitions in a CpG context. To simulate the paternal bias observed in previous studies,[2–4] we randomly assigned 70% of the DNMs to the paternal haplotype and 30% of them to the maternal haplotype. Mutations across the X chromosome were distributed uniformly, as no mutation map was available. We used SimSeq[13] to simulate 100 bp Illumina paired-end reads with an insert size of 250 bp for all 30 samples, within 10 kb regions centred on each simulated DNM (5 kb upstream and 5 kb downstream). We used the SimSeq default Illumina error profile in our simulation, which inserts errors (and their corresponding phred quality scores) in the simulated reads, as a function of the position within the read and the underlying reference base. The reads were aligned to the UCSC human reference sequence build 37 using BWA[14] to produce aligned BAM files. To evaluate the effect of depth of coverage on DNM detection, we downsampled the generated BAM files for each sample during the

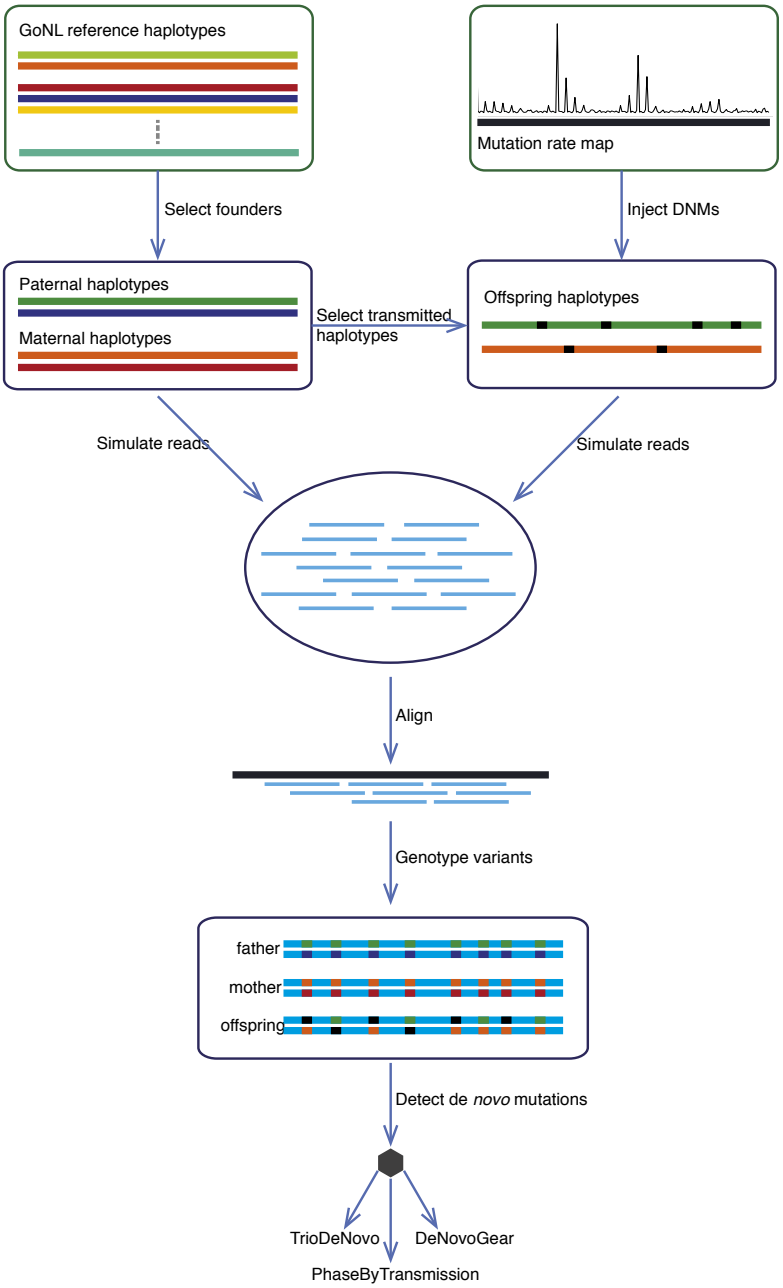


Figure 1: Outline of the pipeline used to generate our simulation data.
The 'mutation rate map' is the autosome-wide GoNL derived mutation rate map, as published before.[12]

variant calling step, to obtain variant call sets for average depths of coverage of 60x, 30x and 15x, respectively. The GATK UG was used on each trio separately to produce the individual genotype likelihoods used as input for PBT.

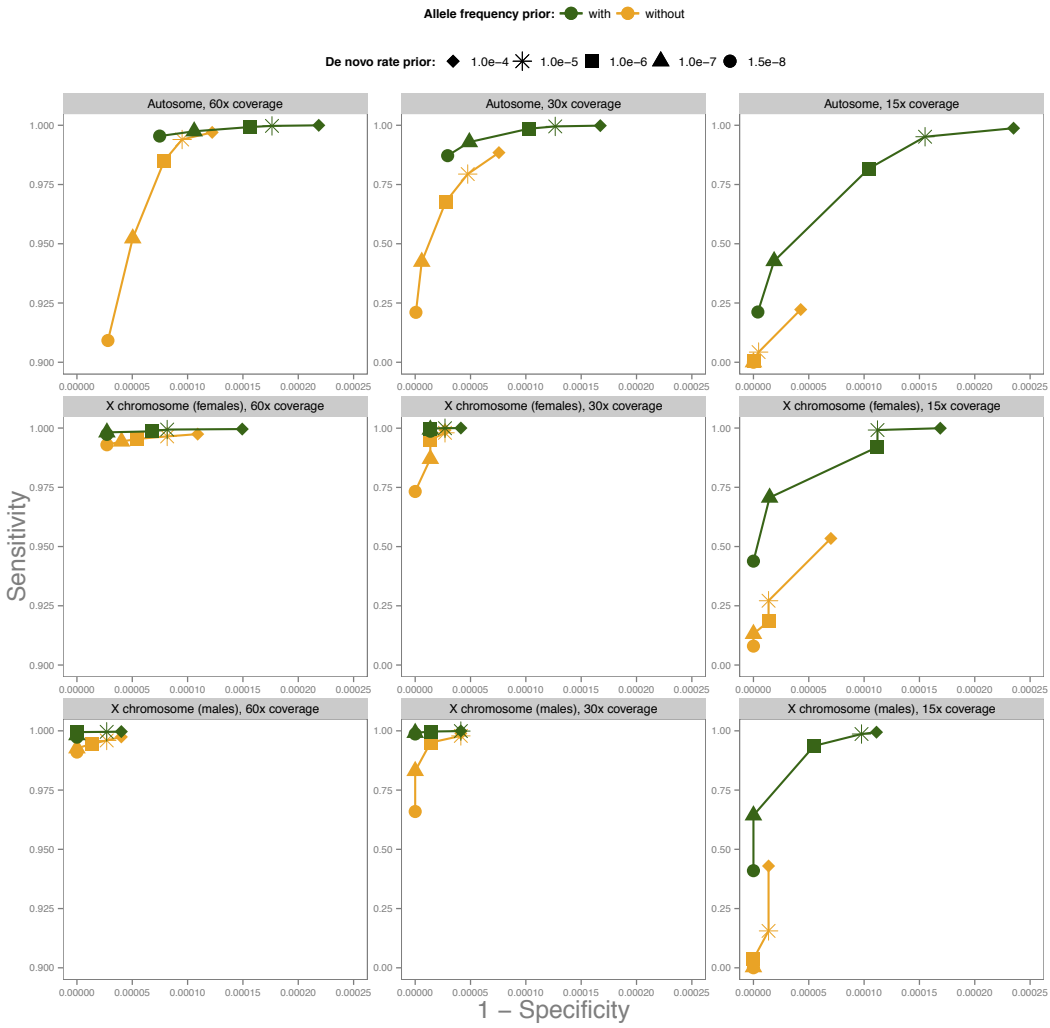


Figure 2: ROC plot showing the performance of PBT, where the mutation rate prior is used as the hidden parameter.
Two scenarios are considered in order to evaluate the relevance of using allele frequency priors (yellow curve: without AF priors, green curve: with AF prior). The analysis is stratified by coverage (columns) and genomic region (rows). The y-scale for the 60x coverage plots is restricted for visibility. Each dot shape corresponds to a specific DNM prior. The allele frequency priors are computed based on 1000 Genomes Phase 3 CEU data.

Using the UG default settings, an average 96% of the simulated DNMs were called as putative variant sites (the remaining 4% were not detected). The VCF file for each trio comprised, on average, 175,458 inherited SNVs and 11,427 Mendelian violations per trio. We ran PBT on the input VCF files using a mutation prior of 1.5×10^{-8} based on estimated per-base human mutation rate estimate.[1–3] We also explored more permissive mutation priors (10^{-7}, 10^{-6}, 10^{-5} and 10^{-4}) and assessed sensitivity and specificity of the downstream results as a function of the depth of coverage and mutation prior. In addition, we ran PBT with and without allele frequency priors based on 1,000 Genomes Phase 3 CEU data.[15] We ran PBT on each set of parameters, and computed the following: the number of simulated DNMs reported as DNMs by PBT (true positives); the number of inherited variants and sequencing errors reported as DNMs by PBT (false positives); the number of inherited variants not reported as DNMs by PBT (true negatives); the number of simulated DNMs not reported as DNMs by PBT (false negatives). From these, we computed the sensitivity as:

$$\frac{\#inherited\ SNVs\ called\ as\ inherited}{\#inherited\ SNVs\ in\ input\ file} \qquad (5)$$

and the specificity as:

$$\frac{\#inherited\ SNVs\ called\ as\ inherited}{\#inherited\ SNVs\ in\ input\ file} \qquad (6)$$

Figure 2 shows the influence of the mutation rate prior and the allele frequency prior on the receiving operator characteristic (ROC) curves for both autosomes and the X chromosome at different depths of coverage. The mutation prior affects the sensitivity and specificity of the resulting DNM calls. As expected, a higher mutation prior increases the sensitivity at the cost of more false positive calls. As a result, the mutation prior value needs to be set depending on the desired output and the sequencing coverage (Figure 2). We note that as coverage increases the optimal value for real data should converge towards the actual human mutation rate (as can be seen for the 60x coverage data). Incorporating allele frequency priors into DNM detection greatly improved the sensitivity at low and medium coverage for both autosomes and the X chromosome. This reflects the higher uncertainty of the parents' genotypes at lower coverage, resulting in poor discrimination between homozygous and heterozygous genotypes. Incorporating the allele frequencies in the model thus leads to a better discrimination between (a) a site that is variant in the population and thus likely to be inherited from one of the parents even though there is little (or no) evidence for the variant allele in (one of) the parents, and (b) a site that is not variant in the population and is likely to be de novo if there is no evidence for the variant allele in either parents.

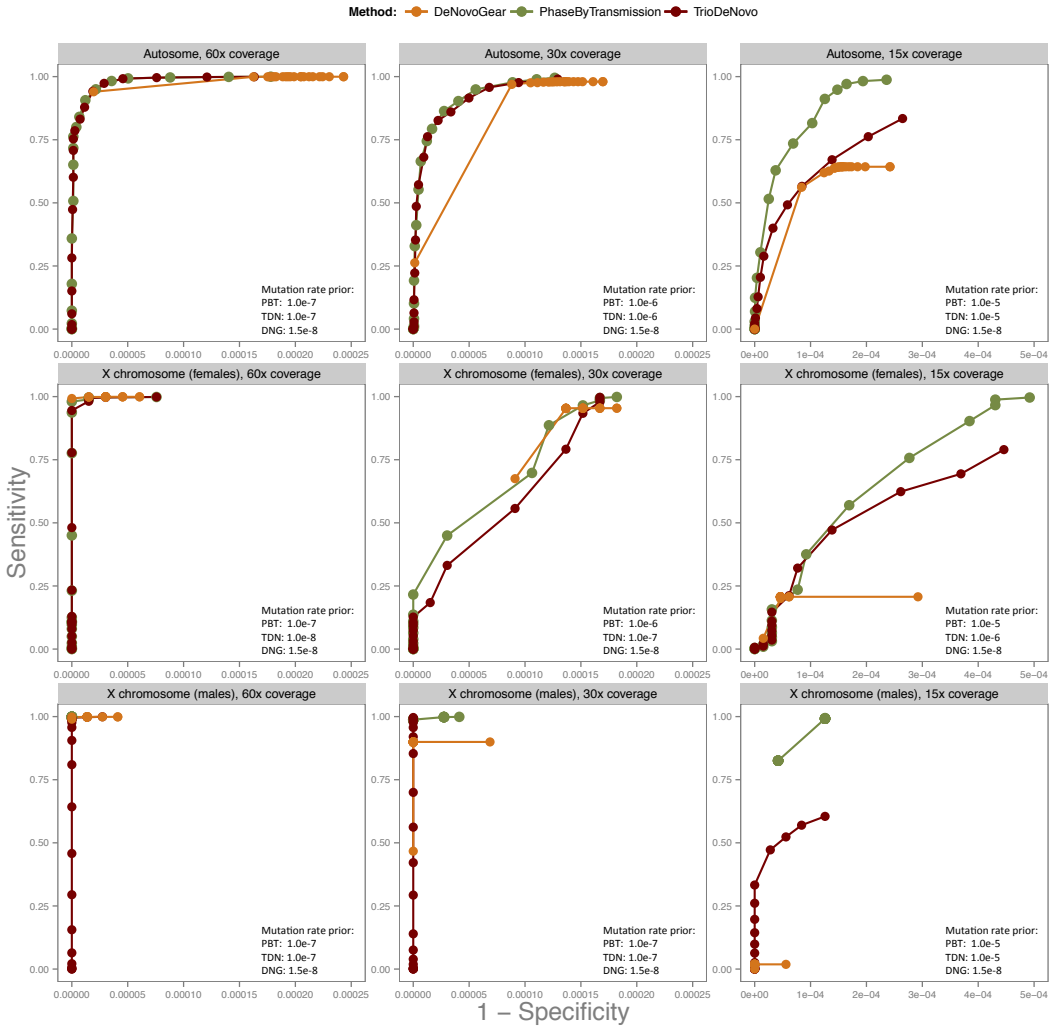


Figure 3: ROC plot illustrating the performance of three DNM calling methods (PhaseByTransmission, TrioDeNovo and DeNovoGear), with respect to each method's DNM output confidence score.
The analysis is stratified by coverage (columns) and genomic region (rows). The posterior cutoffs used for plotting each curve were uniformly distributed across the range of each tool's output DNM confidence scores. Some outlier values where the specificity decreased considerably without any sensitivity gain were removed from the plot and the x-scale for the 60x and 30x coverages is restricted, for visibility purpose. Supplementary Figure SF3 shows the curves with all points. The mutation rate prior values for each scenario, for each tool are selected based on Supplementary Figure SF1.

To compare the performance of PBT against other state-of-the-art DNM callers, we used the same input VCFs to detect DNMs with TrioDeNovo[16] and DeNovoGear.[17] We selected these DNM callers, for their good reported performance as well as similar integration points within analysis pipelines (ie, after individual variant calling is performed). We used the best performing DNM rate prior (out of five predefined priors: $1.5x10^{-8}$, 10^{-7}, 10^{-6}, 10^{-5} and 10^{-4}) to obtain DNM call sets, for each method and coverage. For PBT, we used the allele frequency prior as well. The optimal (in terms of sensitivity versus specificity) mutation rate prior's values were derived from Figure 2 for PBT and from a similar analysis (ie, influence of the mutation rate prior on specificity and sensitivity), on the same simulated dataset, for TrioDeNovo and DeNovoGear (Supplementary Figures SF1 and SF2). The mutation rate parameter value for each method is consistent with documentation or recommendations for each of the tools, where available. Figure 3 shows the ROC curves for the autosomes and the X chromosome at different depths of coverage using the posterior probability reported by each tool as parameter.

All three tools surveyed in this analysis performed very well in terms of the sensitivity at high coverage, while PB T and TrioDeNovo exhibit slightly better specificity. At lower coverage, the differences in sensitivity and specificity become more pronounced. The performance gain achieved by PBT at lower coverage comes from the incorporation of the allele frequencies in the model, which allows for a better discrimination between poorly covered variant sites in the parents and true DNMs. PBT showed good performance in detecting DNMs on the X chromosome even at lower coverage (15x), which was particularly challenging for the other two methods, especially in the male offspring trios. PBT had a sensitivity of 99% in female offspring trios and 98% in male offspring trios. In contrast, TrioDeNovo detects only 77% and 58% of female and male offspring DNMs on the X chromosome respectively, and DeNovoGear sensitivity drops down to 24% for the female offspring DNMs and 3% for the male offspring DNMs, respectively. The better performance of PBT on the X chromosome comes from explicitly modelling the unique mode of inheritance for this chromosome, whereas other tools do not differentiate between autosomes and the X chromosome.

We further evaluated our ability to assign parental origin to the DNMs identified. Assuming sequence reads are of sufficient length, heterozygous variants located close to the DNM can be informative about its parental origin and phase. To this end, we combined trio-based phasing information from PBT and read-based phasing information from ReadBackedPhasing in order to reconstruct the two haplotypes transmitted to the offspring. We only assigned parental origin to sites where all read data spanning adjacent offspring heterozygous positions unambiguously supported the same parental haplotype. We were able to determine parental origin for 14.1% of the simulated DNMs and 81.4% of these were assigned correctly. We note that other tools do not provide automated annotation of the parental origin.

Empirical whole-genome data

In previous work, we have demonstrated the performance of PBT to detect de novo SNVs and indels in 13x coverage autosomal sequencing data of 250 parent-offspring families and

on three parent-offspring families with both whole-exome and whole-genome data from the CLARITY challenge.[18]

Here, we present the application of PBT on the X chromosome sequencing data of 249 parent-offspring families from the GoNL project (230 trios, 11 parent-offspring families with a pair of monozygotic twins and eight parent-offspring families with a pair of dizygotic twins). We used only one randomly chosen offspring from each family with monozygotic twins and used both offspring from families with dizygotic twins. This resulted in a total of 257 offspring (111 males, 146 females) for DNM calling. All GoNL samples were selected without phenotypic ascertainment so as to be representative of the general Dutch population. The DNA samples were extracted from whole blood, and sequenced on Illumina HiSeq2000 using 90 bp paired-end reads with an insert size of 500 bp. The reads were aligned to the UCSC human reference sequence build 37 using BWA and processed using GATK best practices (https://www.broadinstitute.org/gatk/guide/bestpractices). SNVs were called using GATK UG and subsequently filtered using GATK VariantQualityScoreRecalibration (VQSR). We excluded the pseudo-autosomal regions from this analysis since the homology between the X and Y chromosomes in these regions causes ambiguous read mapping and unreliable subsequent genotype calls with current analysis pipelines. The resulting set comprised 701,910 SNVs on the X chromosome and a total of 872,214 Mendelian violations.

We applied PBT to these data using a mutation prior of 10^{-5}, which should provide optimal sensitivity based on our simulations (Figure 2). We also used an allele frequency prior based on the observed allele frequency in all unrelated samples in our study. We applied a posterior cutoff of Q5 for female offspring and kept all DNM calls in male offspring regardless of their posterior, since male offspring calls had lower posteriors in general, due to their overall lower genotype quality. Using these permissive parameters and thresholds, PBT reported a total of 10,380 DNMs. Due to the low depth of sequencing in our data, many of the lower quality calls are likely to be false positives and we thus filtered this set by removing any DNM candidates with any read evidence for the non-reference allele in either of the parents (which in our sequencing context most likely indicates insufficient sequencing of the alternative allele). This resulted in a final set of putative DNMs of 126 male offspring DNMs and 547 female offspring DNMs.

We selected six putative DNMs in male offspring and 54 in female offspring for validation. These candidates were selected randomly from the 66 families where DNA was available for validation using MiSeq deep sequencing (~1,200x coverage). The six male offspring DNMs originate from six different families, whereas the 54 female offspring DNMs originate from 15 families with a median of 3 DNMs per child and a maximum of 7. From the six candidates in male offspring, four could be successfully assayed and all were validated as a true DNM in the offspring. From 54 candidates in female offspring, 43 could be successfully assayed of which 42 (97.7%) were validated as a true DNM. For 10 of the 13 unsuccessfully assayed DNMs, the capture and/or amplification of the locus surrounding the DNM failed for at least

one of the individuals in the trio. In the remaining three cases, the coverage produced by the sequencing run was low in all trio individuals (2–20x). In these three cases, the low coverage data was compatible with a DNM (alternate allele present in child only), but we did not consider the evidence to be sufficient to unambiguously validate the mutation as *de novo*.

We found that male offspring carried on average 1.14 DNMs on the X chromosome, while female offspring carried 1.85 per copy of the X chromosome. Given that male offspring always inherit their X chromosome from their mothers, the much lower average number of DNMs found on the X chromosome of male offspring (1.14), when compared to female offspring (1.85 per copy), is compatible with the paternal germline being highly enriched for DNMs.[1] Despite the limited number of observations in the study, we found a statistically significant

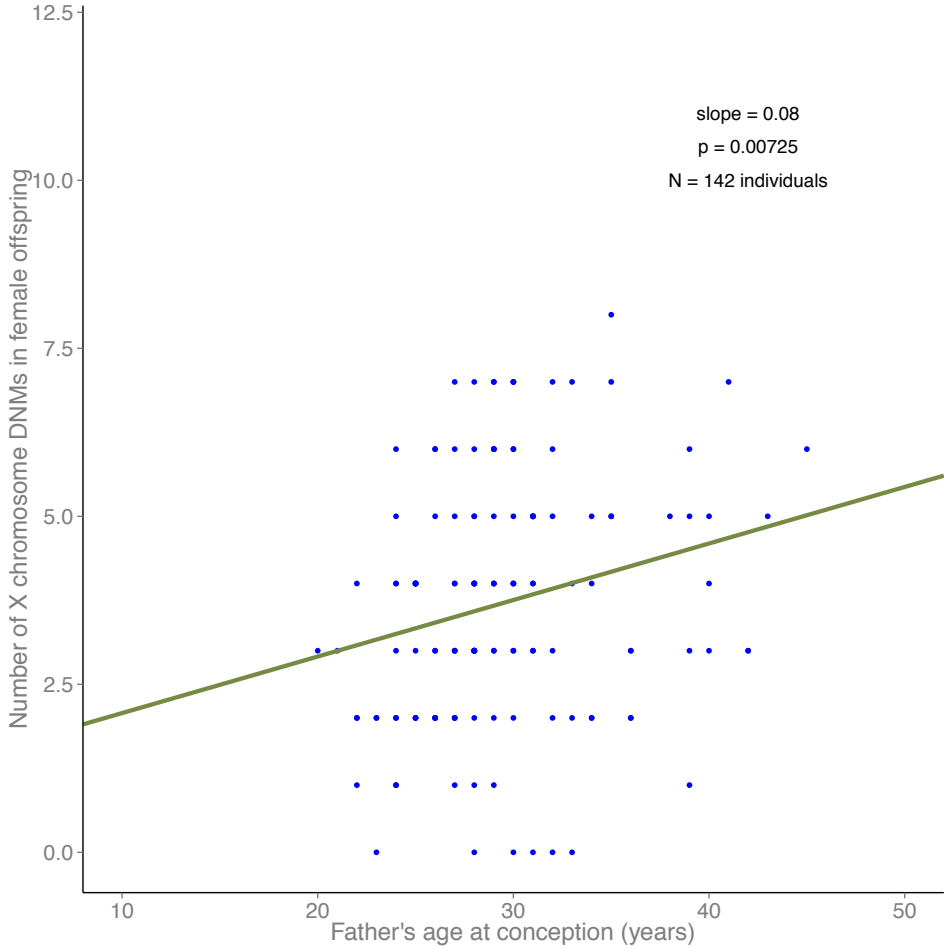


Figure 4: Fitted linear regression line (dark green) of the number of X chromosome DNMs, as a function of father's age at conception.
The data points (blue) represent the set of 547 high confidence DNMs in female offspring. The coefficient estimate is an increase of 0.08 DNMs per year (on the X chromosome).

increase of DNMs on chromosome X with paternal age in female offspring by fitting a linear regression model (P=0.00725), consistent with previous reports[2–4] (Figure 4). As expected, this effect was absent in the male offspring (P=0.24). The linear estimate of 0.08 additional DNMs per year of paternal age on the X chromosome in female offspring data is consistent with previously obtained estimates based on autosomal DNMs (accounting for chromosome sequence length).

Empirical whole-exome data

We evaluated our software on whole exome data in a cohort of 104 trios (single proband and parents). DNA was extracted from wholeblood and exons captured using the Agilent 38Mb SureSelect v2 and sequenced at 60x average depth on the Illumina HiSeq2000 platform for an independent autism study.[19] The sequence data were aligned to the human reference hg19 using BWA,[14] duplicate reads removed, realignment performed around insertions/deletions, and base quality scores recalibrated. Variant discovery and genotyping was performed using the GATK UG across all samples jointly, and calls were subsequently filtered using GATK VQSR.[10]

We ran PBT with a mutation prior of 10^{-7}, on the basis of our simulations (Figure 2), and an allele frequency prior based on the observed data (208 parents). In total, we called 148 putative DNMs, all of which were subjected to experimental validation using Sequenom, and 115 (77.8%) could be assayed successfully. From these, 107 (93%) candidates were validated as true DNMs in the offspring. Looking at false positive calls, five (4.7%) were monomorphic in all samples and three (2.8%) were inherited variants.

Discussion

PhaseByTransmission is an efficient and automated DNM caller using a Bayesian model to estimate the probability of de novo SNVs and/or indel at each site in one or more trios. The model should in principle work with structural variants if genotype likelihoods can be provided. Because PBT works with VCF files as input, it can be integrated into existing NGS analysis pipelines and its results can be annotated using most impact-prediction tools. The PBT algorithm scales linearly with the number of sites and trios. Results on real sequencing data show excellent specificity and sensitivity at both lower and higher coverage in whole-exome and whole-genome data sets. Because PBT explicitly models the inheritance pattern for the X chromosome, it can also be used to derive accurate calls on the X chromosome of both male and female offspring. In addition, due to its integration with the GATK ReadBackedPhasing module, it can provide parent-of-origin information. Finally, PBT can also be used to infer the haplotype phase for most inherited variants in a trio based on the allele segregation within the trio.

Availability of data and materials

PhaseByTransmission and ReadBackedPhasing are available as part of the GATK as a pre-compiled Java package as well as source code at http://www.broadinstitute.org/gatk/download. The GoNL data can be accessed at http://www.nlgenome.nl.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work was funded as part of the Genome of the Netherlands (GoNL) project by the Biobanking and Biomolecular Research Infrastructure (BBMRINL), which is financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007).

Genome of the Netherlands Consortium

Steering committee: Cisca Wijmenga[8,9] (Principal Investigator), Morris A Swertz[8,9],Cornelia M van Duijn[10], Dorret I Boomsma[11], P Eline Slagboom[12], Gertjan B van Ommen[13], Paul IW de Bakker[14,15,16,17]; Analysis group: Morris A Swertz[8,9] (Co-Chair), Laurent C Francioli[14], Freerk van Dijk[8,9], Androniki Menelaou[14], Pieter BT Neerincx[8,9], Sara L Pulit[14], Patrick Deelen[8,9], Clara C Elbers[14], Pier Francesco Palamara[18], Itsik Pe'er[18,19], Abdel Abdellaoui[11], Wigard P Kloosterman[14], Mannis van Oven[20], Martijn Vermaat[21], Mingkun Li[22], Jeroen FJ Laros[21], Mark Stoneking[22], Peter de Knijff[23], Manfred Kayser[21], Jan H Veldink[24], Leonard H van den Berg[24], Heorhiy Byelas[8,9], Johan T den Dunnen[21], Martijn Dijkstra[8,9], Najaf Amin[10], K Joeri van der Velde[8,9], Jouke Jan Hottenga[11], Jessica van Setten[14], Elisabeth M van Leeuwen[10], Alexandros Kanterakis[8,9], Mathijs Kattenberg[11], Lennart C Karssen[10], Barbera DC van Schaik[25], Jan Bot[26], Isaäc J Nijman[14], Ivo Renkens[14], David van Enckevort[27], Hailiang Mei[27], Vyacheslav Koval[28], Karol Estrada[28], Carolina Medina-Gomez[28], Kai Ye[29,12], Eric- Wubbo Lameijer[12], Matthijs HMoed[12], Jayne Y Hehir-Kwa[30], Robert EHandsaker[17,31], Steven A McCarroll[17,31], Shamil R Sunyaev[16,17], Paz Polak[16], Dana Vuzman[16], Mashaal Sohail[16], Fereydoun Hormozdiari[32], Tobias Marschall[33], Alexander Schönhuth[33], Victor Guryev[34], Paul IW de Bakker[14,15,16,17] (Co-Chair); Cohort collection and sample management group: P Eline Slagboom[12], Marian B Beekman[12], Anton JM de Craen[12], H Eka D Suchiman[12], Albert Hofman[10], Cornelia M van Duijn[10], Ben Oostra[35], Aaron Isaacs[10], Najaf Amin[10], Fernando Rivadeneira[28], André G Uitterlinden[28], Dorret I Boomsma[16], Gonneke Willemsen[16], LifeLines Cohort Study[36], Mathieu Platteel[8], Steven J Pitts[37], Shobha Potluri[37], Purnima Sundar[37], David R Cox[37], Whole-genome sequencing: Qibin Li[38], Yingrui Li[38], Yuanping Du[38], Ruoyan Chen[38], Hongzhi Cao[38], Ning Li[39], Sujie Cao[39], JunWang[38,40,41], Ethical, Legal, and Social Issues Jasper A Bovenberg[42], Margreet Brandsma[13]

[8]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; [9]Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; [10]Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands; [11]Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands; [12]Section of Molecular Epidemiology, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands; [13]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; [14]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands; [15]Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands; [16]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [17]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA; [18]Department of Computer Science, Columbia University, New York, NY, USA; [19]Department of Systems Biology, Columbia University, New York, NY, USA; [20]Department of Forensic Molecular Biology, Erasmus Medical Center, Rotterdam, The Netherlands; [21]Leiden Genome Technology Center, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; [22]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; [23]Laboratory for DNA Research, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; [24]Department of Neurology, University Medical Center Utrecht, Utrecht, The Netherlands; [25]Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam Medical Center, Amsterdam, The Netherlands; [26]SURFsara, Science Park, Amsterdam, The Netherlands; [27]Netherlands Bioinformatics Centre, Nijmegen, The Netherlands; [28]Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands; [29]The Genome Institute, Washington University, St Louis, MO, USA; [30]Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands; [31]Department of Genetics, Harvard Medical School, Boston, MA, USA; [32]Department of Genome Sciences, University of Washington, Seattle, WA, USA; [33]Centrum voor Wiskunde en Informatica, Life Sciences Group, Amsterdam, The Netherlands; [34]European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; [35]Department of Clinical Genetics, Erasmus Medical Center, Rotterdam, The Netherlands; [36]A full list of the LifeLines Cohort Study members can be found in the Supplemental Note; [37]Rinat-Pfizer Inc, South San Francisco, CA, USA; [38]BGI-Shenzhen, Shenzhen, China; [39]BGI-Europe, Copenhagen, Denmark;[40]Department of Biology, University of Copenhagen, Copenhagen, Denmark; [41]The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark; [42]Legal Pathways Institute for Health and Bio Law, Aerdenhout, The Netherlands

References

1. Conrad DF, Keebler JEM, DePristo MA et al: Variation in genome-wide mutation rates within and between human families. Nat Genet 2011; 43: 712–714.

2. Michaelson JJ, Shi Y, Gujral M et al: Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 2012; 151: 1431–1442.

3. Kong A, Frigge ML, Masson G et al: Rate of de novo mutations and the importance of father's age to disease risk. Nature 2012; 488: 471–475.

4. Genome of the Netherlands Consortium: Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 2014; 46: 818–825.

5. Nachman MW, Crowell SL: Estimate of the mutation rate per nucleotide in humans. Genetics 2000; 156: 297–304.

6. Hodgkinson A, Eyre-Walker A: Variation in the mutation rate across mammalian genomes. Nat Rev Genet 2011; 12: 756–766.

7. Veltman JA, Brunner HG: De novo mutations in human genetic disease. Nat Rev Genet 2012; 13: 565–575.

8. Gamsiz ED, Sciarra LN, Maguire AM, Pescosolido MF, van Dyck LI, Morrow EM: Discovery of rare mutations in autism: elucidating neurodevelopmental mechanisms. Neurother J Am Soc Exp Neurother 2015; 12: 553–571.

9. McKenna A, Hanna M, Banks E et al: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20:1297–1303.

10. DePristo MA, Banks E, Poplin R et al: A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011; 43:491–498.

11. Li H: A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011; 27: 2987–2993.

12. Francioli LC, Polak PP, Koren A et al: Genome-wide patterns and properties of de novo mutations in humans. Nat Genet 2015; 47: 822–826.

13. Earl D, Bradnam KSt, John J et al: Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res 2011; 21: 2224–2241.

14. Li H, Durbin R: Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010; 26: 589–595.

15. The 1000 Genomes Consortium: An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491: 56–65.

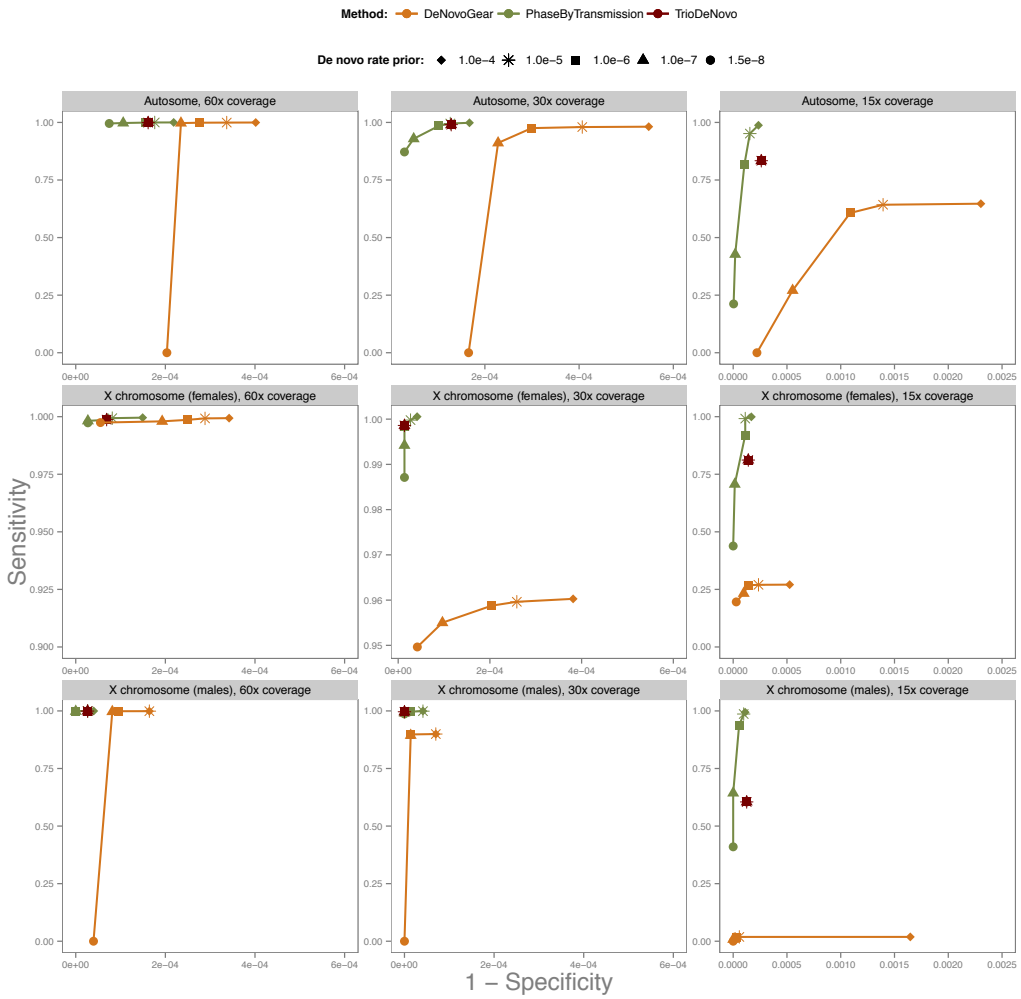
16. Wei Q, Zhan X, Zhong X et al: A Bayesian framework for de novo mutation calling in parents-offspring trios. Bioinformatics 2015; 31: 1375–1381.

17. Ramu A, Noordam MJ, Schwartz RS et al: DeNovoGear: de novo indel and point mutation discovery and phasing. Nat Methods 2013; 10: 985–987.

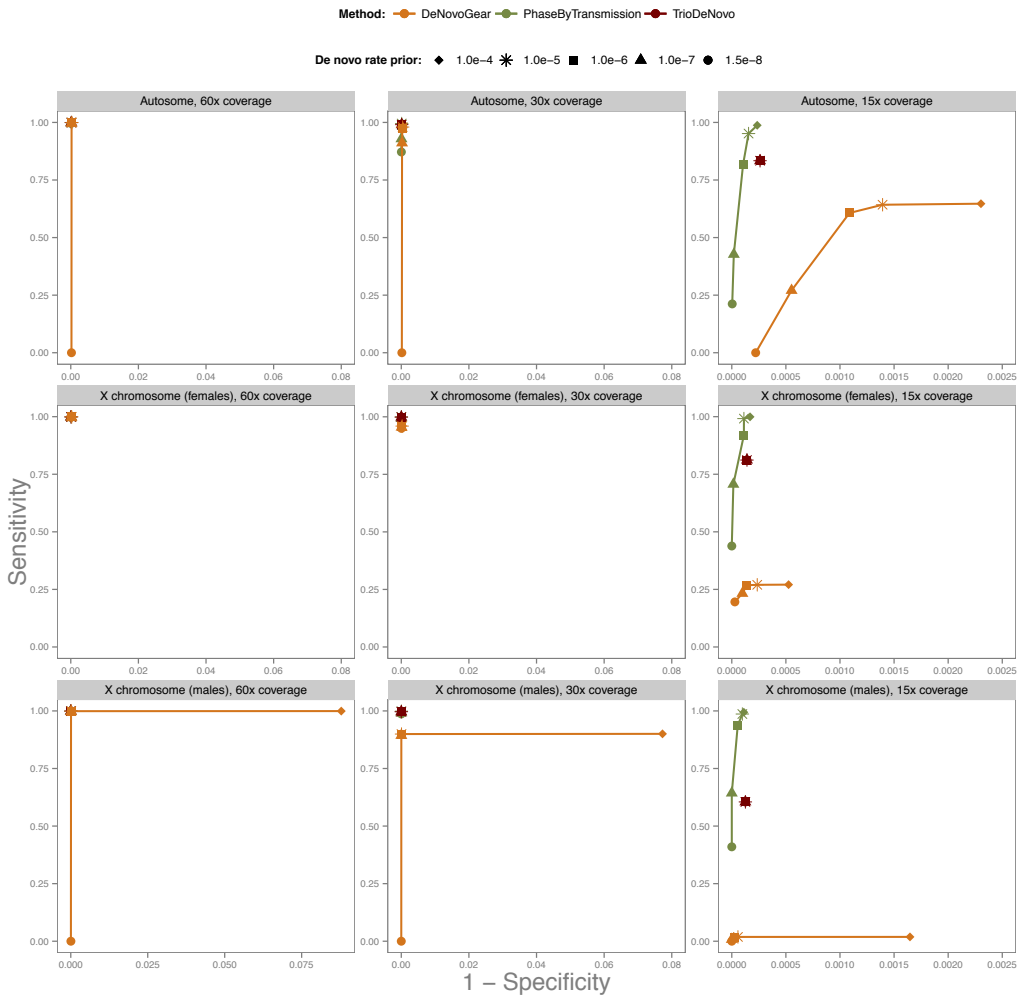
18. Brownstein CA, Beggs AH, Homer N et al: An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. Genome Biol 2014; 15: R53.

19. Neale BM, Kou Y, Liu L et al: Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 2012; 485: 242–245.

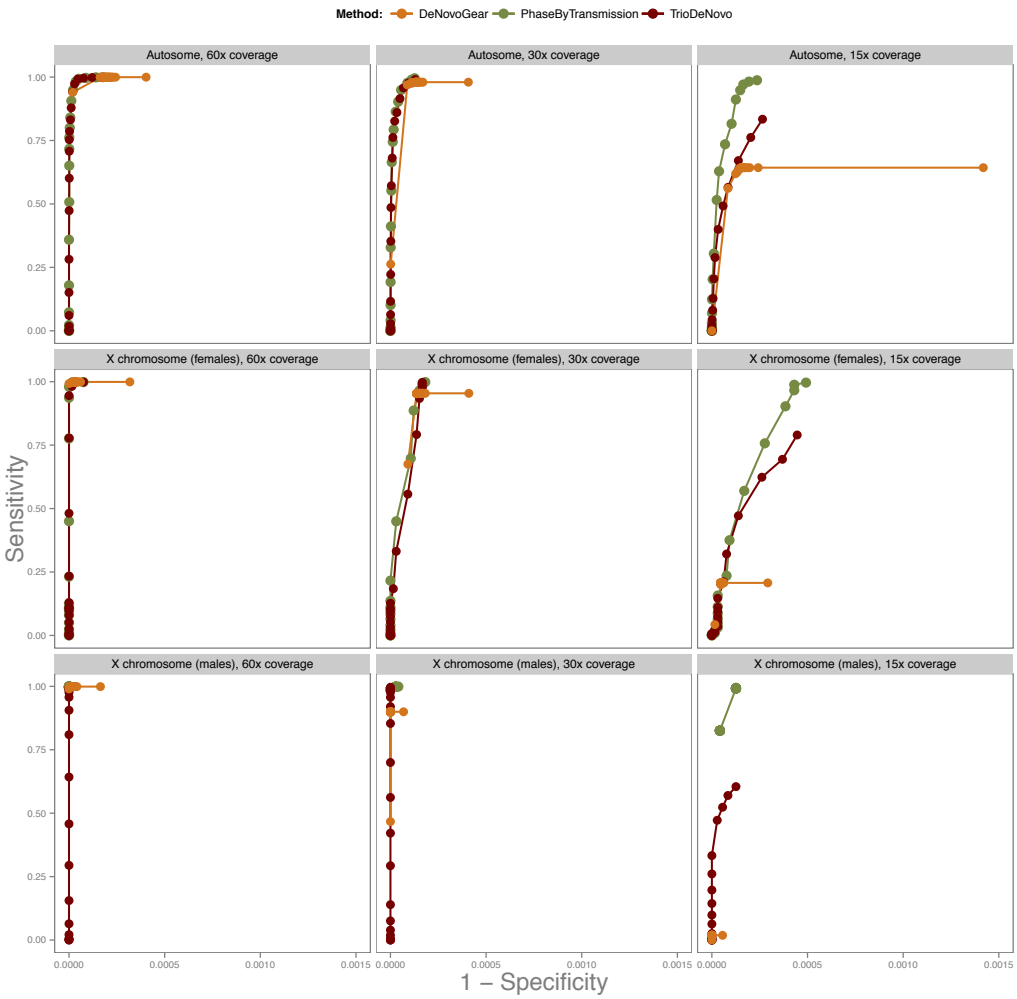
SUPPLEMENTARY INFORMATION TO CHAPTER 2



Supp. Fig. SF1: Receiver Operating Curve (ROC) plot, where the mutation rate prior is used as the hidden parameter.
Homologous to Figure2 in the main text, but evaluating all methods (i.e.: including TrioDeNovo and DeNovoGear). The curve for PBT is using the allele frequency prior and is plotted for comparison. Some outlier values where the specificity decreased considerably without any sensitivity gain were removed from the plot for visibility purposes. The mutation rate prior values for the tool comparison (see main text) were selected based this figure.



Supp Fig. SF2: Homologous to Supp. Fig. SF1, but including all data points



Supp. Fig. SF3: Receiver Operating Curve (ROC) plot illustrating the performance of the three methods being compared (PhaseByTransmission, TrioDeNovo and DeNovoGear), w.r.t. to each method's posterior DNM score. Homologous to Figure3, but showing all data points.

CHAPTER 3

NO EVIDENCE THAT MATE CHOICE IN HUMANS IS DEPENDENT ON THE MHC

Mircea Cretu-Stancu[1], Wigard P. Kloosterman[1,4], Sara L. Pulit[1,2,3,4]

1 | Department of Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands
2 | Li Ka Shing Center for Health Information and Discovery, Big Data Institute, Oxford University, Oxford, United Kingdom
3 | Program in Medical and Population Genetics, Broad Institute, Boston, MA, USA
4 | Corresponding authors. corresponding e-mails: W.Kloosterman@umcutrecht.nl and s.l.pulit@umcutrecht.nl

Manuscript submitted to PLoS Genetics

Abstract

A long-standing hypothesis in biology proposes that various species select mates with a major histocompatibility complex (MHC) composition divergent from their own, so as to improve immune response in offspring. However, human and animal studies investigating this mate selection hypothesis have returned inconsistent results. Here, we analyze 239 mate-pairs of Dutch ancestry, all with whole-genome sequence data collected by the Genome of the Netherlands project, to investigate whether mate selection in humans is MHC dependent. We find no evidence for MHC-mediated mate selection in this sample (with an average MHC genetic similarity in mate pairs (Qc) = 0.829; permutation-based p = 0.703). Limiting the analysis to only common variation or considering the extended MHC region does not change our findings (Qc = 0.671, p = 0.513; and Qc = 0.844, p = 0.696, respectively). We demonstrate that the MHC in mate-pairs is no more genetically dissimilar (on average) than a pair of two randomly selected individuals, and conclude that there is no evidence to suggest that mate choice is influenced by genetic variation in the MHC.

Introduction

The extended major histocompatibility complex (MHC) spans an approximately 7-megabase region on chromosome 6 in humans. The region codes for a series of proteins critical to acquired immune function as well as olfactory genes [1]. Additionally, the MHC contains extensive genetic diversity [2,3], much more so than other regions of the genome; within the human population, the MHC contains thousands of different alleles and haplotypic combinations spanning the frequency spectrum. A plethora of genetic variants in the region have been identified by genome-wide association studies (GWAS) as being associated to a host of diseases [4], both with and without previously-described roles for immune function [5–10].

Some biological studies have proposed that, beyond the direct role in immune function, the MHC may influence mate selection in vertebrate species. Increased MHC diversity is evolutionarily advantageous, as it improves immune response to a wider range of pathogens [11,12]. A number of studies in (non-human) animals indicate that some species of mice, birds, and fish, preferentially mate to maintain or increase MHC diversity [13–17]. For example, studies in sticklebacks [18] indicate that MHC-based mate selection helps to optimize copy number of particular MHC loci between mates. In mice, increased MHC dissimilarity between mates increases diversity of amino acid substitutions within binding-pockets of specific HLA molecules [19,20]. Many of these studies suggest that the observed MHC-dependent mate selection is mediated by the olfactory system, either through detectable residues that mates can smell [21], or because olfactory-receptor genes are often found to cluster in close genomic proximity to the MHC [3].

Evidence for MHC-dependent mate selection in humans is far less conclusive. A study of 411 couples from the Hutterite population, a population isolate in North America, performed HLA typing across all couples and found that couples had more MHC diversity than expected under random mating [22]. Two additional studies, of 200 Amerindian couples [23] and 450 Japanese couples [24], respectively, concluded that the differences between the HLA-types of real couples were not significantly more different than the HLA types of random pairs of individuals. Finally, additional work has investigated whether the remnants of degraded HLA proteins end up in sweat, urine or saliva and can therefore be detected by potential mates through scent. To test the hypothesis that MHC-dependent mate selection in humans is mediated through olfactory processes, researchers have performed so-called 'sweaty t-shirt' experiments, and shown that females indicate an odor preference towards men that carry divergent HLA alleles relative to their own [25,26].

Studies of genetic variation (beyond the classical HLA types) in humans have sought to provide clarity as to whether humans do indeed select mates, at least in part, such that diversity across the MHC increases in offspring. An initial analysis of SNP array-based genotyping data (variation with minor allele frequency (MAF) > 5%) assembled by the HapMap 2 Consortium [27] examined 30 European-ancestry mate pairs and 30 African-ancestry mate

pairs and reported evidence of dissimilar MHC variation in couples of European descent ($p = 0.015$) [17]. Conversely, no such effect was observed in the African-ancestry sample ($p = 0.23$) [17]. A subsequent analysis in the same Hapmap Phase 2 European-ancestry data, but including an additional 24 European-ancestry mate-pairs genotyped as part of HapMap Phase 3 [28], failed to replicate the initial finding [29]. This second analysis demonstrated that the low sample size of the initial analysis, making the study sensitive to small changes in parameter choices, and failure to correct for multiple testing explained the initial report. Neither analysis of the 24 new mate-pairs nor joint analysis of all 54 available European-ancestry mate pairs revealed increased MHC dissimilarity in mates ($p = 0.351$ and $p = 0.143$, respectively).

Here, we aim to test whether human mate pairs are indeed more dissimilar across the MHC, using a sample set that represents an order-of-magnitude increase over the initial reports. Specifically, we test the hypothesis that MHC variation is discordant between couples by analyzing a dataset of 239 unrelated Dutch mate pairs, whole-genome sequenced as part of the Genome of the Netherlands (GoNL) project [30]. The density and resolution of the whole-genome sequence data allow us to test for discordant MHC variation in mate pairs with respect to (a) common variation only (MAF > 1%); (b) the full frequency spectrum of genetic variants, including single nucleotide variants and short insertions and deletions; and (c) imputed amino acids and human leukocyte antigen (HLA) types within the MHC [31].

Results

Reproducing the initial HapMap analysis

We first sought to reproduce the finding of MHC-dependent mate selection in humans reported from an analysis of common variation in the Hapmap Phase 2 data [17], with the goal of not only replicating results but also aligning methodologies. The previous analysis used 30 trios of Northern- and Western-European ancestry living in Utah, USA (called the CEU sample) and 30 trios collected from the Yoruba population in Ibadan, Nigeria (called the YRI sample) [27,32,33] to evaluate MHC genetic dissimilarity in mate pairs. After reproducing the quality control procedures from the initial analysis as closely as possible (**Materials and Methods**), 27 CEU and 27 YRI mate-pairs remained for analysis (**Table 1**).

We used the same measure for genetic similarity between two individuals as defined in the initial report [17]: Qc, defined as 'the proportion of identical genotypes (at variant positions)' between mate pairs (**Materials and Methods**). We compared the average similarity across real couples to the average similarity across randomly generated mate pairs (created by randomly drawing a male and a female from the sample) and obtained results that are close, but not identical to, the initial report (**Figure 1**). We calculated the difference between average genetic similarity across all true mate pairs and average genetic similarity across permuted mate pairs (i.e., average Qc across a null distribution; **Figure 1**) to explicitly quantify how genetic similarity in true mate pairs deviates from the null distribution. We call this metric ΔQc. We found that the CEU mate pairs demonstrated nominally-significant ($p < 0.05$) genetic dissimilarity across the MHC compared to permuted mate pairs (ΔQc = -0.013, 2-sided $p = 0.023$), while mate-pairs in the YRI samples indicated no such relationship (ΔQc = 0.003, 2-sided $p = 0.442$). Genome-wide, CEU mate pairs showed no pattern of genetic similarity or dissimilarity (ΔQc = -0.008, 2-sided $p = 0.100$) while YRI mate-pairs showed a pattern of genome-wide similarity (average Qc = 0.011, 2-sided $p < 10^{-6}$), consistent with the original report [17].

Testing MHC-specific genetic dissimilarity in the Genome of the Netherlands

Next, we sought to test if there was evidence for MHC-dependent mate selection in mate pairs collected as part of the Genome of the Netherlands (GoNL) project [30]. GoNL is comprised of Dutch-ancestry trios (confirmed by principal component analysis [30]) drawn from 11 of the 12 provinces of the Netherlands and whole-genome sequenced at ~14x average coverage on the Illumina HiSeq 2000 [30]. After data quality control and processing in the original project [30], the GoNL dataset contains 248 mate pairs. Because relatedness is a primary confounder for genetic similarity estimations, we calculated sample relatedness in Plink [34] and removed an additional 9 mate pairs with pi-hat > 0.03125 (a threshold corresponding to 5th-degree relatedness; **Materials and Methods**). After this additional quality control, 239 mate pairs remained for analysis. We analyzed the GoNL data (http://www.nlgenome.nl/, see

Online Sources in **Materials and Methods**) from Release 5 of the project, which includes single-nucleotide variants (SNVs) and short (< 20bp) insertions and deletions (indels; **Table 1**).

Sample	Number of mate-pairs	Data type	Variant type(s)	Variant filters	Variant count	ΔQc	p-value
CEU	27	Genotyping (HapMap)	SNVs	MAF > 5%	6,247	-0.0130	0.023
YRI	27				5,773	0.0030	0.442
GoNL	239	Sequencing	SNVs	None	60,339	0.0005	0.702
				HapMap2 sites	8,573	0.0004	0.513
				MAF > 0.5%	44,088	0.0007	0.703
				MAF > 5%	31,145	0.0001	0.709
				Extended MHC	36,413	0.0004	0.696
			SNVs + indels	None	63,357	0.0004	0.693
		HLA imputation	SNVs HLA alleles Amino acids	$r^2 > 0.8$	8,290	-	0.480
				Genic markers $r^2 > 0.8$	2,452	-	0.740

Table 1 : Samples and variants used in analysis. To investigate whether mate selection is MHC-dependent, we analyzed three sample groups: Utah residents with Northern and Western European ancestry (CEU); Yorubans from Ibadan, Nigeria (YRI); and mate-pairs in the Genome of the Netherlands (GoNL) project. The number of mate pairs indicates the number of pairs available after sample quality control. We performed our analysis in common polymorphisms (minor allele frequency (MAF) > 0.05) or common- and low-frequency single nucleotide variants (SNVs, with MAF > 0.5%), as well as including indel variation, where available. For imputed data, we kept only well-imputed data, based on the Beagle imputation quality metric ($r^2 > 0.8$). We additionally restricted the set of variants to only variants within the classical HLA genes including amino acid substitutions, single nucleotide polymorphisms (SNP), insertions and/or deletions (indels) and classical HLA-types ('genic markers').

To test for MHC-dependent mate selection in GoNL, we extracted the MHC (chromosome 6, 28.7 - 33.3Mb on build hg19), calculated Qc across all true GoNL mate pairs, and performed the same permutation scheme as in the HapMap analysis, randomizing the mate pairs and recalculating the average Qc across these randomly-constructed pairs; finally, we calculated ΔQc. All p-values are 1-sided, testing the hypothesis of genetic dissimilarity, unless otherwise stated. Our results showed no evidence for MHC-dependent mate selection (ΔQc = 0.0005, permutation p = 0.702, **Figure 2**). Restricting our analyses to common- and low-frequency SNPs (MAF > 0.5%) or common SNPs only (MAF > 5%) did not change our results (**Table 1**, **Supplementary Figures 1** and **2**), nor did restricting the analysis specifically to the ~2M common SNPs genotyped in HapMap 2 or including the set of ~2M indels sequenced in GoNL into the analysis (**Table 1** and **Supplementary Figure 3**). To test the hypothesis that MHC mating is mediated through olfactory sensory pathways, as hypothesized previously [25,26], we performed the same analysis using an extended definition of the MHC (26.6Mb - 33.3Mb on hg19), which includes a dense cluster of 36 olfactory receptor genes upstream of the HLA Class I region [3]. We observed no statistically significant effect (**Table 1**, and **Supplementary Figures 4** and **5**).

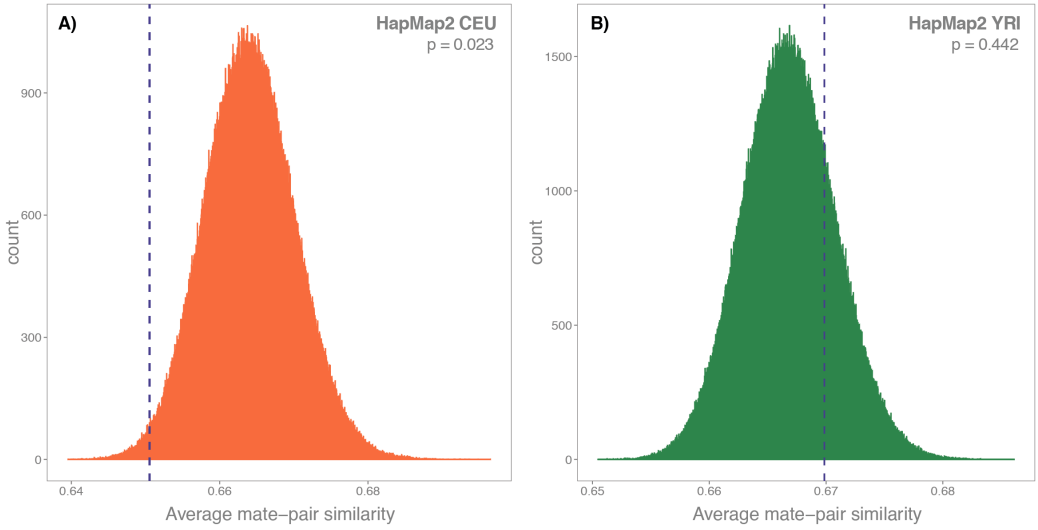


Figure 1 : Genetic similarity across mate pairs in the HapMap 2 data.
The distributions represent the null distribution of average MHC similarity (Qc), across randomly permuted mate pairs from each of the HapMap 2 populations tested (CEU: European samples of Northern and Western descent, orange; YRI: Yorubans in Ibadan Nigeria, green). The average MHC similarity in true mate pairs is marked by the blue dotted line. All p-values are based on 1,000,000 permutations and delta Qc (ΔQc) is the difference between the average real-couple similarity and the average of the distribution or random mate-pair permutations. **(A)** Permutation of the 27 QC-passing HapMap 2 CEU couples. ΔQc = -0.013, 2-sided p = 0.023. **(B)** Permutations of the 27 QC-passing HapMap 2 YRI couples. ΔQc = 0.003, 2-sided p = 0.442.

Though the Netherlands is geographically small and densely populated, both common and rare variation in the GoNL data indicate geographic clustering [30,35–37]. We therefore investigated whether population stratification may explain the discordance between our results and the previous report of MHC-dependent mate selection in humans [17]. We performed genetic similarity analyses in the samples split into three geographic regions ("north," "middle," and "south" as determined by an identity-by-descent analysis [30]), as well as by province. Subsetting by region or province revealed no evidence for subpopulation-specific MHC-dependent mate selection (**Figure 2**). Additionally, accounting for sample ancestry using principal components (**Materials and Methods**) left our results unchanged ($p = 0.78$).

Lastly, we used SNP2HLA [31] to impute 2- and 4-digit HLA alleles, amino acids and SNPs (**Materials and Methods**) into the GoNL samples as a means of evaluating genetic (dis) similarity across imputed HLA types. Given that the dosages output from SNP2HLA are phased, we used the Pearson's correlation (r) across the imputed allele dosages to calculate genetic similarity (instead of the Qc metric). We found no evidence for MHC-dependent mate selection either across all imputed markers ($p = 0.48$, **Table 1**) or by restricting the correlation calculation to only those variants, amino acids, and HLA types within the classical HLA Class I and II gene bodies (and thus more likely to have functional effect; $p = 0.74$, **Table 1**).

Until this point, we had established a null distribution by permuting mate pairs and calculating genetic similarity. To generate an alternative null model for comparison, we randomly sampled 10,000 regions from the genome that either matched the MHC by size (i.e., total span of the region) or by number of variants contained within the region (regardless of the total linear span of the region capturing those markers). For each permutation, we randomly selected the region, computed Qc (averaged across the 239 true mate-pairs) and counted the number of times the mean Qc was as or more dissimilar than that observed in the MHC. We observed no statistically-significant difference when selecting regions based on genomic size or total number of markers in the region, after accounting for multiple testing (one-sided $p = 0.08$ and 0.02, respectively).

Discussion

Using the whole-genome sequencing data of 239 mate pairs, we have performed, to our knowledge, the most comprehensive investigation of MHC-dependent human mate selection to date. The Genome of the Netherlands resource provided both an increased sample size compared to previous efforts [17,29] and high density genetic variation data, allowing for analyses of rare variants, indels, and imputed HLA types. However, despite the size and genomic resolution of the data, our results indicate no evidence for MHC-dependent mate selection in humans. We performed further analyses to investigate the potential effects of geographical clustering of rare variants [30,35], but the results left our results and interpretation unchanged.

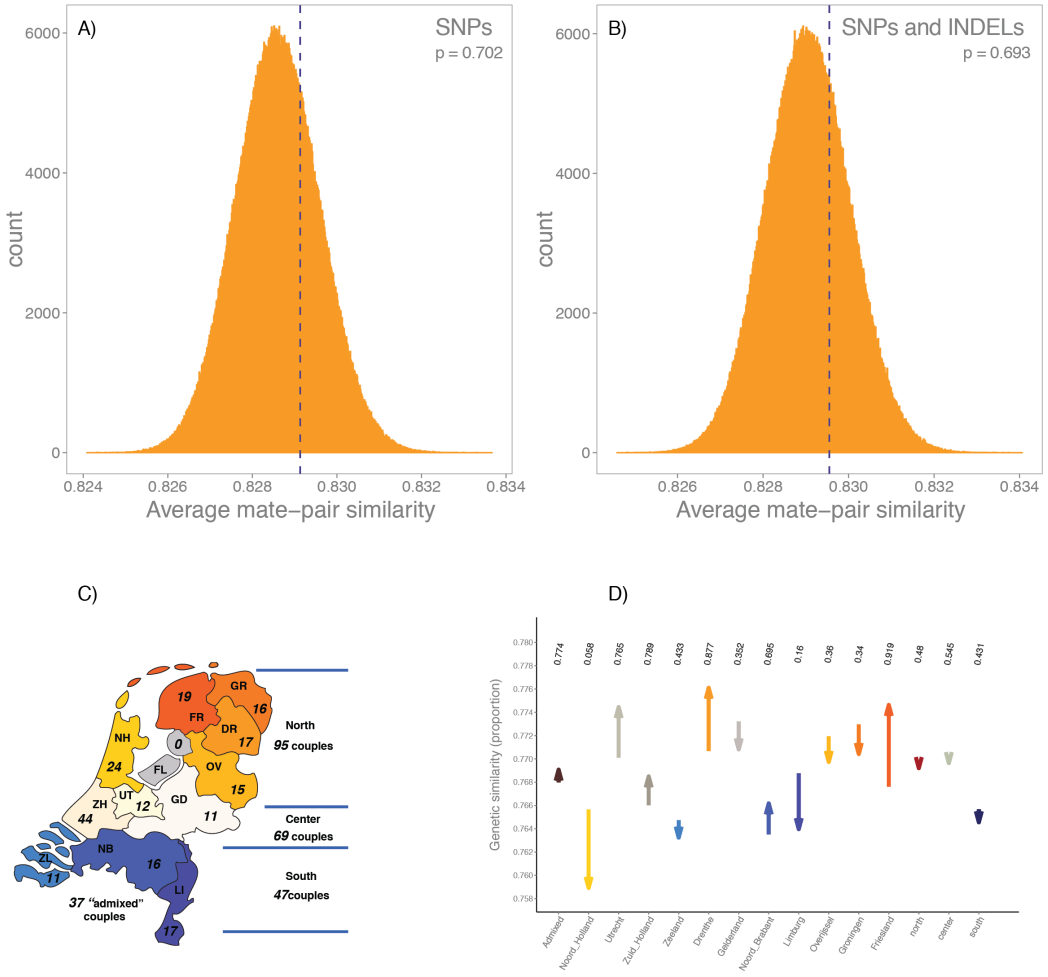


Figure 2 | Genetic similarity across the MHC for 239 Dutch-ancestry mate-pairs.
Panels **(A)** and **(B)** show the null distribution (histograms) of average mate-pair genetic similarity of permuted (i.e., non-real) male-female pairs. We performed a total of 1,000,000 permutations to generate the distribution. The average genetic similarity across 239 real mate pairs is represented with a blue vertical dotted line. **(A)** Genetic similarity measured across all biallelic variants within the MHC (p = 0.702). **(B)** Genetic similarity measured across all biallelic variants and insertions/deletions (indels) in the MHC (p = 0.693). **(C)** The GoNL samples were drawn from 11 of the 12 Dutch provinces. Here, we indicate the number of true mate-pairs available for analysis where both members of the mate-pair come from the same geographic region. These three geographic regions (north, center, and south) are derived from previously-performed population genetic analyses of the GoNL data. **(D)** Genetic similarity of mate-pairs, split by province. The arrows start at the average genetic similarity of permuted (i.e., null) mate pairs and stops at the average genetic similarity across true mate-pairs. Corresponding, one sided p-values for the genetic dissimilarity within couples are marked above.

Notably, our results are inconsistent with an initial investigation of MHC-dependent mate selection using genome-wide genetic variation data [17]. Though these previous findings do not align with our own, the initial report of MHC-dependent mate selection in humans was likely too small (N = 60) to draw conclusive results. Further, potential confounders, including cryptic relatedness and inbreeding amongst the studied samples, along with a lack of multiple testing correction, all likely contributed to this initial positive finding, subsequently contradicted in follow-up analyses of the same samples [29]. By interrogating a larger sample size, more stringently removing samples for relatedness and inbreeding, and performing analyses that account for potential population stratification, we believe our results provide robust information as to whether mate selection in humans is influenced, at least in part, by individuals' genetic composition across the MHC. Additionally, our results are consistent with investigation of MHC-dependent mate selection using HLA types in similarly-sized sample sets [23,24].

While our results indicate that human mate selection is independent of genetic variation in the MHC, a number of studies examining genetic variation and complex traits have found a plethora of positive evidence for assortative mating in humans based on non-MHC genetic factors. Previous studies have shown that human mate choice is associated to quantitative features (such as height) [38], to socioeconomic factors and risk for multifactorial disease [39–41]. A recent analysis in > 24,000 mate pairs, drawn from a number of cohorts including the UK Biobank [42] and 23andMe, focused on genomic loci associated to a number of multifactorial traits and found significant correlation between spouses at loci associated to height and body mass index [43]. By building a genetic predictor in one member of a spousal couple and applying it in the second member, the study also revealed varying degrees of spousal correlation at loci associated to waist-to-hip ratio, educational attainment, and blood pressure [43] in 7,780 couples from the UK Biobank. These correlations represent only a small slice of the numerous factors - both genetic and non-genetic - that contribute to mate selection in the human population. Importantly, however, these observations are correlative; the extent to which these associations are potentially causative remains to be explored.

Though our analysis offers several improvements over previous analyses examining MHC-dependent mate selection, several limitations remain. First, as highlighted by the assortative mating studies discussed above, our sample size may not be large enough to detect a more modest signal for MHC-dependent mate selection, if such a phenomenon exists. Mate selection is likely influenced by a host of hundreds, if not thousands, of factors, all of which likely have modest effect. Therefore, analysis of 239 samples may not be sufficiently well powered to detect such an effect. Further, while we have used permutations of mate pairs to establish a null distribution to which we can compare true mate-pair genetic similarity, this distribution may not be sufficiently informative to detect MHC-dependent effects. Indeed, the authors of the initial analyses [17] reported similar difficulties establishing a null comparator: they sought to additionally use genome-wide genetic similarity as a basis of comparison for MHC similarity, but observed higher genome-wide similarity in YRI samples

compared to the CEU [17]. Given the uniqueness of the MHC, from its gene density and extensive linkage disequilibrium to its high genetic diversity, finding a genomic region with similar properties to use as a null comparator is essentially impossible; permutations of real mate pairs into random pairs, while not ideal, is likely the best null derivation for this experiment. Additionally, our analysis only examines one ancestral population. Analyses extended into other (non-European) samples may result in different findings.

Untested here is the hypothesis that preferential mating may favour specific combinations of HLA alleles that collectively result in an 'optimal' number of antigens that can be presented to T cell receptors. Previous studies indicate that this phenomenon may occur, specifically across Class I classical HLA genes [44], and may provide an alternative mechanism for MHC-mediated mate selection. Given the number of HLA allele combinations that would need to be constructed and analyzed to test such a hypothesis, power (after multiple test correction) would be vanishingly small. We therefore have not tested this specific hypothesis here. However, additional information regarding gene function may make testing this hypothesis feasible in the future.

Despite these limitations, our analysis represents an improved investigation of MHC-dependent mate selection, through interrogated sample size as well as in the spectrum of genetic variation tested. Our data indicate no MHC-mediated preferential mating patterns in our European-ancestry sample. While MHC-mediated preferential mating has been reported in non-human animal models, such a mechanism in humans is either absent or may be one of many subtle contributors to mating patterns and behaviours.

Materials and methods

Quality control of HapMap and Genome of the Netherlands data

Related samples, by definition, are more likely to share more genetic variation compared to two unrelated individuals. To ensure that relatedness was not confounding our analyses, we performed basic quality control (QC) in the CEU, YRI and Genome of the Netherlands (GoNL) sample sets separately. The initial HapMap 2 analysis [17] filtered related couples by looking at the normalized Qc measure and defining outliers. We used the identity-by-descent (IBD) estimates, computed with Plink 1.9 [45] using the --genome command. Though this approach differs from the initial analysis, using IBD estimates are an established means for identifying related samples using genetic variation data.

To estimate relatedness, we first used Plink 1.9 to assemble a set of high-quality SNPs with minor allele frequency (MAF) > 10% and genotyping missingness < 0.1%. We pruned this set of SNPs at a linkage disequilibrium (r^2) threshold of 0.2. Additionally, we removed SNPs in the MHC, lactase (*LCT*) locus on chromosome 2, and in the inversions on chromosomes 8 and 17 (genomic coordinates in **Supplementary Table 1**). We calculated relatedness (--genome in Plink) across all individuals in the CEU and YRI mate pairs. We discarded three mate pairs (N = 6 samples) from the CEU sample and three mate pairs (N = 6 samples) from the YRI sample. We defined relatedness as pi-hat > 0.05 (i.e., shared 1/20th of the genome), close to the 1/22nd threshold used by Derti *et al.* [29]. Our filtering produced nearly identical results to the initial analyses (Supplementary Text S2 of [29]). Due to our slightly more stringent cutoff threshold, we additionally exclude the related pair of samples NA12892 and NA06994.

We filtered for relatedness in GoNL in an identical manner. We used a more stringent cryptic relatedness threshold of pi-hat > 0.03125, corresponding to 5th-degree relatives. We discarded 9 couples from our analysis, leaving 239 QC-passing mate pairs.

Calculating genetic similarity in mate pairs

We define genetic similarity across a mate pair (called Qc, per the initial report [17]) as the proportion of variants that are identical across a pair of individuals. Homozygous genotypes comprised of the same alleles (e.g., AA in sample 1 and AA in sample 2) are considered 100% similar; heterozygous genotypes (e.g., AB in both samples) are considered 50% similar, as they could have either the same or opposite phase; and all other combinations are considered 0% similar.

We note that in the initial report [17], genetic similarity was defined as: R = (Qc - Qm)/(1-Qm), where Qm is the average genetic similarity across all possible mate-pairs (real and permuted) that can be constructed in the sample. We note that the R measure is a linear transformation of Qc measure, as Qm is a constant for the analyzed sample. Further, Qm is not an unbiased estimate of the average genetic similarity within random mate-pairs for

two reasons: (1) because it includes both real mate-pairs and female-male pairs constructed by selecting two random individuals in the dataset; and (2) because the sample pairs over which Qm is averaged are not independent (i.e., the same individual is paired with all possible matches and thus considered multiple times when computing Qm). We therefore perform all our analyses using only the Qc measure of genetic similarity.

Replicating the original HapMap analysis

The HapMap 2 genotyping data is publicly available [27,32,33] and includes a total of 3,965,296 single nucleotide polymorphisms (SNPs). We extracted the MHC region (29.7 - 33.3Mb on chromosome 6, build hg18, as defined in the original analysis) from each population separately: people of Northern and Western European ancestry (the CEU) and Yorubans from Ibadan, Nigeria (YRI). We performed these analyses in 27 CEU mate-pairs and 27 YRI mate-pairs, after filtering on sample relatedness (see *Quality Control of HapMap and Genome of the Netherlands Data*).

Evaluating significance of genetic similarity in true mate-pairs

To evaluate whether genetic (dis)similarity in mate-pairs was significantly different than genetic similarity between two random individuals, we performed a permutation analysis. Specifically, we created 'null' (i.e., non-real) male-female pairs by randomly permuting the individuals in the true mate-pairs. Within any single permutation, we allowed for at most 1 real couple to enable faster sampling of random mate-pairs. We performed a total 1,000,000 permutations to generate a null distribution (**Figures 1** and **2**). Finally, we count the number of permutations that yield an average Qc that is the same or lower than the Qc measured in the true mate-pairs. The total number of such permutations divided by 1,000,000 is the exact p-value of the test. This permutation scheme was used to evaluate the significance of Qc as measured in common variants, all variants, and imputed HLA variants.

Analysis of mate-pairs in the Genome of the Netherlands (GoNL) data

We repeated the same analysis in the Genome of the Netherlands data (GoNL), in the 239 mate-pairs that passed quality control. In the GoNL data, we estimated Qc in three sets of variants (**Table 1**): common biallelic variants only, all available single nucleotide variants regardless of frequency, and in all available variants (including insertions and deletions). For a fourth set of variants - imputed HLA variation - we measured genetic similarity using Pearson's correlation (r), as the imputed variation data was phased and left no ambiguity as to how heterozygous genotypes correlated (e.g., the difference between observing the AB genotype in Sample 1 and the AB genotype in Sample 2; or observing the AB genotype in Sample 1 and the BA genotype in Sample 2). To evaluate the significance of Qc in true mate-pairs, we used the identical permutation scheme as used in the HapMap analysis and described above.

HLA IMPUTATION

We use SNP2HLA [31] and a reference panel built from HLA typing performed in the Type 1 Diabetes Genetics Consortium (T1DGC) (containing 8,961 markers) [31] to impute SNPs, HLA types and amino acid substitutions across 8 classical HLA loci. For imputation, 3,256 SNPs in GoNL overlap the T1DGC reference panel data. After the MHC imputation was complete, we first performed quality control, removing samples where the total number of imputed alleles is > 2.5 (introduced by imprecision in the imputation algorithm) and removing all variants for which the imputation quality ('info') metric is < 0.8.

CORRECTING FOR POPULATION STRUCTURE IN THE GONL SAMPLES

As the Dutch samples are drawn from 11 of the 12 provinces in the Netherlands, subtle population structure can be observed in both common and rare variants [30]. Analysis in the original GoNL effort indicated that the first two principal components reveal a subtle north-to-south gradient, and analysis of rarer (so-called "f_2") variants (two alleles appearing in the entire dataset) indicate strong clustering within geographical regions (north, center, and south, as inferred by IBD analyses) [30]. We thus sought to explore whether population structure, either across the country or by province, may be confounding a potential signal for MHC-dependent mate selection. To do this, we used principal component analysis as well as province-specific analyses.

Genetic PCs are calculated on an individual basis and are an alternative means of unravelling genetic ancestral clustering between individuals. We first needed to collapse individual-level PC loadings into a single value that represented a single mate-pair. We call this collapsed PC the 'mate-pair PC' (PC_{mp}). Assume that the PC1 loading for a female in a given mate-pair is denoted $PC1_f$, and PC1 loading for the male in that mate-pair is denoted $PC1_m$, then $PC1_{mp}$ (continuing up to PC 'n') is defined as follows:

$$PC1_{mp} = (PC1_f - PC1_m)^2$$

...

$$PCn_{mp} = (PCn_f - PCn_m)^2$$

In this way, we used the PCs of the GoNL individuals to obtain, for each (real or permuted) pair of individuals, a PC_{mp} value that is equal to 0 if the loadings of the two individuals in a pair are identical for a given PC, or becomes increasingly large as the two samples' loadings on a particular PC diverge.

We then used the mate-pairs of one random permutation of the 239 mate-pairs in GoNL to train a linear regression model that approximates the genetic similarity between two individuals (Qc), using the mate-pair PCs as defined above:

$$Qc_{hat} \sim PC1_{mp} + PC2_{mp} + \ldots + PC10_{mp}$$

Qc_{hat} estimates the genetic similarity explained by the first 10 PCs for all mate-pairs (real or permuted) as well as residuals (Qc_{res}) from this regression. If there is preferential mating among the true mate-pairs in GoNL, the residuals of this regression model should be systematically different compared to residuals from randomly-assigned male-female pairs. We performed the same initial permutation analysis, on the whole set of 239 true mate-pairs, but using Qc_{res} (instead of Qc) as a measure of genetic similarity adjusted for population stratification. We then compared where the average Qc_{res} across the 239 true mate-pairs falls within the distribution of average Qc_{res} across 239 randomly generated male-female pairs.

Genetic dissimilarity in non-MHC regions

In addition to permuting mate-pairs to establish a null distribution for Qc, we also wanted to establish a null distribution of Qc by randomly sampling regions from the genome that were matched to the MHC based on different characteristics. Because the MHC is an extremely unique genomic region - in gene density, in span of linkage disequilibrium, and in genetic variability - it is nearly impossible to identify regions of the genome that behave identically to the MHC. To identify genomically similar regions to the MHC from which we could construct a null distribution for Qc, we identified regions that either (1) were the same genomic span as the MHC (~3.6 Mb), or (2) contained approximately the same number of markers (~40k), regardless of the linear span of that window. For each criterion (SNP density or span), we randomly sampled 10,000 regions from the genome and computed average Qc across all 239 true mate-pairs, for each region; we compared these distributions to Qc calculated in true mate-pairs across the MHC.

Acknowledgements

The Genome of the Netherlands Consortium generated and analyzed the whole-genome sequencing data analyzed here. A complete list of the Genome of the Netherlands members and affiliations is provided in the Supplementary Materials.

We thank Paul I.W. de Bakker for supporting MCS with funding from VIDI grant 91712354 from the Dutch Organization for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) - ZonMw) and for providing critical feedback on the manuscript.

References

1. De Bakker, P. I. W. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38,** 1166–1172 (2006).

2. Horton, R. *et al.* Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* **60,** 1–18 (2008).

3. Horton, R. *et al.* Gene map of the extended human MHC. *Nat. Rev. Genet.* **5,** 889–899 (2004).

4. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–6 (2014).

5. Pereyra, F. *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330,** 1551–1557 (2010).

6. Hinks, A. *et al.* Fine-mapping the MHC locus in juvenile idiopathic arthritis (JIA) reveals genetic heterogeneity corresponding to distinct adult inflammatory arthritic diseases. *Ann. Rheum. Dis.* **76,** 765–772 (2017).

7. Xie, G. *et al.* Association of Granulomatosis With Polyangiitis (Wegener's) With HLA–DPB1*04 and SEMA6A Gene Variants: Evidence From Genome-Wide Analysis. *Arthritis & Rheumatism* **65,** 2457–2468 (2013).

8. Cortes, A. *et al.* Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with ERAP1. *Nat. Commun.* **6,** 7146 (2015).

9. Zhang, F.-R. *et al.* HLA-B*13:01 and the dapsone hypersensitivity syndrome. *N. Engl. J. Med.* **369,** 1620–1628 (2013).

10. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).

11. Potts, W. K. & Wakeland, E. K. Evolution of diversity at the major histocompatibility complex. *Trends Ecol. Evol.* **5,** 181–187 (1990).

12. Penn, Penn & Potts. The Evolution of Mating Preferences and Major Histocompatibility Complex Genes. *Am. Nat.* **153,** 145 (1999).

13. Potts, W. K., Manning, C. J. & Wakeland, E. K. Mating patterns in seminatural populations of mice influenced by MHC genotype. *Nature* **352,** 619–621 (1991).

14. Olsson, M. *et al.* Major histocompatibility complex and mate choice in sand lizards. *Proc. Biol. Sci.* **270 Suppl 2,** S254–6 (2003).

15. Landry, C., Garant, D., Duchesne, P. & Bernatchez, L. 'Good genes as heterozygosity': the major histocompatibility complex and mate choice in Atlantic salmon (Salmo salar). *Proc. Biol. Sci.* **268,** 1279–1285 (2001).

16. OlsÉn, K. H., Grahn, M., Lohm, J. & Langefors, Å. MHC and kin discrimination in juvenile Arctic charr, Salvelinus alpinus (L.). *Anim. Behav.* **56,** 319–327 (1998).

17. Chaix, R., Cao, C. & Donnelly, P. Is mate choice in humans MHC-dependent? *PLoS Genet.* **4,** 1–5 (2008).

18. Aeschlimann, P. B., Häberli, M. A., Reusch, T. B. H., Boehm, T. & Milinski, M. Female sticklebacks Gasterosteus aculeatus use self-reference to optimize MHC allele number during mate selection. *Behav. Ecol. Sociobiol.* **54,** 119–126 (2003).

19. Yamazaki, K. *et al.* Control of mating preferences in mice by genes in the major histocompatibility complex. *J. Exp. Med.* **144,** 1324–1335 (1976).

20. Penn, D. & Potts, W. K. Untrained mice discriminate MHC-determined odors. *Physiol. Behav.* **64,** 235–243 (1998).

21. Brown, R. E., Roser, B. & Singh, P. B. Class I and class II regions of the major histocompatibility complex both contribute to individual odors in congenic inbred strains of rats. *Behav. Genet.* **19,** 659–674 (1989).

22. Ober, C. *et al.* HLA and mate choice in humans. *Am. J. Hum. Genet.* **61,** 497–504 (1997).

23. Hedrick, P. W. & Black, F. L. HLA and mate selection: no evidence in South Amerindians. *Am. J. Hum. Genet.* **61,** 505–511 (1997).

24. Ihara, Y., Aoki, K., Tokunaga, K., Takahashi, K. & Juji, T. HLA and Human Mate Choice. Tests on Japanese Couples. *Anthropol. Sci.* **108,** 199–214 (2000).

25. Wedekind, C., Seebeck, T., Bettens, F. & Paepke, A. J. MHC-dependent mate preferences in humans. *Proc. Biol. Sci.* **260,** 245–249 (1995).

26. Wedekind, C. & Füri, S. Body odour preferences in men and women: do they aim for specific MHC combinations or simply heterozygosity? *Proc. Biol. Sci.* **264,** 1471–1479 (1997).

27. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).

28. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).

29. Derti, A., Cenik, C., Kraft, P. & Roth, F. P. Absence of Evidence for MHC–Dependent Mate Selection within HapMap Populations. *PLoS Genet.* **6,** e1000925 (2010).

30. Francioli, L. C. *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–825 (2014).

31. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8,** e64683 (2013).

32. The International HapMap Consortium. *et al.* The International HapMap Project. *Nature* **426,** 789–796 (2003).

33. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).

34. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

35. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44,** 243–246 (2012).

36. Mathieson, I. & McVean, G. Demography and the Age of Rare Variants. *PLoS Genet.* **10,** e1004528 (2014).

37. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

38. Silventoinen, K., Kaprio, J., Lahelma, E., Viken, R. J. & Rose, R. J. Assortative mating by body height and BMI: Finnish twins and their spouses. *Am. J. Hum. Biol.* **15,** 620–627 (2003).

39. Vandenburg, S. G. Assortative mating, or who marries whom? *Behav. Genet.* **2,** 127–157 (1972).

40. Hippisley-Cox, J., Coupland, C., Pringle, M., Crown, N. & Hammersley, V. Married couples' risk of same disease: cross sectional study. *BMJ* **325,** 636 (2002).

41. Willemsen, G., Vink, J. M. & Boomsma, D. I. Assortative mating may explain spouses' risk of same disease. *BMJ* **326,** 396 (2003).

42. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12,** e1001779 (2015).

43. Robinson, M. R. *et al.* Genetic evidence of assortative mating in humans. *Nat. hum. behav.* **1,** 0016 (2017).

44. Buhler, S., Nunes, J. M. & Sanchez-Mazas, A. HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics* **68,** 401–416 (2016).

45. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 1–16 (2015).

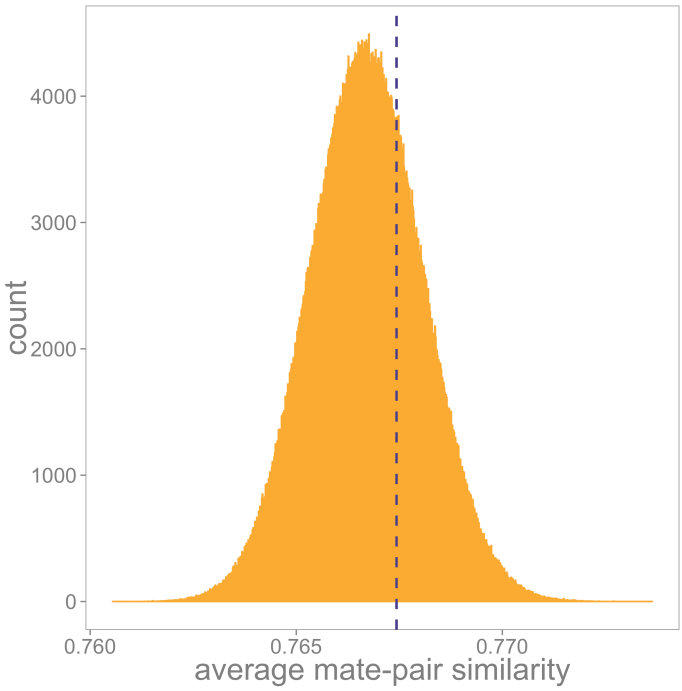
Online Resources

[1] SNP2HLA: http://software.broadinstitute.org/mpg/snp2hla/

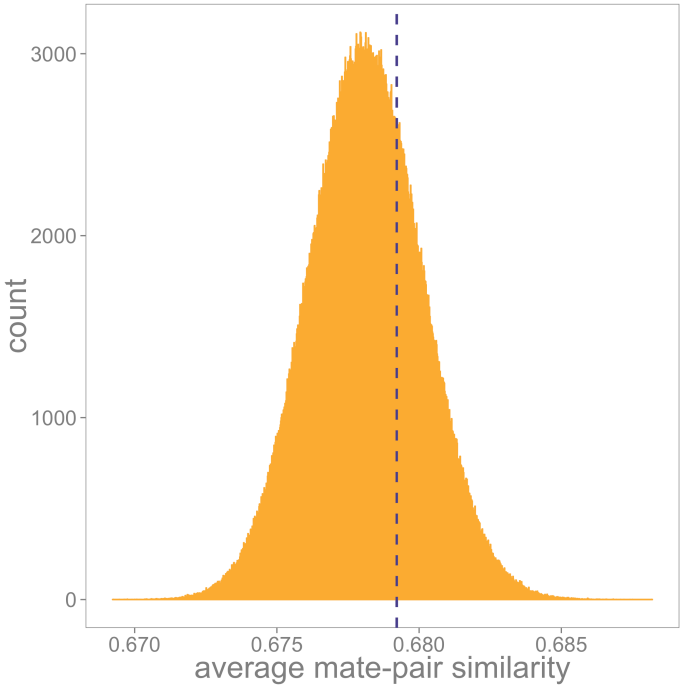
[2] Access to the Genome of the Netherlands data: http://www.nlgenome.nl/

[3] Code used to perform mate-pair permutations and asses similarity: https://github.com/mcretu-umcu/matingPermutations

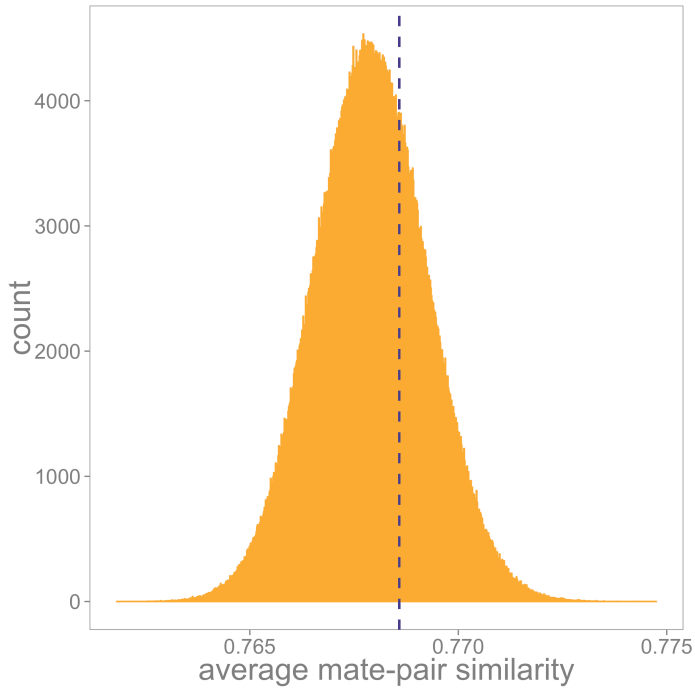
SUPPLEMENTARY INFORMATION TO CHAPTER 3



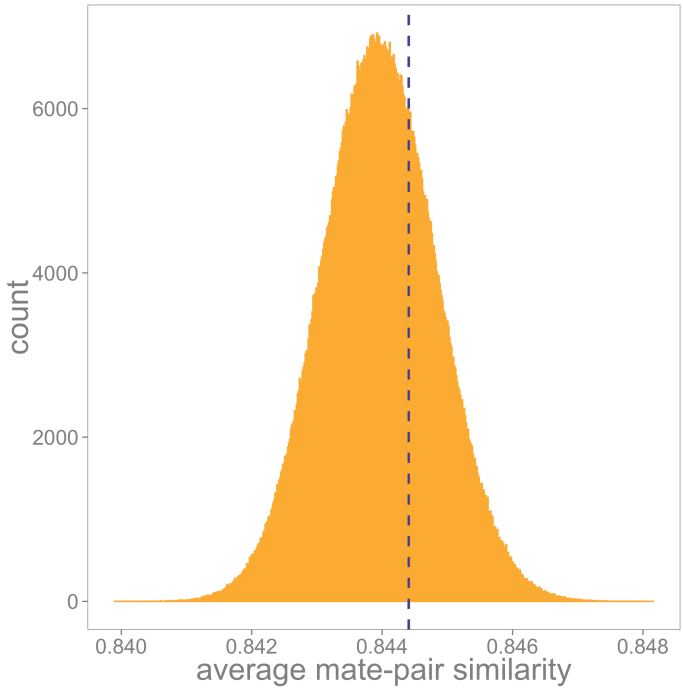
Supplementary Figure1: GoNL mate-pair permutation across the MHC.
Using SNPs with MAF > 0.005; ΔQc = 0.0007, p = 0.703



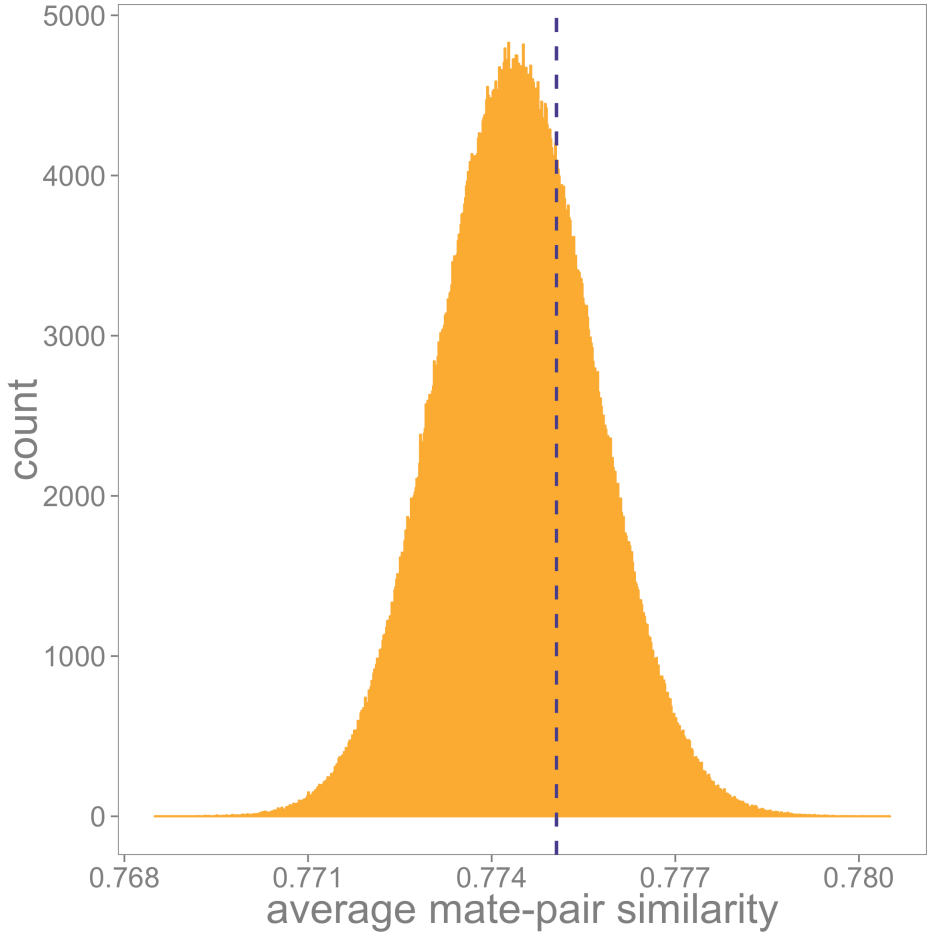
Supplementary Figure2: GoNL mate-pair permutation across the MHC.
Using SNPs with MAF > 0.05; ΔQc = 0.0001, p = 0.709



Supplementary Figure3: GoNL mate-pair permutation across the MHC.
Using all markers with MAF > 0.005; ΔQc = 0.0006, p = 0.693



Supplementary Figure4: GoNL mate-pair permutation across the extended MHC.
Using all SNPs; ΔQc = 0.0004, p = 0.696



Supplementary Figure5: GoNL mate-pair permutation across the extended MHC.
Using SNPs with MAF > 0.005; ΔQc = 0.0006, p = 0.695

CHAPTER4

MAPPING AND PHASING OF STRUCTURAL VARIATION IN PATIENT GENOMES USING NANOPORE SEQUENCING

Mircea Cretu Stancu[1,*], Markus J. van Roosmalen[1,*], Ivo Renkens[1], Marleen M. Nieboer[1], Sjors Middelkamp[1], Joep de Ligt[1], Giulia Pregno[2], Daniela Giachino[2], Giorgia Mandrile[2], Jose Espejo Valle-Inclan[1], Jerome Korzelius[1], Ewart de Bruijn[1], Edwin Cuppen[3], Michael E. Talkowski[4,5,6], Tobias Marschall[7,8], Jeroen de Ridder[1], Wigard P. Kloosterman[1,$]

1 | Department of Genetics, Center for Molecular Medicine, University Medical Center Utrecht, 3584 CG Utrecht, The Netherlands.
2 | Medical Genetics Unit, Department of Clinical and Biological Sciences, University of Torino, Turin, Italy.
3 | Department of Genetics and Cancer Genomics Netherlands, Center for Molecular Medicine, University Medical Center Utrecht, 3584 CG Utrecht, The Netherlands
4 | Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, 02114, USA
5 | Department of Neurology, Harvard Medical School, Boston, MA, 02115, USA
6 | Program in Population and Medical Genetics and Stanley Center for Psychiatric Research, The Broad Institute of M.I.T. and Harvard, Cambridge, MA, 02142, USA
7 | Center for Bioinformatics, Saarland University, 66123 Saarbrücken, Germany
8 | Max Planck Institute for Informatics, 66123 Saarbrücken, Germany
$ | Correspondence: w.kloosterman@umcutrecht.nl
* | Equal contribution

Manuscript adapted from Nature Communications

Abstract

Despite improvements in genomics technology, the detection of structural variants (SVs) from short-read sequencing still poses challenges, particularly for complex variation. Here we analyze the genomes of two patients with congenital abnormalities using the MinION nanopore sequencer and a novel computational pipeline - NanoSV. We demonstrate that nanopore long reads are superior to short reads with regard to detection of *de novo* chromothripsis rearrangements. The long reads also enable efficient phasing of genetic variations, which we leveraged to determine the paternal origin of all *de novo* chromothripsis breakpoints and to resolve the structure of these complex rearrangements. Additionally, genome-wide surveillance of inherited SVs reveals novel variants, missed in short-read datasets, a large proportion of which are retrotransposon insertions. We provide a first exploration of patient genome sequencing with a nanopore sequencer and demonstrate the value of long-read sequencing in mapping and phasing of SVs for both clinical and research applications.

Introduction

Second-generation DNA sequencing has become an essential technology for research and diagnosis of human genetic disease. Sequencing of human exomes has resulted in dramatic increases in novel gene discovery for Mendelian disorders[1], while whole-genome sequencing has revealed that myriad diseases are caused by genetic changes that can occur both within genes as well as in the noncoding genome[2]. As a result, genome sequencing has seen rapid adoption in clinical decision making, as the complete picture of a patient's unique mutation profile enables personalization of treatment strategies[3,4].

Robust methods to detect structural variants (SVs) in human genomes are essential, as SVs represent an important class of genetic variation that accounts for a far greater number of variable bases than single nucleotide variations (SNVs)[5]. Moreover, SVs have been implicated in a wide range of genetic disorders[6].

A particularly revolutionary development in genome sequencing is the usage of protein nanopores to measure DNA sequence directly and in real time[1,7]. The first successful implementation of this principle in a consumer device was achieved in 2014 by Oxford Nanopore Technologies (ONT) with the introduction of the MinION[8]. The MinION can sequence stretches of DNA of up to hundreds of kilobases in length, which already resulted in the sequencing of the genomes of several organisms[9,10]. Because MinION-based sequencing requires almost no capital investment and current devices have a very small footprint, mainstream adoption of these sequencers has the potential to fundamentally change the current paradigm of sequencing in centralized centers.

An important and natural application of the long reads produced by nanopore sequencing is identifying SVs. Long-read sequencing is breaking ground for the discovery of SVs at an unprecedented scale and depth[11]. The first success has been achieved using the Pacific BioSciences SMRT long-read sequencing platform[12,13], and alternative methods to capture long-range information have been introduced such as BioNano optical mapping[14] and 10X Genomics linked-read technology[15]. While short-read next-generation sequencing data rely on multiple (often) indirect sources of information in order to accurately identify SVs, structural changes can be directly reflected in long-read data.

In this work, we demonstrate the sequencing of the whole diploid human genomes of two patients on the MinION sequencer at 11-16X depth of coverage. The two patients suffer from congenital disease resulting from complex chromothripsis. We employ a novel computational pipeline to demonstrate the feasibility of using MinION reads to detect *de novo* complex SV breakpoints, at high sensitivity. The long reads from the MinION allow efficient phasing of genetic variations (SNVs as well as SVs) and enable us to resolve the long-range structure of the chromothripsis in the patients. Moreover, we identify a significant proportion of SVs that are not detected in short-read Illumina sequencing data of the same patient genomes.

Results

Sequencing of patient genomes with nanopore MinION

As a first step toward real-time clinical genome sequencing, we evaluated the use of the MinION device to sequence the genomes of two patients with multiple congenital abnormalities[16], henceforth denoted as Patient1 and Patient2 respectively.

We extracted DNA from patient cells and sequenced this on the MinION. For Patient1, we used R7, R9 and R9.4 pore chemistries (**Supplementary Table 1**) generating a total of 8.2M template sequencing reads from 122 sequencing runs. For Patient2, we exclusively used R9.4 runs and performed only 13 runs (1.89M reads), which required approximately five days of sequencing on seven parallel MinION instruments at a cost of around $7,000 (**Supplementary Fig. 1**) and produced a coverage of 11x. We observed that 82.1% (Patient1) and 98.9% (Patient2) of these reads could be mapped to the human reference genomes and were useful for further analyses. Read lengths were highly variable for Patient1, as a result of differences in library prep methods, with a mean of 6.9kb for template reads, while for Patient2 we reached an average of 16.2kb with consistent read length distributions across each of the 13 runs (**Supplementary Fig. 2**).

Raw sequencing data were transformed into FASTQ format using Poretools and sequence reads were mapped to the human reference genome (GRCh37) using LAST[17]. We uniquely aligned 99% of R7/R9 2D reads or R9.4 1D reads flagged as 'passed' after EPI2ME basecalling, while this dropped to 55% for R9-based 'failed' 2D reads (**Supplementary Fig. 3**). We evaluated the mapping accuracy by calculating the percentages of identical bases (PID) between mapped reads and the reference genome. We observed a mean PID of 90% for R7 2D and R9 2D, 85% for R9 template and 89% for R9.4 template reads based on LAST mapping (**Supplementary Fig. 4**). An analysis of error rates and types, within the Patient2 data (i.e.: R9.4 reads only), shows that from an observed per-base error rate of 15.1%, indel errors were the dominant error class (10%: 9.1% deletions, 0.9% insertions), followed by mismatches (5.1%). We found a 2.6-fold increase in deletion errors for sequences overlapping homopolymers and 1.4-fold for sequences overlapping tandem repeats (**Supplementary Fig. 5**). Furthermore, both deletion and mismatch rates were increased in regions with high GC content (**Supplementary Fig. 6**).

We obtained a mean coverage depth of 16x and 11x for Patient1 and Patient2, respectively (**Supplementary Fig. 7**). Coverage was lower in regions with higher GC content, yet this effect was much less pronounced than for the Illumina sequencing data of the same genomes (**Supplementary Fig. 8**)[12]. This finding was confirmed by analysis of k-mer distributions of MinION and Illumina data (**Supplementary Fig. 9**). We note that while the MinION reads marked as 'fail' show systematic sequencing biases regarding coverage distribution, the quality of the aligned fraction is comparable to the 'pass' reads. We therefore included the

'fail' data of Patient1 that was successfully retrieved through alignment, in all subsequent analyses.

Resolving *de novo* genomic rearrangements with long-read data

Both patients have complex phenotypes involving dysmorphic features and mental retardation, likely caused by their *de novo* complex chromosomal rearrangements, which were karyotypically defined as 46,XX,ins(2;9)(q24.3;p22.1p24.3)dn (Patient1) and 46,XY,t(1;9;5)(complex)dn (Patient2)[16]**.**

We evaluated the performance to detect the breakpoints underlying the complex *de novo* karyotypes of Patient1 and Patient2 using MinION sequencing data, at this medium coverage. Both patients have already been described in recent work, in which Illumina sequencing was used to map the rearrangement breakpoints, as the current gold-standard method for routine genome-wide SV mapping in patient genomes[16,18]. For Patient1, we augmented the previously described data by performing Illumina HiSeq X data for both parents. We performed SV calling with Delly[19] and Manta[20] on the Illumina data from Patient1 and its parents. By integrating SV calls from Delly and Manta and removing calls that were also identified in one or both parents, we obtained a set of 44 putative *de novo* SV breakpoints, 40 of which formed a complex genomic rearrangement, as described previously[16]. These 40 breakpoints were verified by orthogonal breakpoint assays using PCR and MiSeq sequencing (**Supplementary Table 2**). The breakpoints cluster within regions of chromosomes 2, 7, 8 and 9 and are the result of a complex shattering and reassembly process, known as chromothripsis[21,22] (**Figure 1a**).

For Patient2, there were 29 SVs underlying the complex *de novo* karyotype as based on the previously described breakpoint-junctions, which were detected using long-insert mate-pair sequencing and revealed a complex chromothriptic rearrangement involving chromosomes 1, 5 and 9 (**Figure 1a**, **Supplementary Table 2**)[16].

To enable SV detection in nanopore long-read sequencing data, we developed a new bioinformatic tool, NanoSV, tailored to nanopore data. NanoSV uses split-read mapping (obtained from LAST alignment) as a basis for SV discovery (**Methods, Supplementary Fig. 10**), and supports discovery of all defined types of SVs (**Supplementary Fig. 11**). The performance of NanoSV was first evaluated on simulated nanopore long-read data of an artificially rearranged chromosome and benchmarked against two other recently published SV callers, Lumpy[23] and Sniffles[24]. We thus generated 501 simulated rearrangement breakpoints on chromosome 1 and generated equal amounts of simulated nanopore reads of the rearranged, as well as the reference chromosome, using NanoSim[25] (**Methods**). We assessed performance of NanoSV, Lumpy and Sniffles on these simulated data with varying read coverage (1x to 44x). We observed that NanoSV reaches 99.2% recall at 27x coverage (**Supplementary Fig. 12**) with a maximum false positive rate of 1.2% (at 44x coverage). For Lumpy and Sniffles we reached maximum recall rates of 92.4% and 92.6% respectively (at

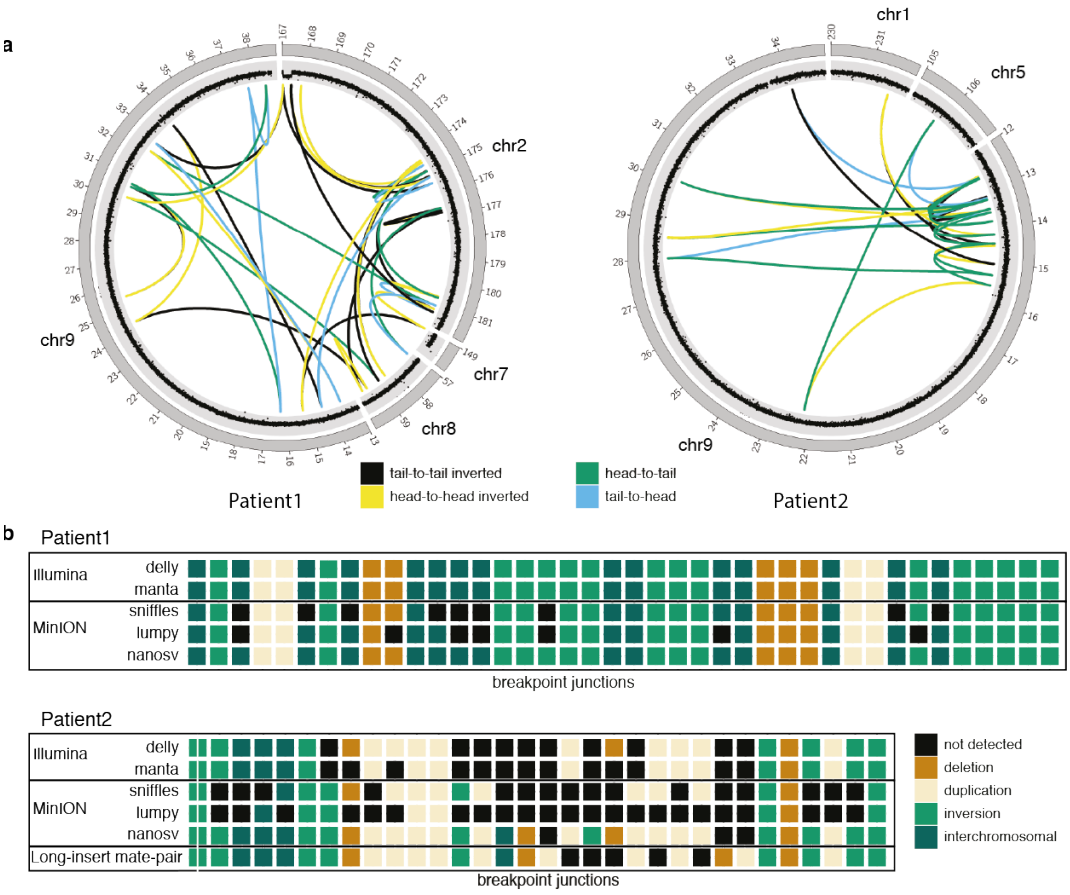


Figure 1: Chromothriptic *de novo* breakpoint junctions of Patient1 and Patient2.
a Circos plots for Patient1 and Patient2 respectively. For Patient1, we took the set of 40 validated *de novo* breakpoint junctions obtained by Illumina whole-genome sequencing of the patient and its parents. For Patient2, we depicted the breakpoint-junctions as published recently[59]. The outer ring of the circos plot shows the chromosomes and the inner ring shows the copy number changes as revealed by FREEC[60] analysis of Illumina whole genome sequencing data for both patients. Colored lines (arcs) indicate breakpoint-junctions. **b** SV genotyping comparison across the chromothriptic breakpoint junctions, between Illumina Hiseq data and MinION data, using various tools tested. The x-axis represents different breakpoint-junctions and the y-axis shows different SV calling methods and datasets. The individual breakpoint-junctions are indicated by colors specifying the type of breakpoint junction.

44x coverage) and maximum false positive rates of 78.8% and 97% respectively.

We went on to apply NanoSV to the complex chromosomal rearrangements data of our patient genomes, and compared results again against Lumpy[23] and Sniffles[24] for the MinION data and Manta[20] and Delly[19] for the corresponding Illumina data. For Patient1, we identified 100% of the 40 validated breakpoint-junctions. Conversely, we discovered 33 (83%) and 31 (78%) of the 40 *de novo* breakpoint junction in the call sets from Lumpy and Sniffles, respectively (**Figure 1b**). For Patient2, NanoSV detected 24 of the 29 previously described breakpoint-junctions. We investigated further why five variants were missed, using Sanger sequencing of PCR products of the respective breakpoint-junctions. We found that two out of the five previously published breakpoint-junctions represent a complex combination of more than two joined segments (**Supplementary Fig. 13**, **Supplementary Table 2**). These short segments were not detectable at the lower resolution of long-insert jumping libraries that were used in the previous analyses compared to the long-read capabilities of MinION

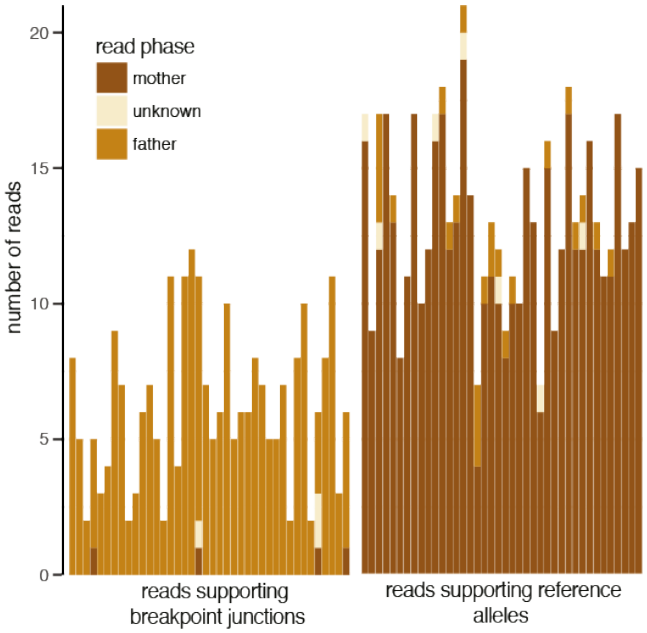


Figure 2: Phasing of chromothripsis breakpoint-junctions.
Phasing of MinION reads overlapping 40 chromothripsis breakpoint-junctions in Patient1. The x-axis displays each of 40 chromothripsis breakpoint-junctions identified in Patient1, stratified by allele (alternative and reference). On the left side only reads supporting the alternative allele are depicted and on the right side reads supporting the reference allele are shown. The y-axis indicates the number of reads supporting each allele, for each of the 40 breakpoint-junctions. Legend colors indicate whether the assigned read phase was paternal, maternal or unknown.

sequencing used here[26]. Based on validation by Sanger sequencing, we retrieved a total of 32 chromothripsis breakpoint-junctions in Patient2 and 29 (91%) of these were detected using NanoSV (**Figure 1b**). For the three remaining breakpoint junctions, insufficient nanopore

read coverage hampered proper genotyping. Nevertheless, for the reads that did span these breakpoints, split read mappings supporting each of these junctions were observed. Lumpy and Sniffles, detected nine (28%) and 16 (50%) of the 32 breakpoints-junctions in the Nanopore data from Patient2, respectively; Manta and Delly detected 19 (59%) and 22 (69%) of the 32 breakpoint-junctions respectively, in the short-insert Illumina data of Patient2. To assess the effect of sequence coverage on breakpoint-junction detection in real data, we subsampled the Patient1 data. This produced an estimate of ~14x for the minimum coverage needed to detect all chromothriptic breakpoint-junctions (**Supplementary Fig. 14**).

Unraveling the long-range structure of chromothripsis

It has been suggested that germline chromothripsis originates from paternal chromosomes[21], but this has previously been inferred from only a few breakpoint-junction sequences or deleted segments. A thorough validation of the conjecture that the origin of chromothripsis is exclusively paternal is lacking. Furthermore, the structure of the chromothripsis rearrangements are typically inferred from the patterns of breakpoint-junctions, under the assumption that the chromothripsis breakpoint-junctions occur on a single haplotype[21,22,27].

We developed a bioinformatic pipeline to augment genome-wide genetic SNP phasing with nanopore read-based phasing of SVs (**Methods**). In a first step we obtained 1.7M heterozygous SNPs from Patient1, that were called from Illumina sequencing data and trio-phased using GATK PBT[28] and Patient1's parents' genotypes. Subsequently, each nanopore read was assigned phase based on a phasing score that takes into account the content and number of overlapping phase-informative SNPs (**Methods**). Per chromothripsis breakpoint-junction, we obtained between two and 11 break-supporting nanopore reads and 85% (195/228) of these overlapped on average of 9.8 phase-informative heterozygous SNPs. We similarly

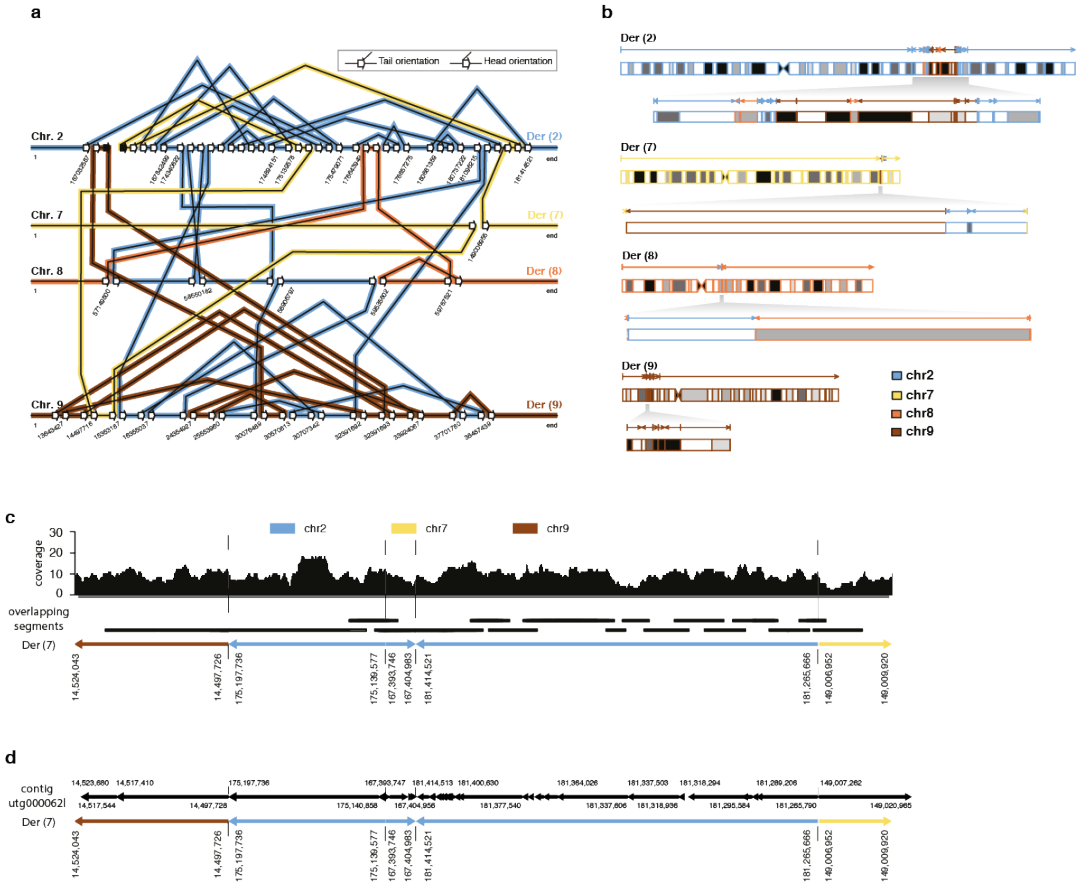


Figure 3: Unraveling long-range chromothripsis structure from the nanopore data.
a Schematic diagram showing the patterns of breakpoint-junctions in Patient1. The human reference genomic regions that are involved in the chromothriptic event are depicted horizontally for each affected chromosome. The slanted lines connecting various reference segments represent breakpoint-junctions. The orientations of breakpoint-junctions are indicated by arrows as shown in the legend. Black (instead of open) arrows indicate the boundaries of a chromosomal deletion resulting from the chromothripsis, whereas open arrows indicate double-stranded DNA breaks. **b** Structure of the chromothriptic derivative chromosomes in Patient1, as inferred from the orientations and order of breakpoint-junctions shown in panel **a**. **c** Reconstruction of a chromothriptic subregion of chromosome 7, involving five chromosomal segments. Overlapping aligned reads originating from Patient1's paternal haplotype were used. Nanopore reads that are instrumental for segment connectivity are indicated by black bars. The coverage track has been generated from all paternal reads mapping to the respective chromosomal segments. The underlying derived chromosome's structure is illustrated on the bottom. **d** Haploid assembly results of the chromothriptic region of Patient1. A 469kb contiguous assembled sequence (utg000062l) spans, through 54 segments that align back to the reference genome, the same chromothripsis subregion illustrated in panel **c**. The assembled contig is fragmented into many (54) aligned segments, as Miniasm does not compute a consensus sequence.

phased the nanopore reads that spanned but did not support the breakpoint junctions (i.e: reference reads). This analysis demonstrated that all 40 *de novo* chromothripsis breakpoint-junctions are of paternal origin (**Figure 2**). A few breakpoint supporting reads point to an origin of some chromothripsis breakpoints on maternal chromosomes. However, these are all reads with three or less overlapping phase-informative SNVs, and likely represent artifacts. These results support earlier hypotheses of a paternal origin of germline chromothripsis, pointing to a breakage and repair process specific for male chromosomes occurring either during spermatogenesis or early zygotic cell divisions[29]. We were further able to reconstruct the affected derivative chromosomes of Patient1 by following the chain(s) of breakpoint-junctions by order and orientation (**Figure 3a-b**). Such a strategy leads to a configuration of four derivative chromosomes for Patient1, each containing one centromere and two telomeric chromosome ends. The chromosomal structure obtained by this procedure matched the observed karyotype (**Supplementary Fig. 15**).

We further sought to investigate the extent to which the derived chromosomal structure could be reconstructed from the MinION sequencing data. We note that a much higher sequencing depth is required in order to accurately reconstruct such large chromothriptic regions through diploid assembly. In order to evaluate the potential of nanopore long-read data to facilitate future analyses, we pre-phased, as described above, all the reads that align within the chromothriptic region (i.e. ~40MB of genomic sequence across four chromosomes) and used only the reads that are known to originate from the paternal haplotype and those that could not be assigned phase (i.e. where the two haplotypes were identical).

We first built contigs by evaluating the read-overlaps from the reference alignment (**Methods**) and obtained contigs that connect between two to five chromothriptic segments, spanning up to 2MB of contiguous DNA sequence (**Figure 3c, Supplementary Fig. 16**). Finally, we used Miniasm[30] to evaluate whether such longer, local haplotype structure can be readily retrieved in a standardized and scalable fashion (**Methods**). The whole 40MB region was assembled into 178 contigs that were subsequently aligned against the human reference genome. We identified three contigs of 241kb, 469kb and 1,217kb in size, each spanning three to five chromothriptic segments. Segment order and orientation in each of the three contigs supports the predicted chromothripsis structure (**Figure 3d, Supplementary Fig. 17**).

Evaluation of SV calling in NA12878 nanopore data

Beyond detection of specific pathogenic SVs, long sequence reads present unique advantages for SV discovery in human genomes, as it has been recently shown from data generated on Pacific Biosciences platforms[12,31]. Here, we assessed whether MinION sequencing data could yield comprehensive and high quality sets of genome-wide SV calls, as well as whether it may yield any novel SVs beyond those found through the Illumina sequencing. To evaluate the performance of NanoSV in a genome-wide analysis, we used the publicly available MinION data for the NA12878 sample[32] and publicly available sets of SV calls, for the same sample, both from short-read Illumina data[33] (referred to as 1KG), as well as from

Pacific Biosciences data[13] (referred to as PB). Based on these calls, we carried out an assessment of both sensitivity, as well as accuracy of our analyses.

We aligned all the fastq MinION R9.4 reads that were generated using normal ONT library preparation, for the NA12878 sample (i.e. we did not include ultra-long read data available for NA12878[34]). We then restricted the analysis to chromosome 1 (as a representative subset) and used NanoSV to produce an initial set of 3,957 genotyped SV calls. Manual inspection of SV candidates within the NA12878 sample as well as within our patients' data showed that MinION sequencing and base calling performs poorly in regions containing homopolymer stretches, which typically lead to a collapse of the whole region into a spurious indel call. This is observed across samples, as well as in MinION sequencing of PCR products (**Supplementary Fig. 18**). Additionally, we noted that SV calling is similarly hampered in tandem repeat regions (**Supplementary Fig. 18**). Based on these observations, we conservatively discarded calls for which both ends of the candidate breakpoint-junction

fall within genomic homopolymer regions or short tandem repeat stretches, resulting in a set of 657 SVs in NA12878. We further filtered for small indels (<40 base pairs) that do not typically result in a split alignment, resulting in a final set of 654 SVs from the chromosome 1 nanopore data of NA12878. We ran Lumpy and Sniffles on the same NA12878 nanopore data and filtered the resulting SV sets, as well as the gold standard truth sets (1KG and PB) using the same criteria, so as to enable an informative comparison. After intersecting the NanoSV callset with 1KG and PB (**Methods**), we observed a sensitivity of 78% (131 out of 168 1KG SVs) and 88% (292 out of 332 PB SVs), respectively. The largest proportion (18/37) of the SV calls that were missed in comparison to 1KG are multiallelic CNVs, which typically require dedicated coverage analysis and are absent from the PB dataset as well. We further missed six indels that were close to the threshold for creating a split read (i.e. 40-50 base-pairs). Identical evaluations of Lumpy and Sniffles revealed sensitivities of 15% (26/168) and 72% (121/168) in the 1KG set of SVs, and 32% (105/332) and 77% (255/332) respectively, in the PB dataset. We note that Lumpy was designed and tested on short-read paired-end sequencing data and we used it on long-read data as the algorithm is conceptually applicable, rather than specifically tailored.

For all subsequent analyses of the NA12878 sample, we considered all the SVs also preset in the 1KG or PB datasets as true positive SV calls (TPs) and any additional SV calls made by NanoSV as false positive calls (FPs). Out of our set of 654 NanoSV calls, 354 overlap with an SV call in the 1KG or PB datasets, resulting in an estimated precision of 54%. Similarly, Lumpy and Sniffles show precisions of 2% and 50% respectively.

To further improve post-calling filtering, we train a random forest model that produces a high confidence set of SVs, with a precision beyond the 54% mark. The features based on which the model is trained are extracted from the aligned sequencing data and are designed to be sequencing read-depth and read-length independent, such that the model is applicable to

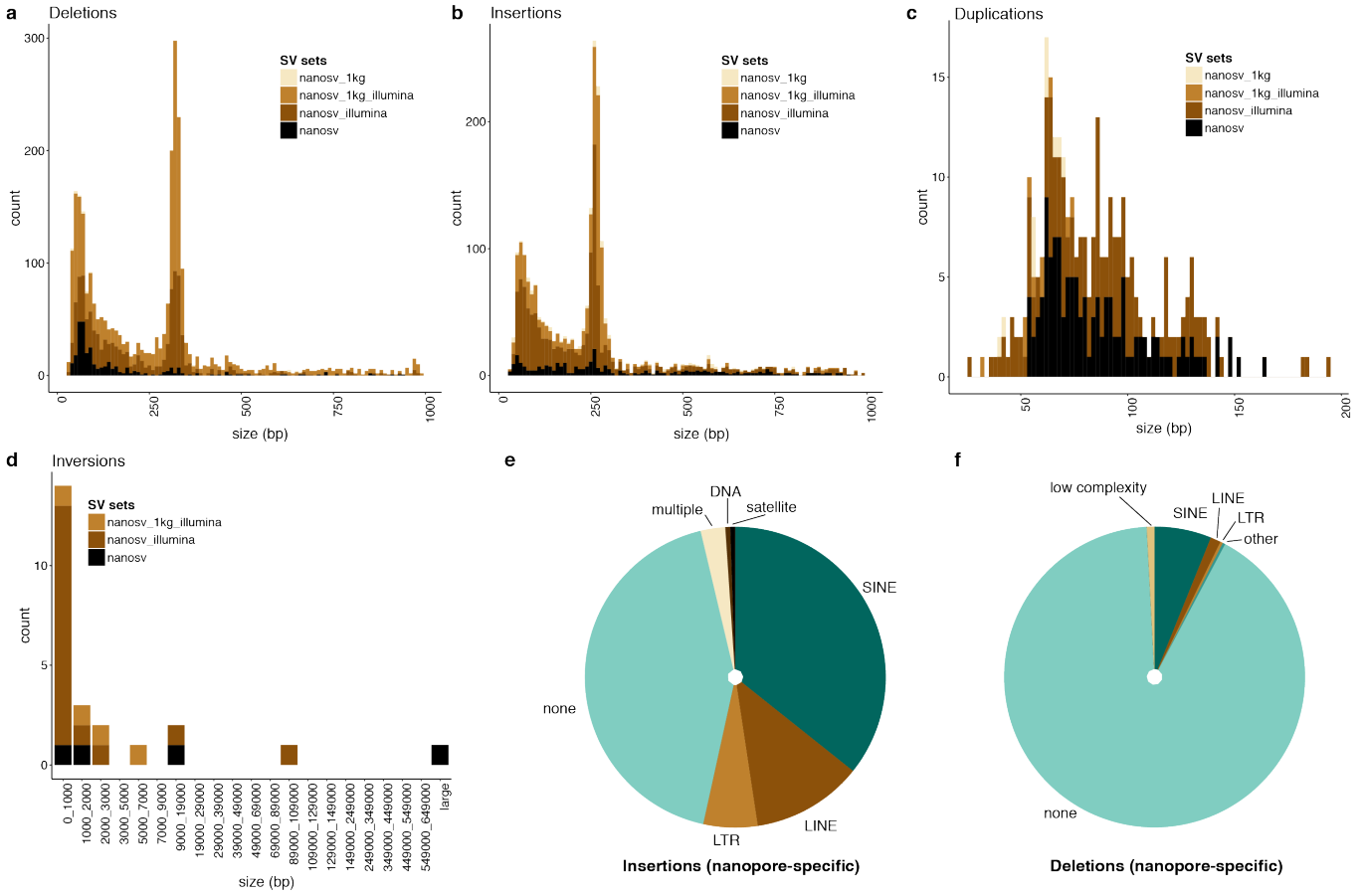


Figure 4: Genome-wide detection of SVs using nanopore sequencing data.
a-d The total amount of high confidence NanoSV SV calls for Patient1 and Patient2 jointly, across different SV size bins and stratified by SV type as follows: **a** deletions, **b** insertions, **c** duplications and **d** inversions. The NanoSV calls were intersected with SV calls from other data sources (Illumina data of Patient1 and Patient2 and 1KG phase 3 sites). For panels **a** and **b**, the x-axis was trimmed to 1000bp for visibility and a small number of variants beyond this size are not displayed in the figure. Similarly, for panel **c** the x-axis was limited to 200bp. **e** The repeat content of nanopore-specific insertions. **f** The repeat content of nanopore-specific deletions. Repeat annotation was obtained from the UCSC repeat masker table (GRCh37).

any MinION sequencing setting (**Methods**). We select as optimal, a random forest model with 82% precision and 75% sensitivity, on our training data (**Supplementary Fig. 19**). The data used for training are the 354 TP and 300 FP NA12878 SV calls described above.

Genome-wide SV discovery from MinION reads

We went on to analyze the whole genome MinION data of Patient1 and Patient2. We ran NanoSV and obtained initial callsets of 36,959 and 36,321 SVs respectively. Filtering for all SVs that do not overlap homopolymers or simple repeats, we obtain 8,578 and 6,791 SVs in Patient1 and Patient2 respectively. Finally, we ran the random forest model trained on the NA12878 data, as described above, and obtained final callsets of **3,271** and **3,345** SVs, for Patient1 and Patient2.

To further evaluate the robustness of our analysis pipeline, we performed multiple rounds of orthogonal validation, on a random sample, spanning all SV classes and size ranges (**Methods**). We obtained validation status for 274 SVs, regardless of the random forest prediction outcome, for Patient1, and 77 SVs predicted as true by the random forest, for Patient2. Based on these sets we obtained precision estimates of 95% and 96% for Patient1 and Patient2 and a sensitivity estimate of 72% for Patient1.

We intersected the SV callsets of Patient1 and Patient2 with calls generated by Lumpy and/or Sniffles. Furthermore, we performed SV calling on the corresponding Illumina data of both patients using six tools (Pindel, Manta, Delly, FREEC, Mobster and GATK HaplotypeCaller) that are commonly used in human genome sequencing studies and which represent different methods to detect SVs (and/or indels) from whole genome short-read Illumina sequencing data, that collectively capture most classes of SVs[19,20,35–37]. An SV is considered to be overlapping with the Illumina dataset if the nanopore data SV call matches an SV call in any of the tools used on the Illumina data (**Methods**). We further considered as overlapping Illumina data (i.e. “detectable” through short-read sequencing) any NanoSV-called variant that can be matched within the 1KG SV and indel sites respectively (**Supplementary Fig. 20**)[38]. Finally, we annotated the SVs from both patients for overlapping repeat elements from the UCSC repeat masker track or the DFAM database (**Methods**).

We identified 14% (944) of SVs in Patient1 and Patient2 nanopore data that were not observed in Illumina data nor are they 1KG variant sites (**Figure 4**). A comparison of the two sets of SV calls shows that nanopore specific SVs are on average located at sites with a higher GC content (i.e. than SVs also genotyped from Illumina data), which are typically hard to sequence with short-read technologies (**Supplementary Fig. 21**). The most frequent class of SVs in the set of 6,616 predicted true-positive SVs are deletions (54%), of which 10% (360) are novel variants detected by nanopore data (**Figure 4a, Supplementary Fig. 22**). We observed that SINE elements were proportionally less abundant among nanopore-specific deletions (6% vs 30% among calls overlapping with Illumina data, **Figure 4f**). The major fraction (91%) of nanopore-specific deletions is not overlapping a repeat feature, most

likely due to our very stringent initial filtering of simple repeats. In fact, the majority (66%) of the nanopore-specific deletions are smaller than 200bp, while only 27% of all deletions are smaller than 200bp. Short deletions are known to be hard to detect using short-read sequencing[39]. Insertions represent the largest fraction among the nanopore-specific set of variants (382, **Figure 4b**). We observed a proportional increase in the amount of LINEs among nanopore-specific insertions compared to calls overlapping Illumina data (12% vs 8%), while SINEs are proportionally underrepresented in nanopore-specific insertions (36% vs 42%) (**Figure 4e, Supplementary Fig. 22**). Finally, 41% of all detected (tandem) duplications (337) are novel variants detected by nanopore data (**Figure 4c**).

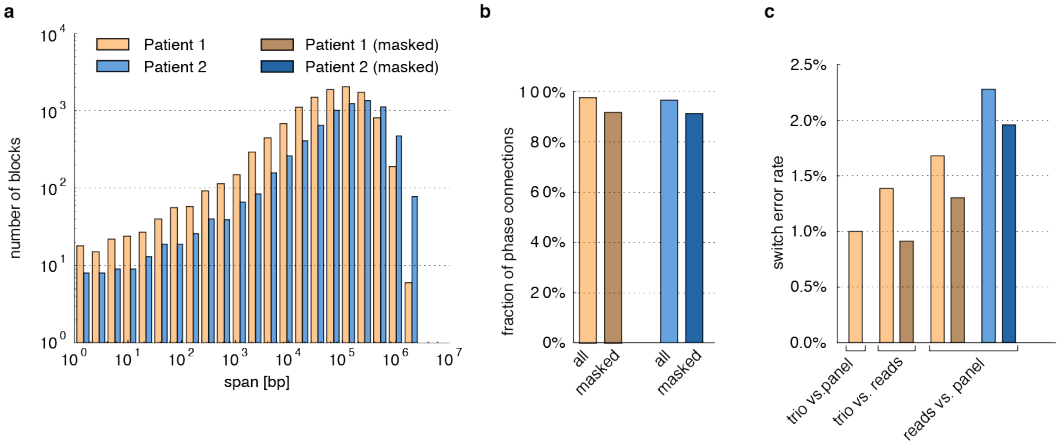


Figure 5. Performance of SNV phasing using nanopore reads.
a Distribution of phased block lengths resulting from read-based phasing by WhatsHap. Patient1 and Patient2 are shown in brown and blue, respectively. **b** Fraction of phase connections (i.e. pairs of consecutive SNVs phased with respect to each other) established in the two patients and with/without masking repeats (light/dark colors). **c** For Patient1, switch error rates of all pairs of trio-based (PBT), population-based (ShapeIt), and read-based (WhatsHap) phasing are shown. For Patient2, where no family data is available, read-based phasing is compared to population-based phasing.

MinION read-based phasing of SNVs

Phasing genetic variation is critical for human disease studies[40,41]. To demonstrate the potential of long-read nanopore sequencing data for direct read-based phasing of genetic variation, we employed WhatsHap, an algorithm that we recently established[42,43]. Using WhatsHap, we phased a set of high-quality genome-wide SNVs from both patients (**Methods**)

and obtained haplo-blocks with N50=126kb for Patient1 and N50=305kb for Patient2 respectively. The distribution of block lengths is shown in **Figure 5a**. We were able to establish 97.5% (96.5%) of all possible phase connections in Patient1 (Patient2), where a phase connection is defined as the relative phase between two consecutive heterozygous SNVs (**Figure 5b**). For Patient1, where Illumina sequencing data was available for the parents, we produced a ground-truth phasing by genetic haplotyping, that is, by using the SNV genotypes and the family relationship[28]. Additionally, we phased both samples using ShapeIt2 and the 1KG phase 3 reference panel[33]. **Figure 5c** shows pairwise comparisons of the obtained haplotypes, with switch error rates of 1.7% and 2.3% when comparing read-based and pop-

ulation-based phasing for Patient1 and Patient2, respectively. We observed a lower switch error rate of 1.4% between trio-based and read-based phasing, which points to a significant amount of switch errors in the population-based phasing (1.0% when comparing trio-based vs. population-based phasing). Therefore, a significant amount of differences between read-based and population-based phasing is most likely due to errors in the population-based phasing. Since MinION reads are especially prone to errors in homopolymer regions, we investigated the effect of excluding all SNVs in such regions from phasing (see **Methods** for a precise definition). This resulted in a decrease in the number of established phase connections from 97.5% to 91.7% for Patient1 and from 96.5% to 91.1% for Patient2 (**Figure 5b**) and a decrease in the switch error rate with respect to the pedigree-based phasing from 1.4% to 0.9% in Patient1, see **Figure 5c**. This shows that switch errors are indeed often found at such homopolymer sites and that masking those sites significantly reduces switch error rates at the expense of only a moderate reduction of phased variants.

MinION read-based phasing of SVs

While structural variation has recently been integrated in larger population genetic reference panels, which enables their imputation for genetic association studies[18,38], building these panels often requires statistical phasing approaches, which drop accuracy for low allele frequency SV sites. Read-based phasing of SVs using long reads will enhance our ability to include SVs in high-quality reference panels, where structural variation is still underrepresented[18].

We apply the same methodology as above (i.e. used for phasing chromothriptic breakpoints) to evaluate genome-wide SV phasing accuracy. A total of 3.8M MinION reads overlapped one or more of the 1.7M genome-wide phase-informative SNPs. As estimated from reads overlapping at least 20 phase-informative SNPs, an average of 85.2% of the SNPs spanned by a read consistently support a particular phase assignment, which is in line with the reported error rate of MinION sequencing data (**Supplementary Fig. 23**). A distinction between reads originating from paternal or maternal haplotypes can be readily made, particularly if reads overlap with multiple phase-informative SNPs (**Supplementary Fig. 24**). We then selected a set of 2,389 heterozygous SVs that overlap between Manta (Illumina) and NanoSV (nanopore) call sets. Each SV was assigned a phase and a phasing quality (**Methods**), by combining information from all phase-informative SNPs falling within the breakpoint-junction supporting reads and reference supporting reads respectively. In this way, we phased 1909 (78.7%) SVs and could assign 971 and 938 to paternal and maternal chromosomes, respectively. For the remainder of 480 SVs, spanning reads did not overlap any phase-informative SNP and therefore a phase could not be assigned to these SVs. Using the SV phasing produced by PBT as ground truth, our long-read based phasing of SVs had an accuracy of 98.5%.

Discussion

In this work, we show the first stand-alone analysis of MinION whole genome sequencing data of human, diploid, patients' genomes, demonstrating the feasibility of long-read sequencing of human genomes on the MinION real-time portable nanopore sequencer. Given the long-read nature of the MinION platform, we focused the analysis on the detection of clinically relevant SVs, a diverse category of genetic variation that is often causal to human genetic disease[44]. Hundreds to thousands of such patients are routinely screened annually for pathogenic SVs in clinical genetic centers, most often by copy number profiling or karyotyping. Although these methods are robust and relatively cost-efficient, they are not capable of mapping small or copy-balanced SVs, nor do they provide base-pair resolution accuracy, or the possibility to resolve complex SVs[45].

Here we show that MinION sequencing provides an attractive alternative approach for genome-wide detection of clinically relevant SVs, which could be implemented as a clinical screening tool for patients with congenital phenotypes, such as intellectual disability[46]. We developed a robust SV discovery and genotyping pipeline that can produce SV calls matching any state of the art precision benchmark (>95% precision). Due to the medium coverage, some intrinsic nanopore sequencing biases and for benchmarking purposes, we employ extremely stringent filtering that results in a good estimated sensitivity (~75%), which can be further increased through higher sequencing depth, or by relaxing our post calling filtering steps.

We were able to extract all known *de novo* breakpoint junctions for Patient1 (**Figure 1**), even at relatively low coverage (16X). The long reads identified additional complexity for several breakpoint-junctions of Patient2. Moreover, 32% (29 vs 22) more chromothripsis breakpoint-junctions were detected with MinION compared to short-insert Illumina sequencing. Our work also confirms previous data that revealed up to 89% novel SVs and indels discovered from PB long-read sequencing of haploid human cells[12,31]. Here, we perform the first standalone genome-wide analysis of SVs in (diploid) patient genomes using nanopore sequencing and show that long-read nanopore data can be readily applied for any research question for which SVs may play a role. We observed that 14% of the high confidence set of SVs in the nanopore data could not be found in matching Illumina sequencing data (despite extensive variant calling using six different variant calling methods as well as comparison to 1KG variant sites). Although this percentage of novel variants is lower than for previously reported PB data, this is primarily due to our conservative SV calling and post-calling filtering steps. Long MinION sequencing reads thus enable a straightforward and homogeneous analysis of SVs, while retaining a very high accuracy in the final set of variants.

Phasing of genotyped SVs - relevant for mapping disease associations - is commonly done using statistical methods or by employing family-relationships among sequenced individuals[18]. We here devised a computational strategy that allowed accurate phasing of SVs

directly from the long nanopore reads using flanking (phased) heterozygous SNPs. Read-based phasing of SVs is advantageous particularly for classes of SVs with a low population frequency and for *de novo* variations. This is exemplified by the evidence provided here for the paternal origin of all *de novo* breakpoint-junctions in Patient1, whereas previous work on chromothripsis has not provided robust support for the parental origin of chromothripsis[21].

If MinION/ONT data quality and throughput increase at a similar pace as we have observed recently (**Supplementary Figs. 1 and 4**), SNV calling and genotyping may be directly performed based on the Nanopore reads. Even though our data are of relatively low coverage, we were already able to obtain a good genotype concordance (96%) with the Illumina based pipeline, for existing SNV calls in Patient1 (not further investigated here). SNV calling combined with accurate genome wide-phasing, as we demonstrated here, will enable simultaneous long-read-only genetic variation discovery and phasing.

Long sequencing reads facilitate personal genome assemblies and emerging new ways of dealing with genetic variation discovery and representation[47,48]. Efforts to obtain full-length haplotype resolved chromosomal sequences are continuously advancing and the combination of multiple long-range sequencing and mapping approaches have recently led to diploid human genome assemblies with contig N50 size of well over 10MB[13,49]. We have not attempted a full human genome assembly using the MinION reads in this work (primarily due to insufficient coverage). However, we were able to separate reads by haplotype, which formed the basis for a reconstruction of the long-range structure of chromothripsis rearrangements. Such information is essential for interpretation of clinical phenotypes[50].

A drawback of current short-read genome sequencing technology is the need for high capital investment, which often leads to sequencing infrastructure being located in dedicated sequencing centers. This is associated with a complex logistic workflow and relatively long turnaround times. Our results show that such limitations can be overcome by the use of the portable MinION sequencing technology. Since the start of this project in April 2016, we have seen a tenfold increase in throughput per MinION sequencing run (**Supplementary Fig. 1**) and an increase in sequencing quality to 90% accuracy for high output 1D runs (450b/s). In practice this means that 10x coverage of the human genome can be reached using 10-15 MinION flowcells at a cost of 5,000$ to 8,000$ within one week of overall sequencing time.

This work demonstrates the potential of long-read, portable sequencing technology for human genomics research and clinical applications. Creating larger catalogues of SVs, in complex repeat regions and segmental duplications, is a particular challenge in the coming years. We foresee that population-scale genome sequencing by ONT or other long-read technology will facilitate such discoveries, leading to further understanding of the role of SVs in the human genome in general and in genetic disease in particular.

Methods

Sample source

The DNA for human genome sequencing in this study was obtained from two patients with congenital abnormalities and the parents of one of them. Informed consent for genome sequencing and publication of the results was obtained from all subjects or their legal representatives. The study was approved by Institutional Review Boards of San Luigi University Hospital and Brigham and Women's Hospital and Massachusetts General Hospital. Both patients have been previously described by Redin et al[16].

DNA extraction

DNA of Patient1 was obtained from either peripheral blood mononuclear cells (PBMCs) derived from blood and from renal epithelial cells obtained from urine. Renal cells were cultured up to eight passages as reported previously[51]. Cells were harvested after reaching confluency by trypsinization with TrypLE Select (Thermo Fisher Scientific) and centrifugation at 250g for five minutes. DNA from the parents was obtained from PBMCs. PBMCs were collected by a ficoll gradient. In brief, the blood was diluted 4x with phosphate buffered saline (PBS). Subsequently 13 mL of Histopaque®-1077 (family 1; Sigma-Aldrich 10771-500ML) was added per 35 mL of diluted blood. The resulting mixture was centrifuged at room temperature for 20 minutes at 900 x g, followed by recovery of the PBMC layer. PBMCs were washed twice using PBS, centrifuged at 750 x g for five minutes and resuspended in PBS with 50% DMSO. For Patient2, DNA was obtained from a lymphoblastoid cell line, which has not been tested for mycoplasma contamination. The cell line was authenticated by whole genome sequencing. DNA extraction from cultured cells and PBMCs was performed using DNAeasy (Qiagen) or Genomic-tip (Qiagen) according to manufacturer's specifications with exclusion of vortexing to maintain DNA integrity.

MinION library preparation and sequencing

Isolated DNA was sheared to ~10-20kb fragments using G-tubes (Covaris). Subsequently, genomic libraries were prepared using the Oxford Nanopore Sequencing kit (SQK-MAP006 for R7 or SQK-NSK007 for R9), the Rapid library prep kit (SQK-RAD001) or the 1D ligation library prep kit SQK-LSK108. A 0.4x (instead of 1x) ampure cleanup was introduced after the FFPE DNA repair and the end-repair steps in the protocol to ensure removal of small DNA fragments. Genomic libraries were sequenced on R7.3, R9 and R9.4 flowcells followed by base-calling using either Metrichor workflows or MinKnow software. For Patient2 we introduced a DNA size selection step prior to library preparation using the Pippin HT system (Sage Science).

Illumina whole genome sequencing

Genomic DNA of the patients and the parents was sheared to 400-500bp fragments using the Covaris. Subsequently, genomic libraries were prepared using the nano library prepa-

ration kit. Genomic libraries were sequenced on an Illumina HiSeq X instrument to a mean coverage depth of ~30x.

Nanopore data mapping

FASTQ files were extracted from base-called MinION sequencing data using Poretools (version 0.6.0)[52]. Subsequently, fastq files were used as input for mapping by LAST (version 744)[17], against the GRCh37 human reference genome. Prior to mapping the full dataset, we used the *last-train* function to optimize alignment scoring parameters using a sample of 1200 nanopore reads. Nanopore sequencing data were also mapped using BWA-MEM with the -x ont2d option, as required by Lumpy and Sniffles. MinION 2D runs can produce 2D sequence reads, i.e. data where both forward and reverse reads of a DNA duplex are collapsed into a single sequence read, which produce three sequences in a fastq file, termed 1D template, 1D complement and 2D. Therefore, we filtered the LAST and BWA BAM files by only retaining one of these three "versions" for each read based on the following order of preference: 2D > 1D template > 1D complement.

Illumina data mapping

Illumina HiSeq X ten data were mapped to the reference genome using BWA-0.7.5a using "BWA-MEM -t 12 -c 100 -M -R". Reads were re-aligned using GATK IndelRealigner[53] and deduplication was performed using Sambamba markdup[54]. Short indels and SNPs were genotyped using GATK HaplotypeCaller, jointly for the Patient1 trio and individually for Patient2.

Analysis of MinION sequencing error rates

We generated a set of 1,064,470 random positions on chromosome 1 and excluded sites that were regarded as polymorphic based on Illumina GATK variant calling. For each of the remaining positions, the mismatch rate, deletion rate and insertion rate were calculated using samtools mpileup. All positions were overlapped with a bed file consisting of homopolymers longer than or equal to 5 base pairs. Additionally, we retrieved the simple repeat track from the UCSC table browser for overlapping all genomic positions with simple repeats. GC content was calculated using a window size of 10bp surrounding each genomic position.

NanoSV algorithm

The NanoSV algorithm developed here (https://github.com/mroosmalen/nanosv) uses LAST BAM files as input. We did not use BWA-MEM alignments as NanoSV input, because the reads are not always split in non-overlapping segments. More precisely, we observed that the following two (oversimplified) CIGAR strings may be produced, for two aligned segments originating from the same sequencing read: 6M4S and 4S6M respectively. While at least some of these alignments are marked as secondary by BWA and can be simply discarded from the analysis, we found that the LAST alignment splitting of the same reads leads, in some cases, to identification of otherwise high confidence structural variants. This observation was not further investigated for the purpose of this project. See **Supplementary Fig. 25**

for a real example extracted from our data.

NanoSV uses clustering of split reads to identify SV breakpoint-junctions. In a first step, all mapped segments of each split read are ordered based on their positions within the originally sequenced read. The aligned read may contain gaps between its aligned segments, i.e. parts of the read that do not align anywhere on the reference genome, for example due to insertions (**Supplementary Fig. 10**, **Supplementary Fig. 11**) or simply due to low quality sequencing.

Let tuple $x = (c,s,e,k)$ describe an aligned sequence segment, where the chromosome and genomic start and end coordinates of the segment are specified by c, s and e respectively, and the mapping orientation by k in {+,-}. The coordinates s and e represent the start (lowest) and end (highest) coordinate of the mapped segment on the reference genome. Now, read R_i can be described in terms of the ordered list of aligned segments and alignment gaps $X_i = [u_1, x_1, u_2, x_2, u_3, \ldots, x_N, u_{N+1}]$, where the ordering is determined based on their occurrence in the read, u is the gap (i.e.: unaligned sequence preceding segment x) and N is the total number of aligned segments for read R. Alignment gaps are defined as read segments that are either unaligned or segments that fail to reach the mapping quality threshold Q_1 (default: 20). The size of an unaligned segment is denoted as |u|, and can be zero in case two adjacent segments align successfully.

Any two consecutive aligned segments $[x_n, u_n, x_{n+1}]$ in a read define a candidate breakpoint-junction.

We further aggregate evidence from different reads supporting the same candidate breakpoint-junction. This is achieved by clustering all candidate breakpoint-junctions that have the same orientation and have start and end coordinates that are in close genomic proximity. In order to facilitate clustering of reads that cover the same breakpoint-junction but that map to opposite strands of the reference human genome, order and orientation of the aligned segments is reverse complemented if for the genomic coordinates $\{p,q\}$ mapping to the two closest bases of segments x_n and x_{n+1}, respectively, within a given sequence read R_x, at least one of the following conditions is met:

1. p and q are on the same chromosome and $p - q > 0$

2. p and q are on different chromosomes and p has a higher chromosome number

The clustering is initialized by assigning each pair of consecutive aligned segments $[x_n, u_n, x_{n+1}]$ to a separate cluster. The resulting clusters are then recursively merged. Any two clusters (C_x and C_y) are merged if and only if, there exists a candidate breakpoint-junction tuple (x_n, x_{n+1}) that belongs to cluster C_x and a candidate breakpoint-junction tuple (y_m, y_{m+1}) that belongs to cluster C_y, such that the following conditions are met:

$x_n(c) = y_m(c)$ (segments n,m map to same chromosome)

$x_{n+1}(c) = y_{m+1}(c)$ (segments $n+1,m+1$ map to same chromosome)

$x_n(k) = y_m(k)$ (segments n,m have same orientation)

$x_{n+1}(k) = y_{m+1}(k)$ (segments $n+1,m+1$ have same orientation)

$\min_{x,y}(|x_n(e) - y_m(e)|) \leq d$ if $x_n(k) = +$ (n,m segment-ends are in close proximity)

$\min_{x,y}(|x_n(s) - y_m(s)|) \leq d$ if $x_n(k) = -$ (n,m segment-starts are in close proximity)

$\min_{x,y}(|x_{n+1}(s) - y_{m+1}(s)|) \leq d$ if $x_{n+1}(k) = +$ ($n+1,m+1$ segment-starts are in close proximity)

$\min_{x,y}(|x_{n+1}(e) - y_{m+1}(e)|) \leq d$ if $x_{n+1}(k) = -$ ($n+1,m+1$ segment-ends are in close proximity)

Where d is the threshold that we set for the maximum distance between the alignment coordinates of two segments if they are to be considered as supporting the same breakpoint-junction (default: 10 base-pairs). Iterative clustering continues until no more clusters can be merged. Each final cluster represents one candidate SV, which is described by tuple $b = (c_1,c_2,p_1,p_2,k_1,k_2,g)$, with p_1, p_2 the medians of the start and end coordinates of all candidate breakpoint junctions contained in the cluster, c_1, c_2 the chromosomes associated to these coordinates and k_1, k_2 the orientation of the breakpoint-junction. Finally, the gap size g denotes the median size of the unaligned segments u_n, between the two consecutive aligned segments x_n and x_{n+1} of all the tuples within the respective cluster.

A true SV is called when a candidate SV is supported by more than T reads (default: 2). Moreover, SVs with median mapping quality of the supporting reads not exceeding Q_2 are still reported, but flagged as “MapQual” in the VCF FILTER field. SV-types can be determined from tuple b. Breakpoint-junctions where c_1 and c_2 point to different chromosomes are considered interchromosomal SVs (e.g. chromosomal translocations), which can have one of four possible orientations (3’to3’, 3’to5’, 5’to5’, 5’to3’). Similarly, breakpoint-junctions where c_1 and c_2 point to the same chromosome are intrachromosomal SVs, which can have one of four possible orientations (inversion type=3’to3’ or 5’to5’, deletion/insertion type=3’to5’, tandem duplication type=5’to3’). Insertions and deletions are discerned based on the relation between the gap size, g, and the reference-length $l = |p_1\text{-}p_2|$, where an insertion is called if $g > l$ and a deletion is called when $g <= l$ (**Supplementary Fig. 11**).

We only consider two possible alleles for each SV candidate (present = ALT/absent = REF). The reads supporting the alternative allele contain the segments constituting the breakpoint-junction cluster. We consider as supportive of the reference allele all reads for which there is an aligned segment crossing one of the ends of the breakpoint junction (or both).

More formally, a read is defined as crossing a breakpoint if it contains at least one aligned segment x_n for which holds: $(p_1 - x_n(s) > 100$ and $x_n(e) - p_1 > 100)$ or $(p_2 - x_n(s) > 100$ and $x_n(e) - p_2 > 100)$. Reads not supporting the reference allele according to this definition are ignored. SV genotypes (homozygous alternative, heterozygous, homozygous reference, not-called) are assigned based on a Bayesian likelihood similar to the one used (and formally defined) by the SVTyper[23]. SV calls are reported in VCF format following the VCF standards as maintained by samtools specifications[55]. To facilitate reporting of complex SV types, such as inversions or reciprocal translocations, individual breakpoint-junctions that bridge the same chromosomal regions, but are opposite in orientation (e.g. 3'to3' and 5'to5' for inversions), are linked using an identifier.

Nanopore data SV calling

We run NanoSV on the MinION data of each patient using the default parameters : "-t 8 -s 10 -p 0.70 -m 20 -d 10 -c 2 -f 100 -u 20 -r 300 -w 1000 -n 2 -q 80 -i 0.80 -g 100 -y 20". We discarded all sites where the alternative allele count was 0 in the resulting genotype (i.e.: HOM_REF) and further filtered the resulting call sets for SVs tagged as "Cluster". The "Cluster" VCF INFO-field tag is added to all SV calls which lie inside a (default) 1000 base-pair region containing three SVs or more. These clusters of SVs are most likely either sequencing errors or located in highly repetitive and/or decoy regions of the human reference. We used Lumpy[23] and Sniffles[24] (specifically designed for ONT and PB data) to call SVs in both samples using BWA-MEM alignments (instead of LAST alignment, as per requirement of the respective callers) of the same data and settings that match our own (liberal) NanoSV settings as closely as possible, as follows. For Lumpy: "-mw 2 -tt 0 -e", requiring that at least one read supports each candidate breakpoint and clustering breakpoints within 10 base-pairs (back_distance=10). For Sniffles: "-s 2 --max_num_splits 10 -c 0 -d 10"[24]. At the time of our analysis SVTyper was not supporting nanopore reads (i.e. it required paired-end reads), therefore we considered the Lumpy, ungenotyped, SV candidate sites as final calls for all subsequent analyses/comparisons. This implies that all sensitivity estimates for Lumpy are upper bounds and that precision estimates are most likely under-estimated.

Simulation of nanopore sequencing reads and genomic structural variants

We took the human reference chromosome 1 sequence (GRCh37) and introduced 501 breakpoints, followed by random reshuffling of chromosomal segments into a new sequence. The breakpoints were introduced in a 20MB region (chr1:51707947-71707947), similar in size as a typical chromothripsis region. Subsequently, NanoSim[25] was used to simulate nanopore reads. We used 400 random reads from Patient2 to build the error profiles for the simulated reads. Simulated read sets were generated for both the reshuffled chromosome 1 and the reference chromosome 1, in order to be able to introduce heterozygous structural variations in the simulated read data. Simulated reads were mapped using LAST and BWA, followed by SV calling using NanoSV, Lumpy and Sniffles, as described above. We performed down-

sampling of the reads to evaluate the effect of coverage on simulated breakpoint detection. Four of the randomly generated SV breakpoints produced small events (~40-50 basepairs), for which the LAST alignment does not result in a split read; these events were missed by NanoSV, regardless of the coverage used.

Random Forest variant filtering model

We trained a random forest (RF) model that we subsequently used to filter out false positive SV calls from our Nanopore data, such that we obtained a high precision set of variants for downstream analysis. The training data for our model consists of 354 true positive (TPs) SVs and 300 false positives (FPs). These 654 training data SVs are the NA12878 SV genotypes described in **Results**, where any SV overlapping any of 1KG or PB datasets is considered a TP and all other SVs are considered FPs. Our training data is conservative in the sense that while all SVs considered TPs are based on previously curated data, we denote false positive SVs solely by overlap with other (different data) datasets (i.e.: we performed no validation on NA12878 to evaluate if all/most novel variants that we find are indeed FPs).

We supply the following features to the RF model (where side1 and side2 refer to the lowest and highest genomic coordinates of a breakpoint-junction, respectively; and the Mean Decrease Gini for each feature - proportional to the efficiency of splits in the model based on the respective feature - following in bold):

- **Mapq1:** average mapping quality over all reads supporting side1 of the breakpoint junction (5.78)
- **Mapq2:** average mapping quality over all reads supporting side2 of the breakpoint junction (4.39)
- **Pid1:** average percent identity (i.e.: to the reference) over all reads supporting side1 of the breakpoint junction (27.10)
- **Pid2:** average percent identity (i.e.: to the reference) over all reads supporting side2 of the breakpoint junction (31.73)
- **Cipos1:** genomic distance from the median start position of the SV to the lower bound of its associated confidence interval (21.22)
- **Cipos2:** genomic distance from the median start position of the SV to the upper bound of its associated confidence interval (i.e.: confidence interval width = cipos1 + cipos2) (17.08)
- **Plength1:** average proportion of the aligned segment (i.e.: relative to the entire read length), across all segments supporting side1 of the breakpoint junction (32.02)
- **Plength2:** average proportion of the aligned segment (i.e.: relative to the entire read length), across all segments supporting side2 of the breakpoint junction (43.44)
- **Ciend1:** genomic distance from the (median) end position of the SV to the lower bound of its associated confidence interval (17.73)

- **Ciend2:** genomic distance from the (median) end position of the SV to the upper bound of its associated confidence interval (i.e.: confidence interval width = ciend1 + ciend2) (26.72)
- **TotalCovNorm:** depth coverage summed across both ends of the breakpoint junction, divided by the average depth of coverage across the sample (13.44)
- **Vaf:** percentage of the reads spanning either end of the breakpoint junction that support the variant allele (i.e.: the presence of a breakpoint junction) (82.81)

We found that most of our selected features were statistically significantly different across the sets of TPs and FPs respectively (**Supplementary Fig. 26**), thus informative to our model.

The precision-recall curve of the model, and its 95% confidence interval, displayed in **Supplementary Fig. 19** is derived from 100 bootstrapping runs where the whole training data was split into 90%-10% train-test subsets. The optimal operating point was chosen at 82% precision and 75% recall, and the final model used was trained again using all the training data available.

We compared the distributions of the random forest features used, across the training data of NA12878 and the test data of Patient1 and Patient2 respectively, and observed no statistically significant difference (**Supplementary Fig. 27**), We note that some difference should in fact be expected in the Patient1 comparison, in the average read percent identity related features used (pid1 and pid2), given the different chemistries and nanopores used to generated these data.

Illumina data SV calling

SV calling for Illumina data was done using Manta[20], Delly[19], FREEC[36], Mobster[35] and Pindel[37]. For Manta we used version 0.29.5 with standard settings, for Delly we used version 0.7.2 with "-q 1 -s 9 -m 13 -u 5", for FREEC we used version 7.2 with window=1000, for Mobster we used version 0.1.6 with standard settings (Mobster properties template), for Pindel we used version v0.2.5b8 with standard settings and excluding regions represented by the UCSC GRCh37 gap table using the -c option. Homozygous reference calls (genotype = 0/0) were omitted from the call sets for each of these tools.

PCR, primer design and SV validations

Primers for breakpoint-junction validation were designed using Primer3 software[56]. Breakpoint-junction coordinates and orientations were used as input for primer design. Amplicon sizes varied between 500-1000bp. PCR reactions were performed using AmpliTaq gold (Thermo Scientific) under standard cycling conditions. PCR products were sequenced using MiSeq (TruSeq library preparation, Illumina), Sanger sequencing (Macrogen) or MinION sequencing (2D library preparation and sequencing).

We perform extensive and heterogeneous validation on the SV calls of Patient1, in order to obtain a thorough and reliable characterization of our dataset and an informative comparison to standard approaches. We first randomly selected 384 NanoSV candidate calls (uniformly distributed across the observed size-range of SVs) from the call set of Patient1 and performed validation with Illumina MiSeq. We further selected 400 candidate calls (uniformly distributed across the observed size-range of SVs) exclusively from the nanopore specific SV calls and validated them. Deep coverage MinION sequencing was used for this second round of validation, under the assumption that a long-read accessible only set of variants would be less likely to validate using the short-read Illumina sequencing. A third round of validation was performed, also by MinION deep coverage sequencing, on a set of 192 non-random variants; namely, 96 variants were expected to be true positive SV calls and 96 false positive SV calls, as of an initial attempt to build a discriminative model. Upon inspection of these validation results, SVs falling within homopolymer stretches (see above) and/or short tandem repeats (UCSC tandem repeat table) were considered unreliably genotyped (i.e. even in the validation data) and were subsequently discarded from the dataset altogether (see main text - **Results**).

All of the above three rounds of validation are thus restricted to the sites that fall outside homopolymers and/or short tandem repeats and SVs for which we did not obtain a specific PCR product are discarded. This is the subset that is referred to as validation data throughout the text, when evaluating precision and it consists of 274 SVs (185 true positives and 89 false positives).

Finally, we selected 14 large inverted breakpoint-junctions (>1000bp) plus 82 randomly selected SV candidates, all of which were predicted as true by the random forest model from Patient2. We performed PCR for each of these 96 SV breakpoint-junctions and sequenced the resulting amplicons using deep coverage MinION sequencing. We were able to successfully produce and sequence amplicons for 77 of the variants, and 74 of them validated. Out of the 14 large inverted breakpoint-junctions, eight produced a PCR product and seven of these were validated as true.

A structural variant was considered validated as a true positive if there exists an SV call, in the validation SV call set, that overlaps (in the meaning described below) the original SV validation candidate. The validation SV call set is produced similarly to the initial analysis, where Manta is used for genotyping SVs in the MiSeq validation data and NanoSV is used for the nanopore data respectively, with the note that deep coverage (i.e.: ~1,000 for MiSeq and MinION runs) enables accurate genotyping.

Annotation of repeat elements

All deletions from our NanoSV callset were annotated as overlapping a repeat element, if the sequence of the variant overlaps an entry of the repeat masker table of UCSC (GRCh37). For all NanoSV variants reported as insertions, we extracted the inserted sequence as iden-

tified in each supporting nanopore read, used Muscle[57] to generate a multiple sequence alignment for all the sequences supporting the same insertion and obtained a consensus sequence by a simple majority vote. Subsequently we interrogated the DFAM[58] database and annotated all insertions which contained sequence of a repeat element.

Calculating overlap between SV datasets

To calculate the intersection between SV call sets, we considered each SV call as a set of breakpoint-junction start and end coordinates s and e, and orientation k. For any SV call i, each breakpoint-junction coordinate (s_i and e_i) is the median of an associated confidence interval, ($s_{i,l}$,$s_{i,h}$) and ($e_{i,l}$,$e_{i,h}$) respectively, as derived from the evidence cluster C_i. SV calls i and j are overlapping if the confidence intervals of their start and end coordinates are closer together than 101bp. For SVs smaller than 1000bp (excluding insertions), we additionally required that SVs overlap each other with a reciprocal overlap of at least 70%, otherwise, considering the 100 base-pair margin that we use when comparing breakpoint junction borders, different SVs that happen to be in genomic close proximity may, incorrectly, be considered the same event .

GC bias

The GC content (i.e. percentage of guanine or cytosine bases within a certain DNA sequence) was computed for 100,000 5kb intervals of the reference genome (build GRCH37). These intervals were chosen such that they do not overlap sequencing gaps in the reference, as defined in the UCSC table browser, including telomers, centromeres and other gaps. The average depth of coverage across each interval was then computed from the HiSeq alignment data and the MinION alignment data respectively (stratified by sequence reads tagged as “passed” and “failed” by the Metrichor basecalling for Patient1). The GC content was binned into 30 uniformly spread bins, between the minimum and the maximum GC content derived from the data. Six GC-content bins were discarded - i.e. those where GC-content < 0.26 or GC-content > 0.66 - as too few sampled intervals fell within these bins and a coverage distribution cannot be robustly estimated (i.e.: 1 - 18 intervals per bin, **Supplementary Fig. 6**).

A linear regression model with average coverage as the dependent variable and GC-content as the independent variable was fitted, in order to quantify the GC bias of the two sequencing technologies, respectively. The average coverage values were normalized (0 mean, 1 variance) for Illumina and MinION data respectively, because of the different sequencing average depth of coverage, such that the regression coefficients for the two technologies be comparable (i.e. the resulting regression coefficients express the number of standard deviations that the coverage varies, per GC content percentage).

Genetic phasing of variants from Illumina sequencing data

We used the Illumina whole genome sequencing data of Patient1 and both its parents to

obtain a high confidence set of phased genotypes (including SNVs, short indels and SVs), against which we subsequently evaluated the nanopore data analysis. We used GATK PhaseByTransmission (PBT)[28] to correct genotypes based on trio information and to obtain deterministic phasing for most loci. PBT settings were: "-prior 0.000001 -useAF GT -af_cap 0.0001". The PBT-phased SNVs were used to evaluate the genome-wide read-backed phasing from nanopore data as well as for phasing the nanopore reads and the PBT-phased SVs were used to evaluate the nanopore read-backed phasing of the SVs (i.e.: evaluation was limited to the SVs detected in both nanopore and Illumina data). PBT was run with a *de novo* mutation prior of 10e-6 and supplied with the population allele frequencies of 1KG phase 3 European population.

Nanopore read-based phasing of SNVs using Whatshap

For both patients, all bi-allelic heterozygous SNVs were phased from the aligned MinION reads using WhatsHap[42,43] (version 0.13+21.g45bd7f8, command line "whatshap phase --distrust-genotypes --reference <ref.fasta> <variants.vcf> <reads.bam>"), i.e., with realignment mode enabled. That is, reads were realigned against reference and alternative alleles at variant sites, which is critical for phasing performance of noisy long reads[43]. For comparison purposes, we used SNV genotypes to obtain a population-based phasing with respect to the 1KG phase 3[33] reference panel by running ShapeIt with default parameters. We excluded from the comparison all variants that fell within homopolymer runs longer or equal to five base-pairs, due to both genotyping accuracy, but mostly because of the known drop in sequencing accuracy of MinION reads for longer homopolymer sequences. The homopolymer bed-track used was computed genome-wide, incorporating a one base-pair border around the homopolymer, such that relatively frequent sequences of the form "XXXXXYZZZZZ" be merged into one homopolymer region for the final result.

Phasing of nanopore reads and SVs

Individual nanopore reads from Patient1 were phased using a set of 1.7M heterozygous SNVs that were genetically phased by GATK PBT[28]. Individual nanopore reads were phased using the genetically phased SNVs by determining the basecall and corresponding base-quality at each SNV position within each read. Let *b(i)* and *q(i)* be the basecall and associated quality value for some SNV *i* in some read under evaluation. Further let *BM(i)* and *BP(i)* be the maternal and paternal alleles respectively (i.e.: as phased by PBT), for SNV *i*. The information from all SNVs spanned by a read is then aggregated and the likelihood that read r originates from the paternal or the maternal haplotype respectively is computed:

$$Lp(r) = \prod_{i=1}^{n} P(\, b(i) | BP(i) \,)$$

$$Lm(r) = \prod_{i=1}^{n} P(\, b(i) | BM(i) \,)$$

Where *n* is the total number of SNVs that read *r* overlaps and

$$P(\, b(i)|base\,) = 1 - \, 10^{\frac{-\,q(i)}{10}}, \qquad if\ b(i) = \, base$$
$$P(\, b(i)|base\,) = 10^{\frac{-\,q(i)}{10}}, \qquad if\ b(i) \neq \, base$$

Is the probability that a read supports a specific phased allele at a SNV. The likelihoods that the SV resides on the paternal or the maternal haplotype respectively are then computed:

$$Lp(SV\,) = \prod_{r=1}^{R_{SV}} L_p(r)$$
$$Lm(SV\,) = \prod_{r=1}^{R_{SV}} L_m(r)$$

Where *Rsv* denotes the set of all reads supporting the breakpoint junction. The two likelihood scores are then transformed to probabilities (i.e.: normalized to sum up to 1) and phase for the set of breakpoint-junction supporting reads is assigned as indicated by the highest likelihood score. Phase is assigned identically to the set of reference-supporting reads spanning the breakpoint junction.

An SV is then considered phased if the two phases, for the set of breakpoint supporting reads and reference supporting reads respectively, correspond to different parental haplotypes and the (phred scaled) phasing posterior quality is defined as:

$$PP = \, -10 * \log_{10}(\, \max\Big(L_p(SV), L_m(SV)\Big) * \max\Big(\, L_p(REF), L_m(REF)\Big)\,)$$

Construction of chromothripsis structure using reference-based MinION read overlap

To obtain evidence for the long-range structure of the chromothripsis breakpoint-junctions in Patient1, we first extracted the set of (aligned) nanopore reads that span the chromothripsis regions on chromosomes 1, 7, 8 and 9. Separation of reads by phase was done as described above. The mapped segments were ordered by left genomic mapping coordinate of each segment to produce an ordered list of segments L={s(1), s(2), …, s(n)}, from all reads jointly. We then built an oriented graph, where each aligned segment in L is initially a node and nodes were iteratively merged as follows: Let i and j represent the start (left) and end (right) coordinate of each segment (i.e.: or, subsequently, nodes). In order for *s(x)* to be merged into a node *s(y)* there must exist at least two other segments *s(z)* and *s(t)* supporting the same node, such that ($s(z)_j$-$s(x)_i$) >= 20bp and ($s(t)_j$-$s(x)_i$) >= 20bp. Edges were then added, to the final, reduced set of nodes, by evaluating each read's segmentation across supported nodes. Namely, an edge is added between any two nodes for which there is a read such that one segment of the read supports one node and another segment of the same read supports the other node. Each edge was then weighted by the number of reads supporting that connection.

Finally, contigs were built by evaluating all maximal length paths through the graph, where only edges of weight at least two are considered and branching is resolved in a greedy way, by selecting the maximum weight path at each step.

Using the above algorithm, individual breakpoint junctions were connected together, providing support for the order of the joined segments in the chromothripsis chromosomes of Patient1.

Assembly of MinION sequencing data

Nanopore reads of Patient1 were separated into three bins by phase, as described above. The reads that were assigned a paternal phase and the unphased reads were used as input for *de novo* assembly using Miniasm[30], with settings: minimap -S -w 5 -L 100 -r 500 -m 0 and miniasm -c 1 -m 100 -h 20000 -s 100 -r 1,0 -F 1. The mentioned parameters were found to produce the longest contigs from our data. These Miniasm contigs were subsequently aligned to the human reference genome (GRCh37) using LAST, with settings: -s 2 -T 0 -Q 0 -p [*last_parameters*]. The *last_parameters* were obtained as described above (i.e.: the same used for aligning the initial MinION data of Patient1 and Patient2). LAST aligments (SAM format) were processed by custom scripts to evaluate the presence of chromothripsis segments from Patient1 based on chromosomal coordinate overlap.

Data availability

Illumina and nanopore whole genome sequencing data used in this study can be accessed through the European Genome-phenome Archive under accession number

EGAS00001002333.

Competing interest statement

WK and JdR have received financial compensation for travel and accommodation expenses to speak at an Oxford Nanopore Technologies-organised meeting.

Acknowledgements

We thank the Bioinformatics Expertise Core of the UMC Utrecht for setting up part of the computational infrastructure and software to analyze Nanopore sequencing data. This work was supported by funds from the Utrecht University to implement a single-molecule sequencing facility. MCS is supported by VIDI grant 91712354 from the Dutch organization for scientific research (NWO-ZONMW). MET was supported by grants from the National Institutes of Health (GM061354 and HD081256). GP is supported by a fellowship of the Associazione San Luigi Gonzaga. We thank Eleonora Di Gregorio, Alfredo Brusco and Elisa Savin for their contribution to the identification of the complex chromosomal rearrangement in Patient1.

Author contributions

SM, GP, DG, GM, JK, MET provided access to patient cells and DNA. IR and WPK generated MinION sequencing data. EB, JK and EC provided Illumina sequencing data. MCS, MJR, MN, JL, JEVI, TM, JR and WPK performed nanopore data analysis. MCS, WPK, TM and JR wrote the manuscript. All authors contributed to the final version of the manuscript.

References

1. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Unlocking Mendelian disease using exome sequencing. *Genome Biol.* **12,** 228 (2011).

2. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17,** 507–522 (2016).

3. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).

4. Pierce, B. L. & Ahsan, H. Clinical assessment incorporating a personal genome. *Lancet* **376,** 869; author reply 869–70 (2010).

5. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).

6. Stankiewicz, P., Paweł, S. & Lupski, J. R. Structural Variation in the Human Genome and its Role in Disease. *Annu. Rev. Med.* **61,** 437–455 (2010).

7. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34,** 518–524 (2016).

8. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17,** 239 (2016).

9. Quick, J., Joshua, Q., Quinlan, A. R. & Loman, N. J. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *Gigascience* **3,** (2014).

10. Goodwin, S. *et al.* Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* **25,** 1750–1756 (2015).

11. English, A. C. *et al.* Assessing structural variation in a personal genome—towards a human reference diploid genome. *BMC Genomics* **16,** (2015).

12. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517,** 608–611 (2015).

13. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12,** 780–786 (2015).

14. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30,** 771–776 (2012).

15. Mostovoy, Y. *et al.* A hybrid approach for de novo human genome sequence assembly and phasing. *Nat. Methods* **13,** 587–590 (2016).

16. Redin, C. *et al.* The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet.* (2016). doi:10.1038/ng.3720

17. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21,** 487–493 (2011).

18. Hehir-Kwa, J. Y. *et al.* A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* **7,** 12989 (2016).

19. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).

20. Chen, X. *et al. Manta: Rapid detection of structural variants and indels for clinical sequencing applications*. (2015). doi:10.1101/024232

21. Kloosterman, W. P. *et al.* Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *Cell Rep.* **1,** 648–655 (2012).

22. Chiang, C. *et al.* Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat. Genet.* **44,** 390–7, S1 (2012).

23. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15,** R84 (2014).

24. fritzsedlazeck. fritzsedlazeck/Sniffles. *GitHub* Available at: https://github.com/fritzsedlazeck/Sniffles. (Accessed: 21st October 2016)

25. Yang, C., Chu, J., Warren, R. L. & Birol, I. NanoSim: nanopore sequence read simulator based on statistical characterization. *Gigascience* (2017). doi:10.1093/gigascience/gix010

26. Redin, C. *et al.* The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet.* (2016). doi:10.1038/ng.3720

27. de Pagter, M. S. *et al.* Chromothripsis in healthy individuals affects multiple protein-coding genes and can result in severe congenital abnormalities in offspring. *Am. J. Hum. Genet.* **96,** 651–656 (2015).

28. Francioli, L. C. *et al.* A framework for the detection of de novo mutations in family-based sequencing data. *Eur. J. Hum. Genet.* (2016). doi:10.1038/ejhg.2016.147

29. Kloosterman, W. P. *et al.* Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *Cell Rep.* **1,** 648–655 (2012).

30. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32,** 2103–2110 (2016).

31. Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* (2016). doi:10.1101/gr.214007.116

32. nanopore-wgs-consortium. nanopore-wgs-consortium/NA12878. *GitHub* Available at: https://github.com/nanopore-wgs-consortium/NA12878. (Accessed: 19th June 2017)

33. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

34. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. (2017). doi:10.1101/128835

35. Thung, D. T. *et al.* Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biol.* **15,** (2014).

36. Boeva, V. *et al.* Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27,** 268–269 (2011).

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

38. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).

39. Marschall, T., Hajirasouliha, I. & Schönhuth, A. MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics* **29,** 3143–3150 (2013).

40. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nat. Rev. Genet.* **12,** 215–223 (2011).

41. Corradin, O. *et al.* Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. *Nat. Genet.* **48,** 1313–1320 (2016).

42. Patterson, M. *et al.* WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol.* **22,** 498–509 (2015).

43. Martin, M. *et al.* WhatsHap: fast and accurate read-based phasing. (2016). doi:10.1101/085050

44. Lupski, J. R. Structural variation mutagenesis of the human genome: Impact on disease and evolution. *Environ. Mol. Mutagen.* **56,** 419–436 (2015).

45. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12,** 363–376 (2011).

46. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511,** 344–347 (2014).

47. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* (2016). doi:10.1093/bib/bbw089

48. Church, D. M. *et al.* Extending reference assembly models. *Genome Biol.* **16,** 13 (2015).

49. Seo, J.-S. *et al.* De novo assembly and phasing of a Korean human genome. *Nature* **538,** 243–247 (2016).

50. Middelkamp, S. *et al.* Molecular dissection of germline chromothripsis in a developmental context using patient-derived iPS cells. *Genome Med.* **9,** 9 (2017).

51. Zhou, T. *et al.* Generation of human induced pluripotent stem cells from urine samples. *Nat. Protoc.* **7,** 2080–2089 (2012).

52. Loman, N. J. & Quinlan, A. R. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **30,** 3399–3401 (2014).

53. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

54. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31,** 2032–2034 (2015).

55. [No title]. Available at: https://samtools.github.io/hts-specs/VCFv4.2.pdf. (Accessed: 24th January 2017)

56. Primer 3. Available at: http://primer3.sourceforge.net/. (Accessed: 10th January 2017)

57. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

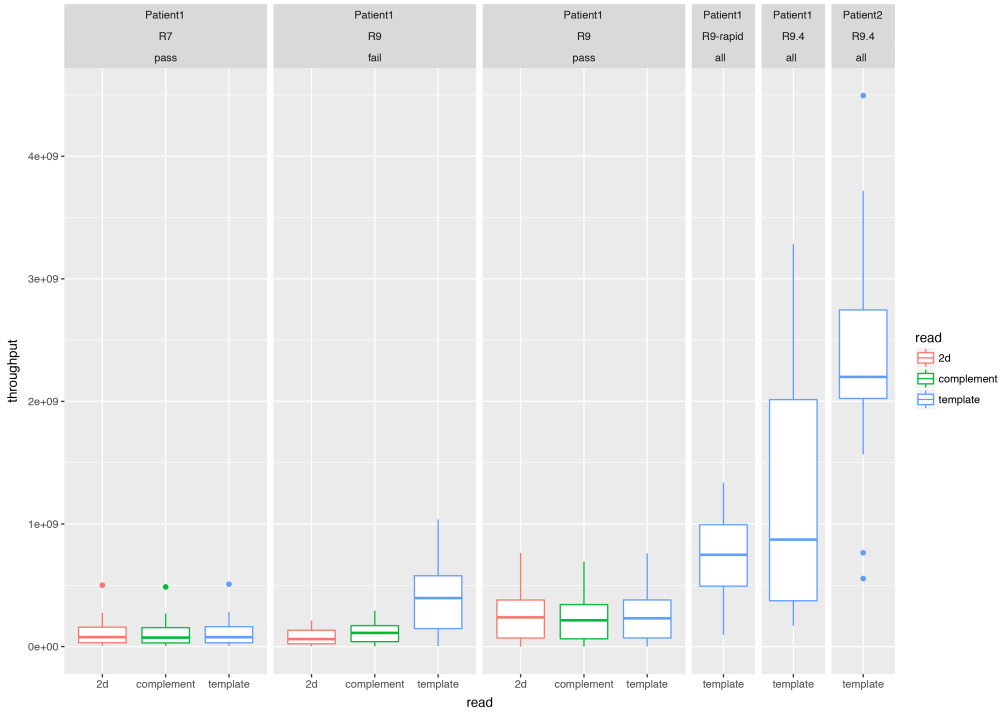
58. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44,** D81–9 (2016).

59. Redin, C. *et al.* The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet.* (2016). doi:10.1038/ng.3720

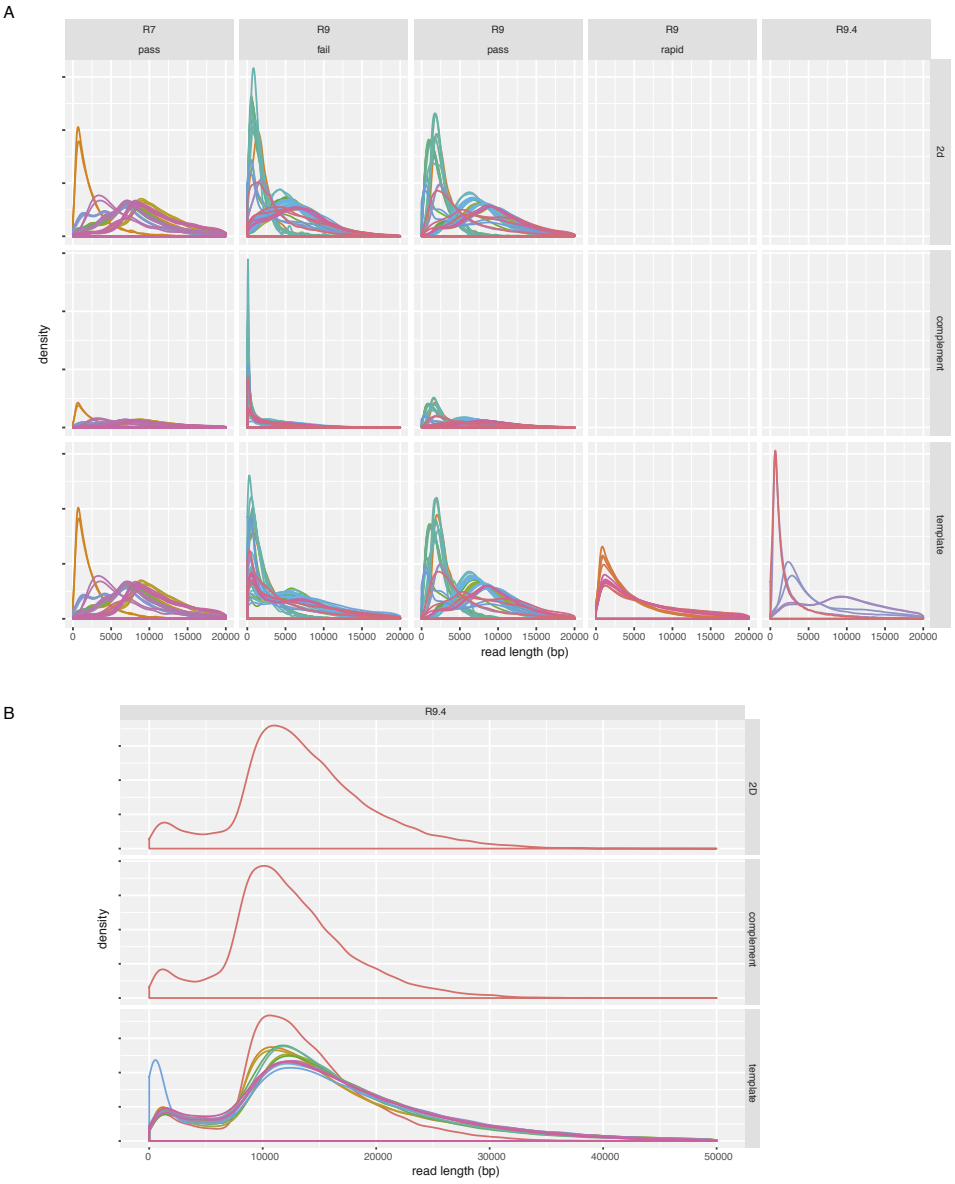
60. Boeva, V. *et al.* Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27,** 268–269 (2011).

SUPPLEMENTARY INFORMATION TO CHAPTER 4

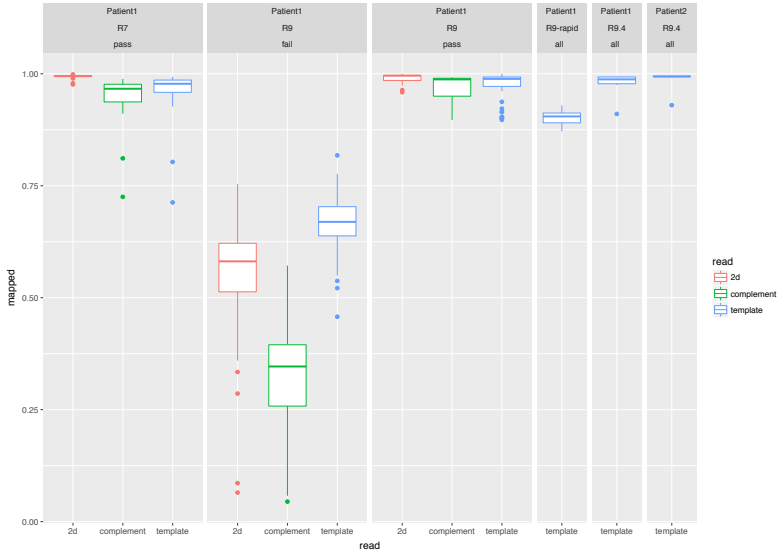
Supplementary Figures



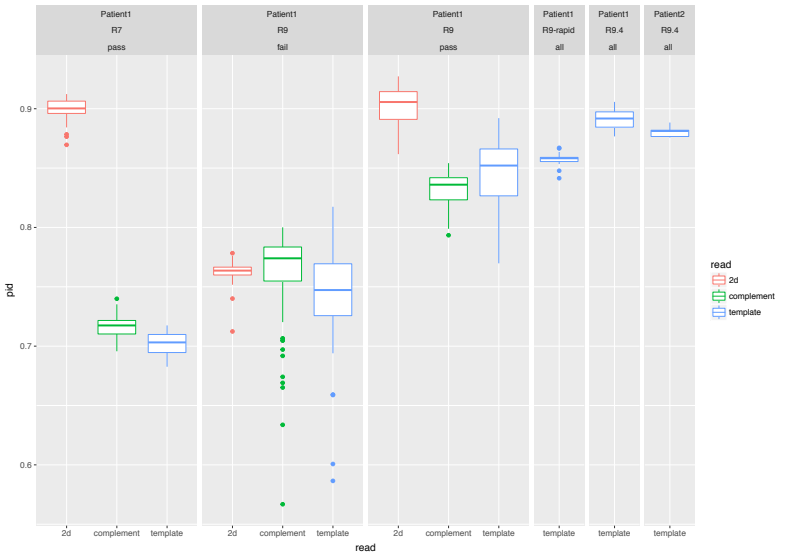
Supplementary Figure 1: Distributions of sequencing throughput for MinION sequencing runs.
For each sequencing run the throughput was calculated as the sum of the read lengths for each type of read (template, complement and 2D). The boxplots indicate the distribution of the throughput for multiple runs, stratified by data type (R7, R9_pass, R9_fail, R9_rapid and R9.4). The y-axis indicates throughput in bases and the x-axis shows read types. R7, R9 and R9.4 represent different nanopore sequencing chemistries for the MinION. Pass and fail indicates reads that were classified as either 'pass' or 'fail' following Metrichor basecalling. 2D indicates consensus reads generated from a template and complement read of a DNA duplex. 1D template and complement indicate reads derived from only one of the two strands (template or complement) of a DNA duplex. 'Rapid' means data from a rapid nanopore library preparation.



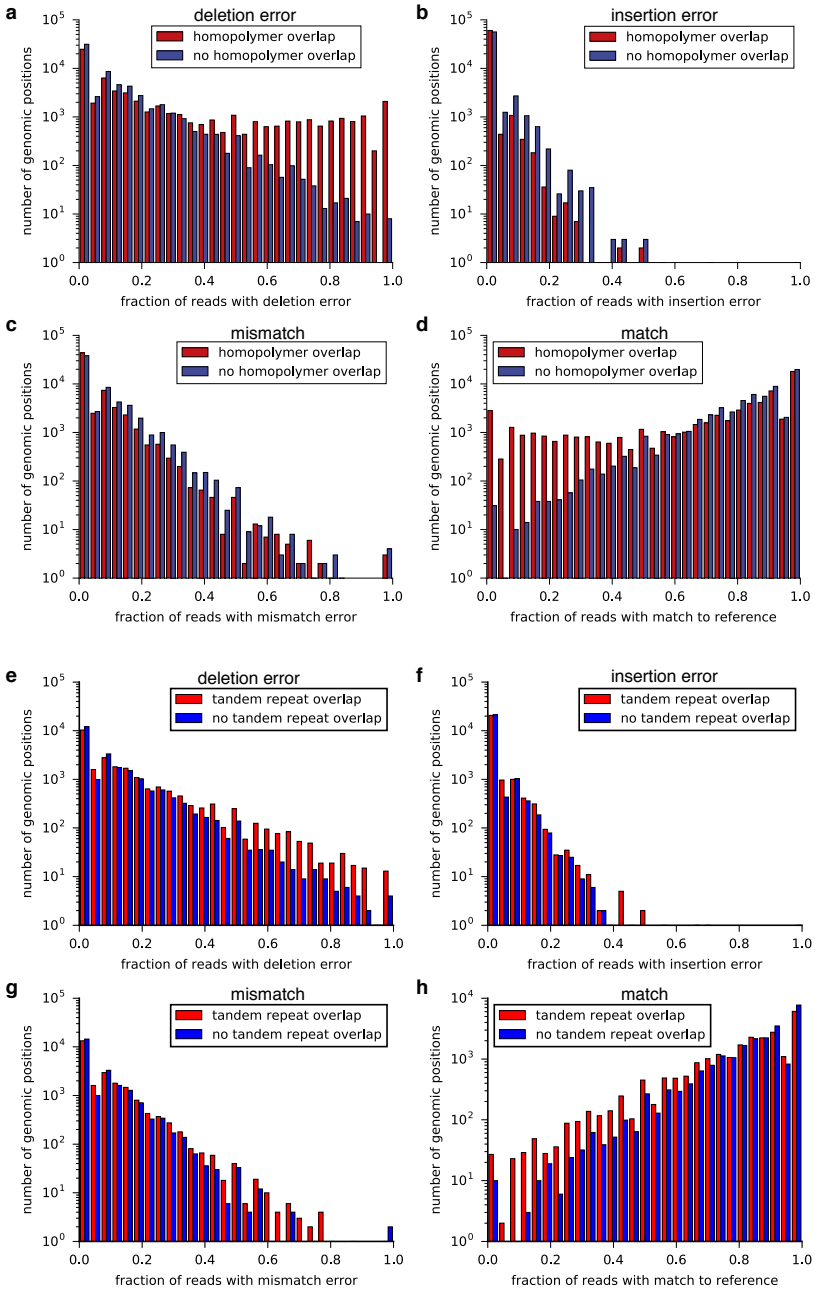
Supplementary Figure 2: MinION run read length distributions.
A Read length distributions for runs for Patient1. **B** Read length distributions for runs for Patient2. MinION sequence runs are indicated by different line colors. Read lengths were calculated for all 1D and 2D reads in the fastq file for each run. Plots are stratified by run type (R7, R9, R9_rapid and R9.4) and data quality ('passed' and 'failed' R9 reads following EPI2ME basecalling) in the horizontal direction and by read type (2D, 1D template, 1D complement) in the vertical direction. R7, R9 and R9.4 represent different nanopore sequencing chemistries for MinION. Pass and fail indicates reads that were classified as either 'pass' or 'fail' following Metrichor basecalling. 2D indicates consensus reads generated from a template and complement read of a DNA duplex. 1D template and complement indicate reads derived from only one of the two strands (template or complement) of a DNA duplex. 'Rapid' means data from a rapid nanopore library preparation.



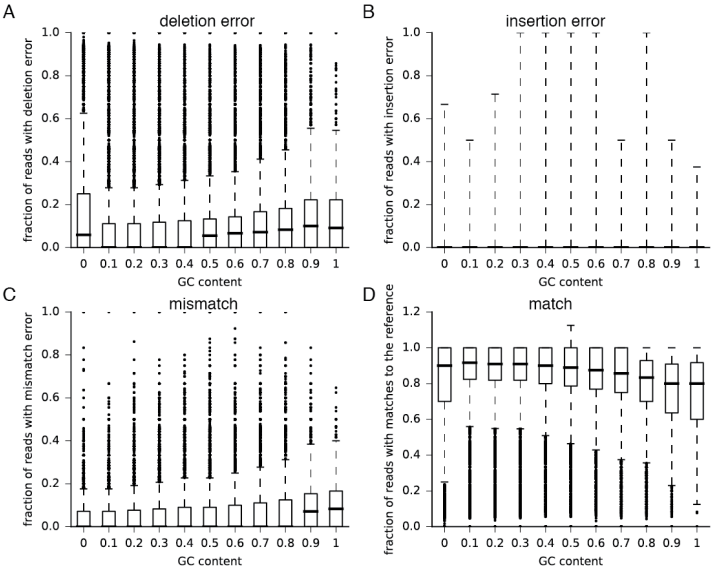
Supplementary Figure 3. Distributions of the percentages of reads mapped by LAST.
For each sequencing run the percentage of mapped reads was calculated. The boxplots indicate the distribution of the percentages of mapped reads for multiple runs, stratified by data type (R7, R9_pass, R9_fail, R9_rapid and R9.4). R7, R9 and R9.4 represent different nanopore sequencing chemistries for MinION. Pass and fail indicates reads that were classified as either 'pass' or 'fail' following Metrichor basecalling. 2D indicates consensus reads generated from a template and complement read of a DNA duplex. 1D template and complement indicate reads derived from only one of the two strands (template or complement) of a DNA duplex. 'Rapid' means data from a rapid nanopore library preparation.



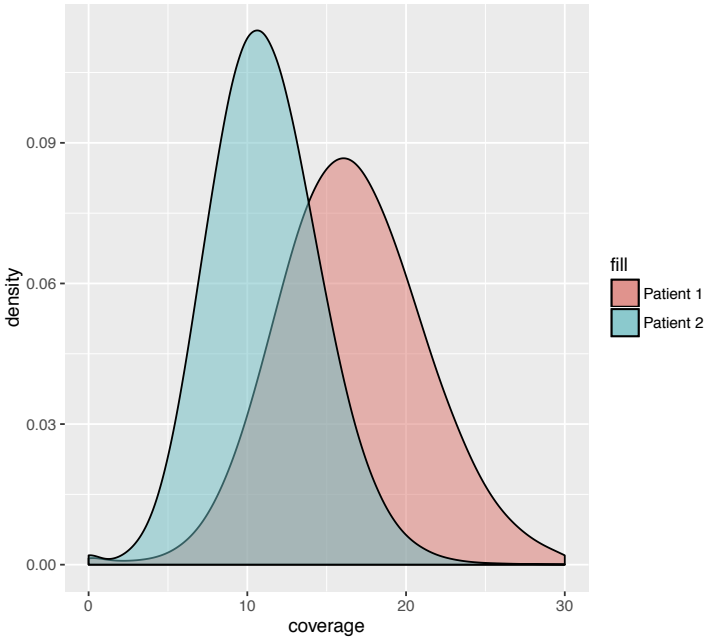
Supplementary Figure 4: Distributions of alignment accuracies of LAST alignments.
For each sequencing run the percentage of identical bases (PID) between reference and read sequences were calculated in the alignments. The calculation was done per mapped segment by dividing the amount of identical bases by the length of the mapped segment. Boxplots show the distribution of percentages stratified by run type (R7, R9, R9_rapid and R9.4) and data quality ('passed' and 'failed' R9 reads following EPI2ME basecalling) and by read type (2D, 1D template, 1D complement). R7, R9 and R9.4 represent different nanopore sequencing chemistries for MinION. Pass and fail indicates reads that were classified as either 'pass' or 'fail' following Metrichor basecalling. 2D indicates consensus reads generated from a template and complement read of a DNA duplex. 1D template and complement indicate reads derived from only one of the two strands (template or complement) of a DNA duplex. 'Rapid' means data from a rapid nanopore library preparation.



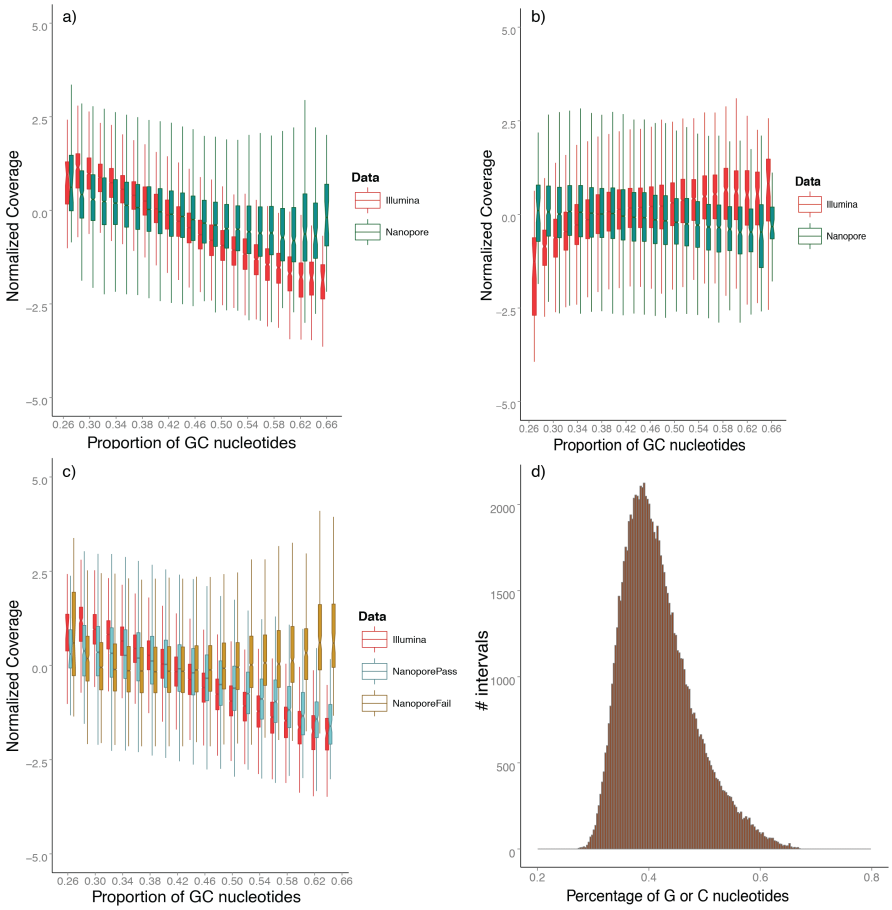
Supplementary Figure 5: Error profiles of R9.4 MinION sequencing data related to homopolymer and tandem repeat context.
A set of 1,064,470 randomly generated genomic positions (excluding polymorphic sites) were sampled from chromosome 1. For each of these positions the fraction of reads with deletion errors, insertion errors and mismatches was determined, based on MinION data from Patient2 (R9.4). In addition, the distance to the closest homopolymer (Methods) or tandem repeat (UCSC Simple Repeats track) was calculated. a-d Overlap of genomic sites with and without homopolymer overlap (without: >200bp away from nearest homopolymer), stratified by error class (a deletion, b insertion, c mismatch, d fraction of bases matching the reference). e-h Overlap of genomic sites with and without tandem repeat overlap (with: tandem repeat overlap and no homopolymer overlap, without: >300bp away from nearest tandem repeat and no homopolymer overlap), stratified by error class (e deletion, f insertion, g mismatch, h fraction of bases matching the reference).



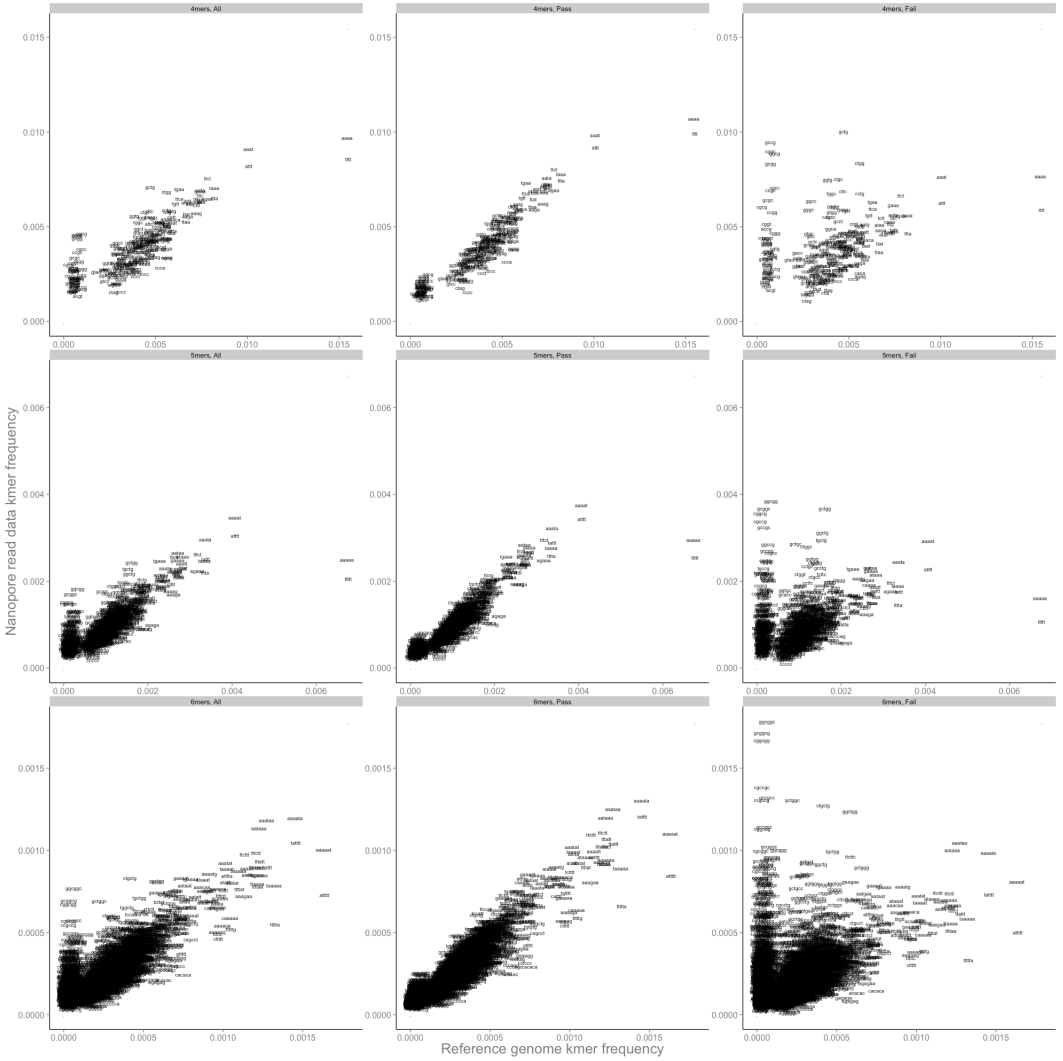
Supplementary Figure 6: Genomic GC content versus error rate in R9.4 MinION sequencing data.
A set of 1,064,470 randomly generated genomic positions (excluding polymorphic sites) were sampled from chromosome 1. For each of these positions the fraction of reads with deletion errors, insertion errors and mismatches was determined, based on MinION data from Patient2 (R9.4). In addition, the GC content of the reference genome was calculated based on a window of 10bp at each examined genomic position. a The fraction of deletion errors, b insertion errors, c mismatches and d matches to the reference genome are depicted (y-axis) as a function of genomic GC content (x-axis). For deletion errors, a linear regression model shows a statistically significant dependency of the error rate on the GC content ($p < 10^{-16}$). The estimated coefficient, as change of error fraction per percent of GC-content, is 0.0072 (std. error = 0.0007).



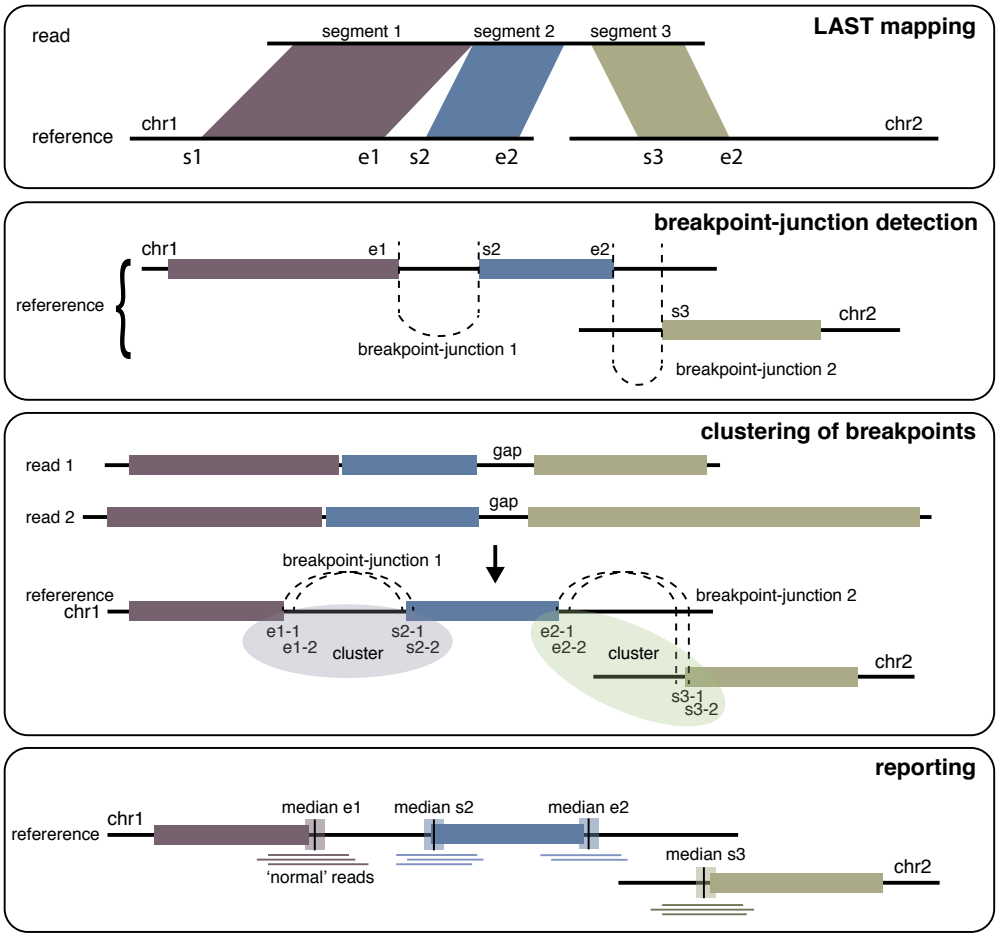
Supplementary Figure 7: Coverage distribution for sequencing data from Patient1 and Patient2.
Coverage distributions were generated by calculating the coverage for 1,000,000 random genomic positions, excluding positions in the gap table downloaded from the UCSC genome browser (GRCh37 gaps in golden path).



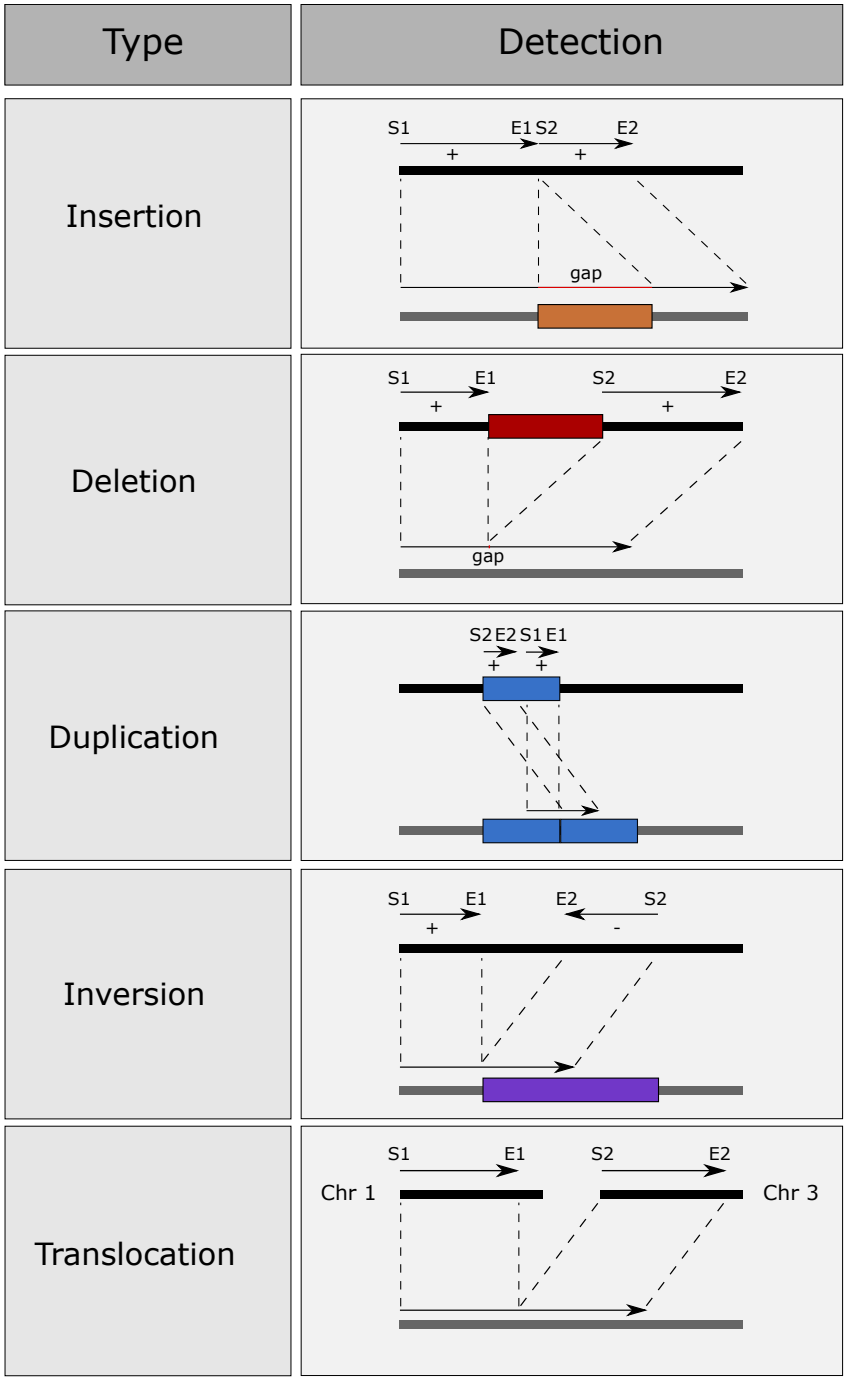
Supplementary Figure 8: Average coverage distribution as a function of GC-content for MinION and Illumina sequencing data of Patient1 and Patient2.
Panels a and b show statistics of depth of coverage for Illumina data (red) and MinION data (green) for Patient1 and Patient2 respectively. Panel c shows statistics of depth of coverage for Illumina data (red), MinION nanopore "pass" data (green) and MinION nanopore "fail" data (dark yellow) of Patient1. Panel d shows the GC content distribution across our randomly sampled intervals. The average coverages across 100,000 randomly sampled 5kb genomic intervals were used in each panel. Average coverage outliers, defined as 6 or more interquartile distances away from the median, were discarded for each technology respectively. The remaining data were normalized to N(0,1), to account for different genome-wide sequencing average coverage and binned by GC-content. A linear regression model shows a statistically significant dependency of coverage depth on the GC content expressed as percentage, for both technologies ($p < 10^{-16}$). The estimated coefficients, as number of standard deviations of change, per percentage of GC-content, are -0.094 (std. error = 0.0004) and -0.029 (std. error = 0.0004) for the Illumina and MinION data of Patient1 respectively (panel a non-binned data). Conversely the estimated coefficients for Patient2 are 0.033 (std. error = 0.0004) and -0.018 (std. error = 0.0004) for the Illumina and MinION nanopore data respectively (panel b non-binned data).



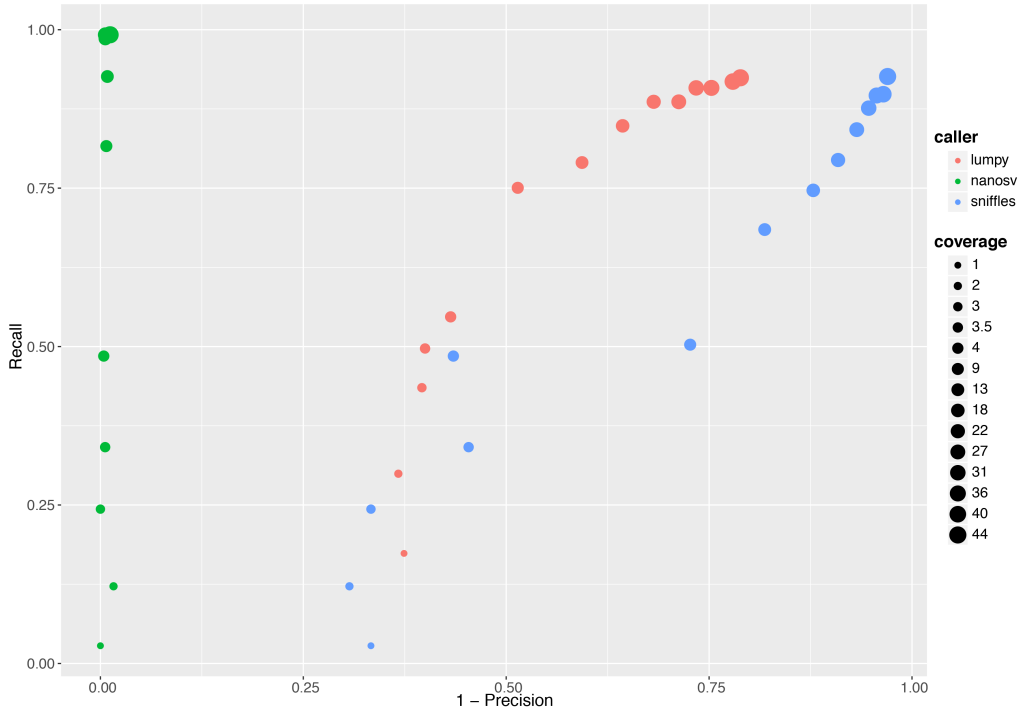
Supplementary Figure 9: K-mer distribution in MinION sequencing data of Patient1.
Plotted observed (MinION data) versus expected (GRCh37 reference genome) relative k-mer frequencies for 4-mers (top), 5-mers (middle) and 6-mers (bottom). The expected kmer frequencies are computed from the relative frequency of each kmer on the reference genome primary assembly for each k-mer size. The MinION data k-mer frequency was similarly computed, across all MinION reads, further stratified by “pass” (middle) or “fail” (right) read status. The “All” (left) represents the aggregate “pass” and “fail” MinION data.



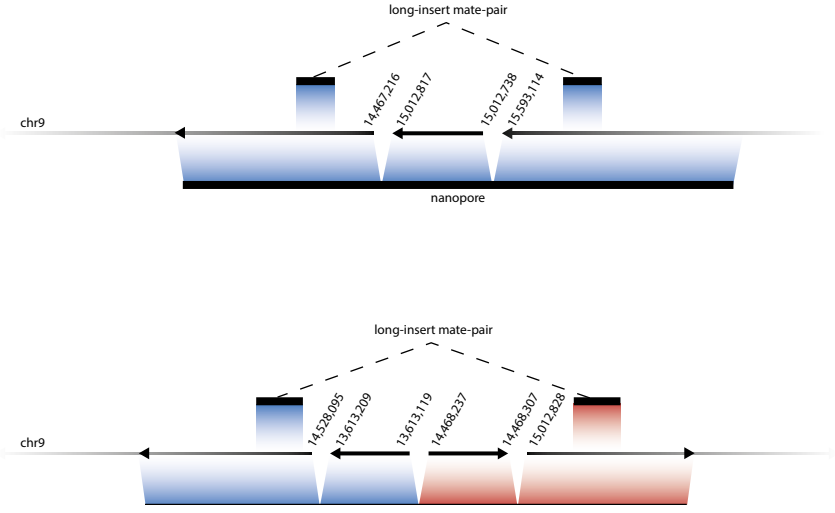
Supplementary Figure 10: Overview of NanoSV algorithm.
NanoSV uses LAST mapping output for discovery of SVs. In a first step candidate breakpoint junctions are detected using split read mappings. Candidate breakpoint junctions are subsequently clustered across multiple reads based on the overlap of junction coordinates and orientation. Clusters of breakpoint junctions are reported as SVs in VCF format. The tool is available on github: https://github.com/mroosmalen/nanosv.



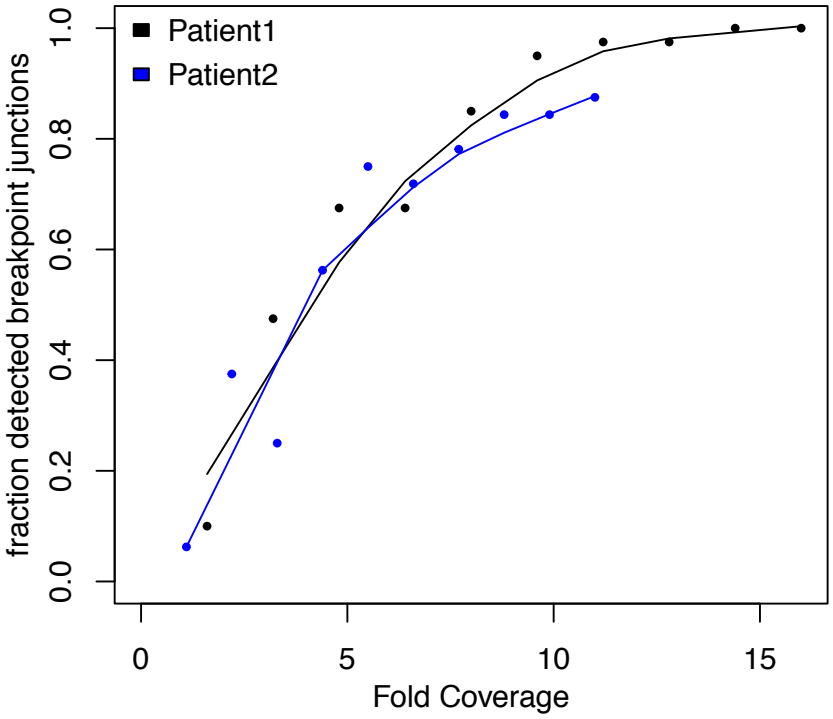
Supplementary Figure 11: Detection of different SV types by NanoSV.
NanoSV detects most types of breakpoints junctions with the exception of insertions consisting of unmapped repeat elements which are longer than the nanopore read lengths, e.g. LINE insertions may be missed if the read length falls below the typical length of LINE elements (~6kb). Genomic coordinates of mapped segments are indicated by s1/s2 (start of segments) and e1/e2 (end of segments). Gaps within reads represent unmapped segments, which may result from repeat insertions or complex variations. Deletions are discerned from insertions if the gap length is smaller than the distance between the joined genomic positions (s2-e1 which represents SV size for variants other than insertions).



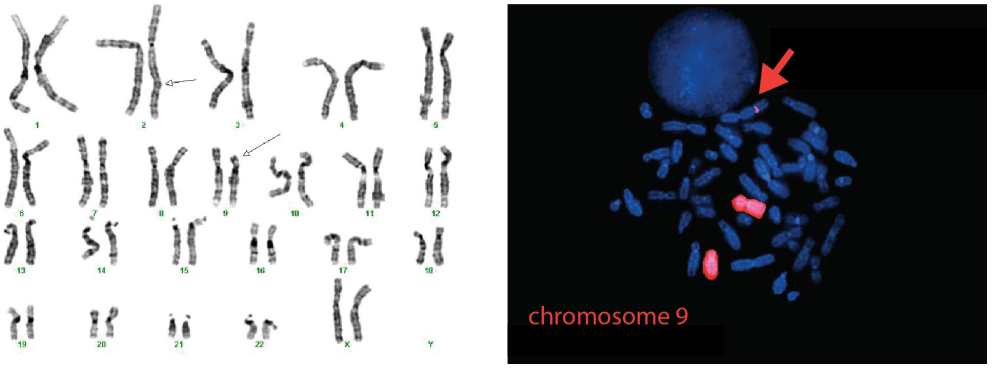
Supplementary Figure 12: Recall-precision curve for SV calling performance on simulated nanopore data.
Breakpoints (501) were simulated on reference chr1 and based on the resulting chromosomal sequence nanopore reads were simulated using NanoSim1. SV calling using Lumpy2, Sniffles3 and NanoSV was performed on subsets of the simulated nanopore reads to estimate the effect of read coverage. The recall (true positives/true positives + false negatives) and precision (true positives/true positives + false positives) was calculated for each call set, without any additional post-calling filters being applied.



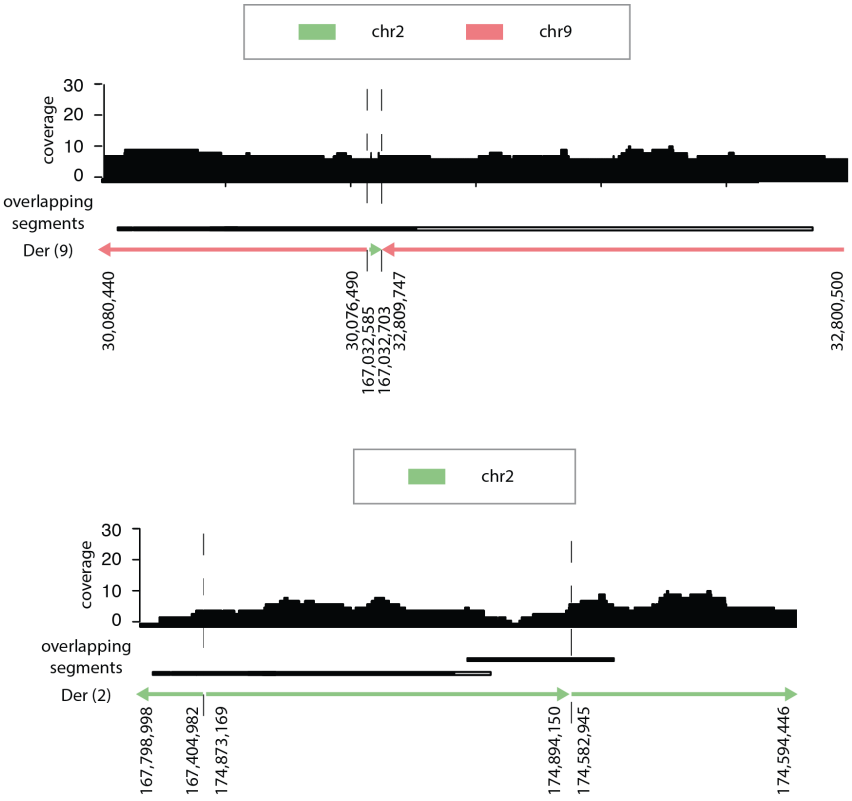
Supplementary Figure 13: Structure of two complex breakpoint-junctions in Patient2 chromothripsis.
Long-insert mate-pair sequencing was previously used to study the chromothripsis in Patient24. The long-insert size of these mate-pair libraries hampers detection of short chromosomal segments, because the short sequence reads can jump over the short segments and only reveal the connection between the segments flanking these short segments. In the upper panel, an 80bp segment from chr9 is depicted, which was identified using nanopore reads and confirmed by Sanger sequencing. The lower panel highlights two adjacent short genomic segments - both from chr9 - that were missed by the long-insert mate-pair sequencing, but detected by nanopore reads and subsequent PCR and Sanger sequencing.



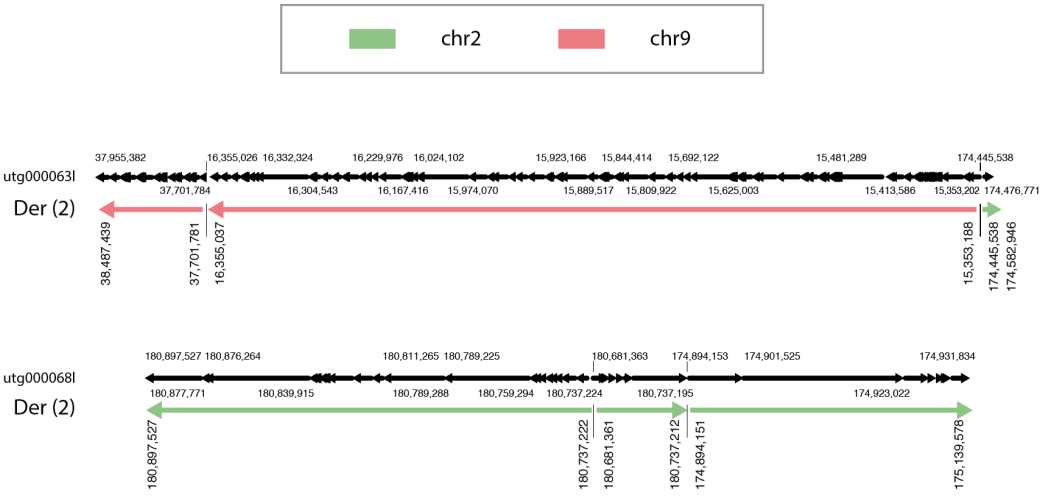
Supplementary Figure 14: The effect of subsampling the MinION sequencing data on chromothripsis breakpoint-junction detection.
Nanopore sequencing reads were subsampled from 10% to 90% of the original data and each subsampled dataset was analyzed using NanoSV to determine the fraction of known chromothriptic breakpoint-junctions that could be detected. Below a coverage of ~14x (Patient1), the fraction of detected breakpoint junctions drops below 1.



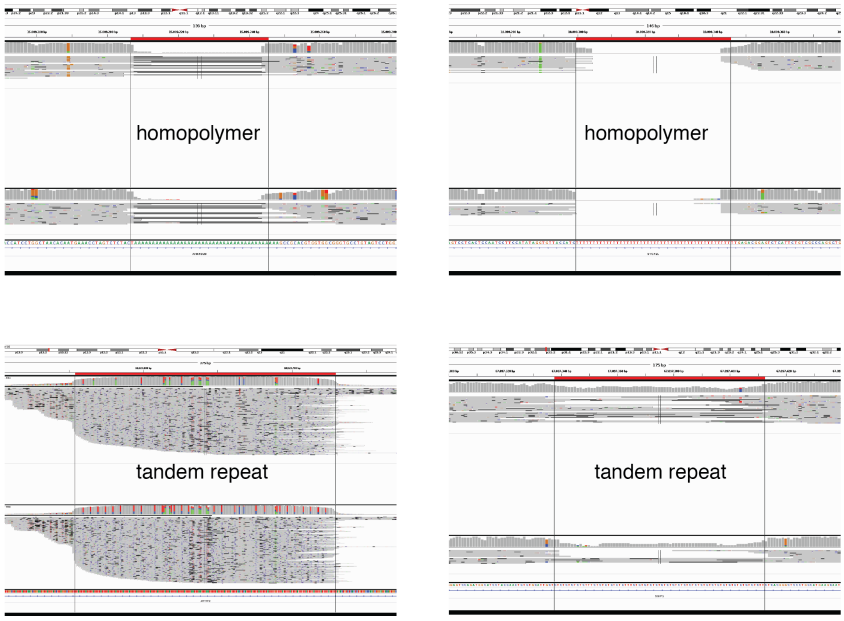
Supplementary Figure 15: Karyotype and chromosome 9 painting derived from Patient1 chromosome spreads.
Left panel shows the patient karyotype. Arrows indicate chromosome 2 and chromosome 9. The right panel displays a chromosome 9 paint (red) demonstrating an insertion of a part of chromosome 9 into chromosome 2 (arrow).



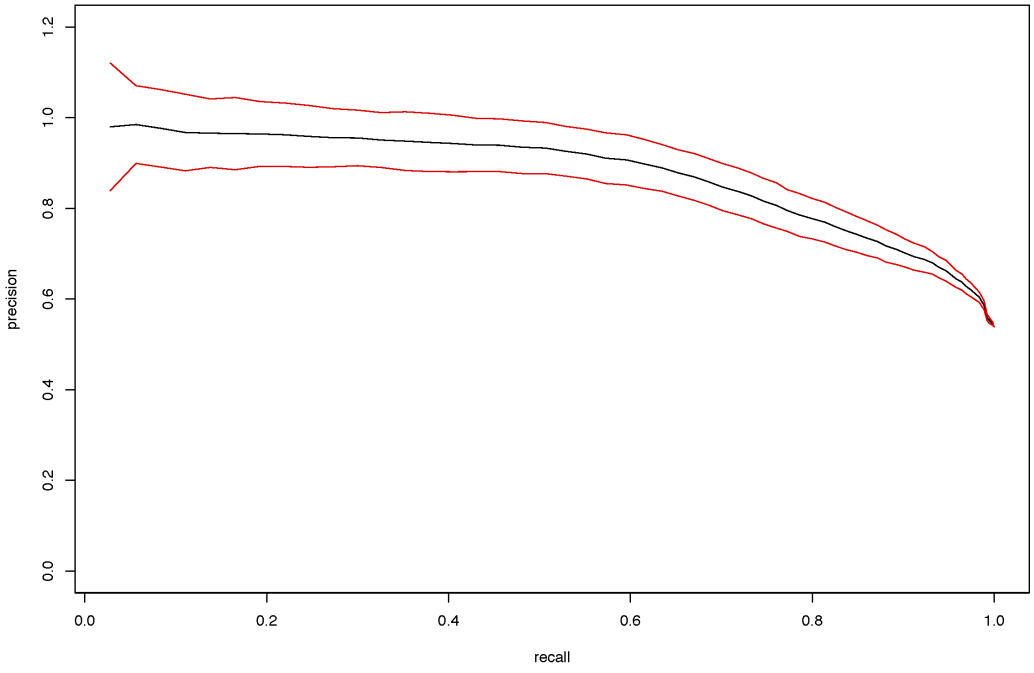
Supplementary Figure 16: Reference-assisted assembly of chromothripsis regions in Patient1.
Order and orientation of chromosomal regions involved in the chromothripsis rearrangements of Patient1 is depicted by colored lines with arrowheads. The resulting chromosomal configuration is based on overlapping nanopore reads derived from the paternal haplotype of Patient1. Nanopore reads that are instrumental for segment connectivity are indicated by black bars. The coverage track has been generated from all paternal reads mapping to the respective chromosomal segments. The order and orientation of the joined chromosomal segments matches the chromothripsis structure that is described in Figure 3.



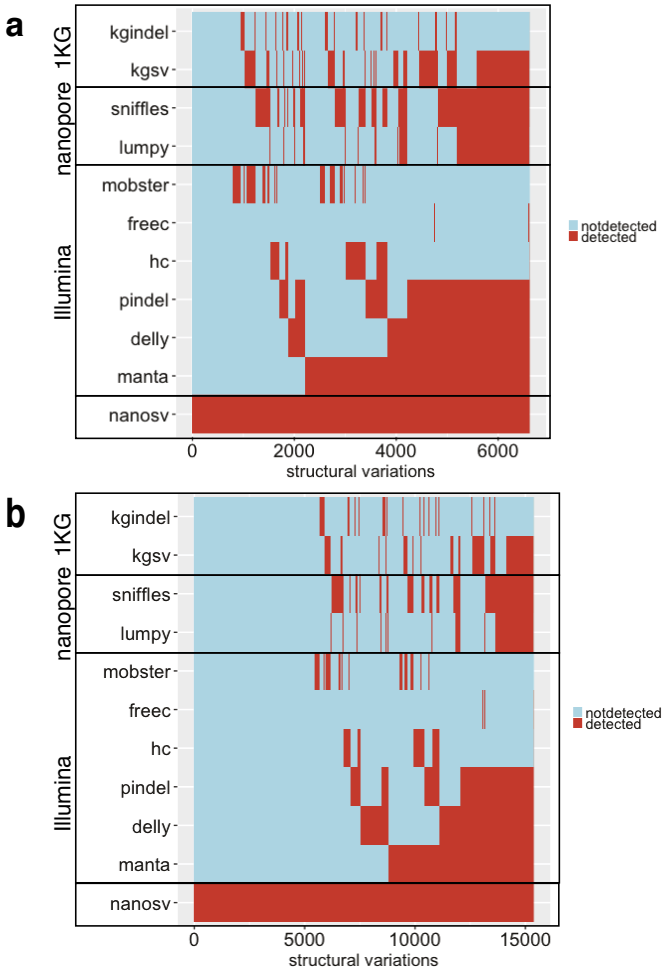
Supplementary Figure 17: Contig structure produced by Miniasm assembly of chromothripsis regions in Patient1.
Order and orientation of chromosomal regions involved in the chromothripsis rearrangements of Patient1 is depicted by colored lines with arrowheads and was obtained as for Figure 3. The structure for two chromothriptic regions, containing three genomic segments each, was supported by contiguous sequences (contigs) resulting from Miniasm5 assembly of nanopore reads, excluding reads that were assigned to the maternal haplotype. Both contigs (utg000068l and utg000063l) support part of the structure of derivative chromosome 2. The black arrows indicate the positions and orientations of contig segments mapped to the human reference genome (GRCh37).



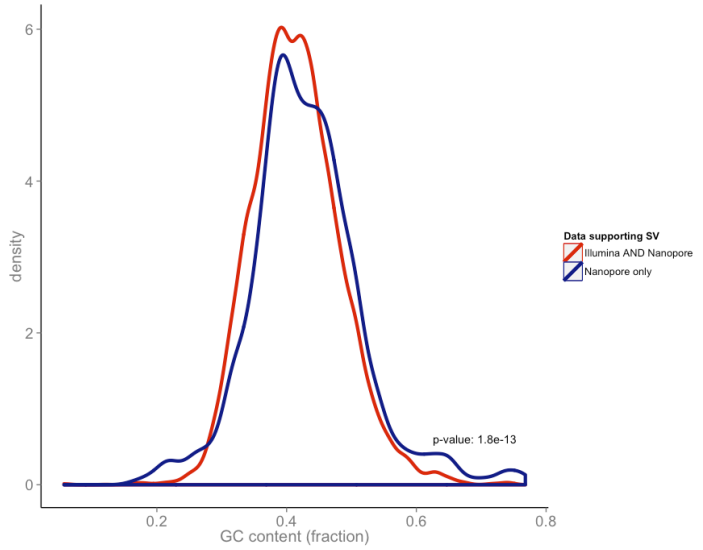
Supplementary Figure 18: IGV screenshots showing MinION nanopore read alignments in homopolymer and tandem repeat regions.
For each panel the upper alignments are from Patient1 and the lower alignments are from Patient2 MinION read data. The lower left panel represents a NanoSV predicted duplication call and the remaining three examples represent NanoSV predicted deletions.



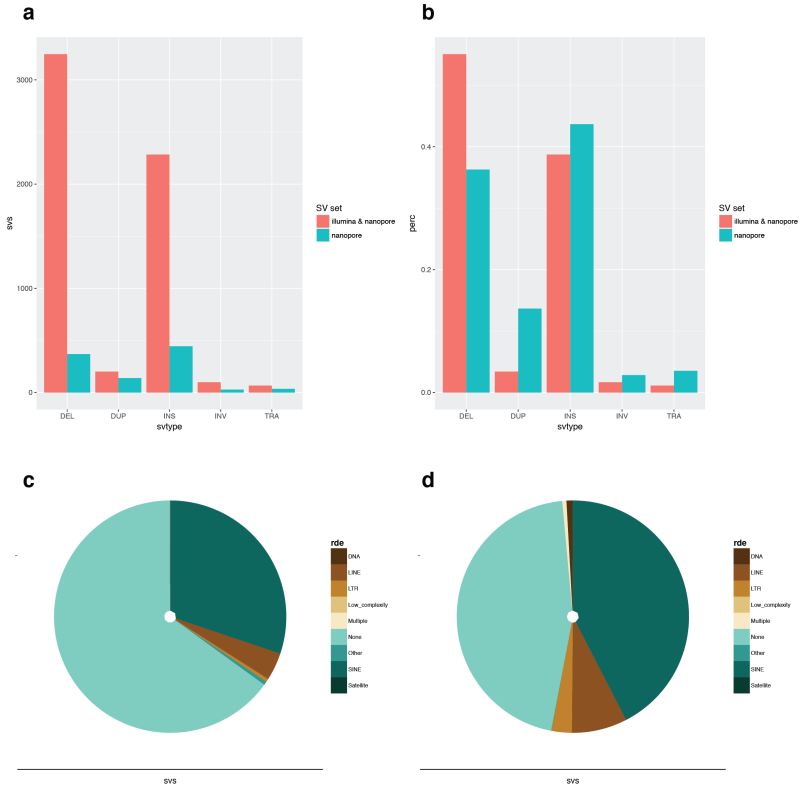
Supplementary Figure 19: Recall-precision curve obtained from training and cross-validation on NanoSV SV calls.
The illustrated ROC curve is obtained from 100 cross-validation random forest training runs (split 90%-10% for training-testing) from the total set of 354 true positive and 300 true negative SVs from the NA12878 sample. The chosen, optimal operating point has a precision of 82% at a recall rate of 75%.



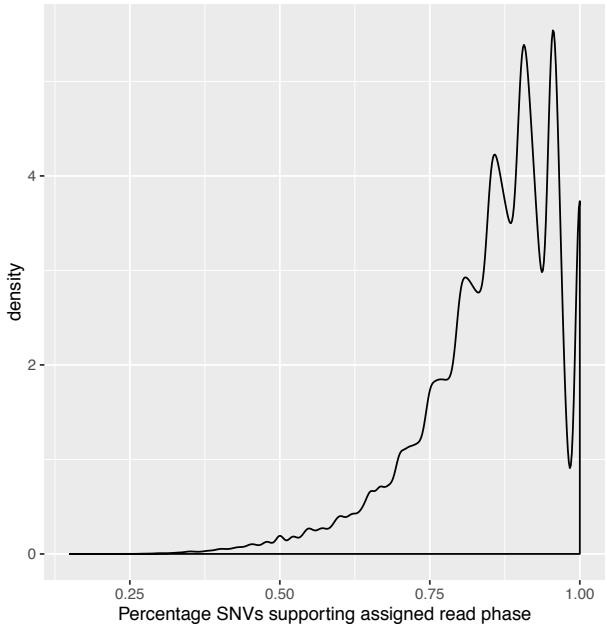
Supplementary Figure 20: Heatmap showing the overlap of SV calls between different callers and SV datasets.
We used the NanoSV SV call set of Patient1 and Patient2 as a basis for intersection with SV call sets generated from Illumina data, using six different tools. Additionally, we used two tools for detection of SVs in the Nanopore data from Patient1 and Patient2. Finally, we intersected the NanoSV calls with the 1000 Genomes phase 3 consensus calls6. a Heatmap showing overlaps of 6,616 NanoSV SVs predicted as true positive by a random forest classifier (Methods). b Heatmap showing overlaps of the initial call set consisting of 15,369 candidate NanoSV SVs, following filtering for SVs that overlap homopolymers and tandem repeats (Methods).



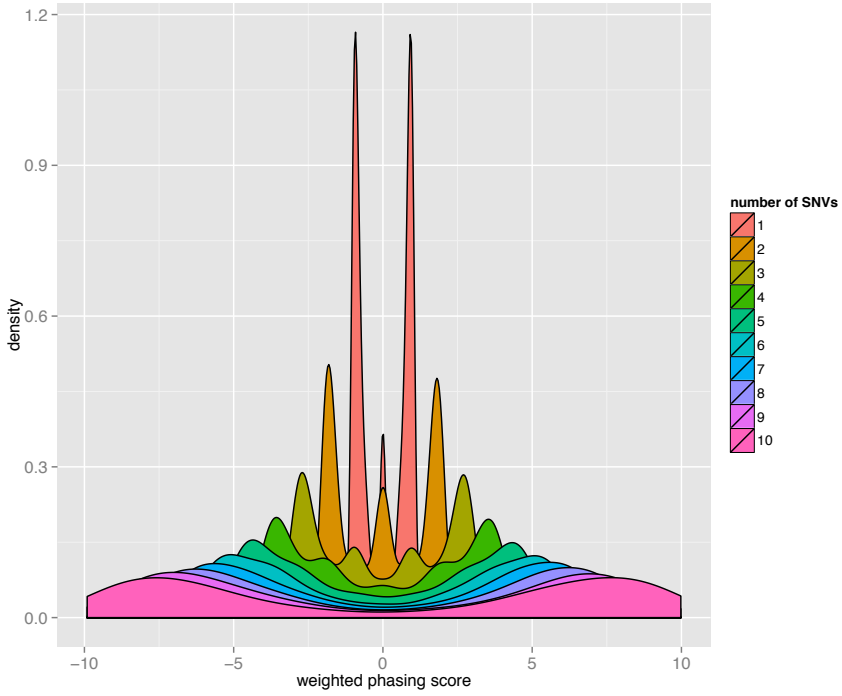
Supplementary Figure 21. GC bias of nanopore specific SVs.
GC content distributions across 500 base-pair windows around the high confidence set of SV calls that are detected in both Illumina and MinION nanopore data (red) and nanopore data only (blue). The average GC content in the regions where an SV is detected only in the nanopore data is 1.4% higher than the the average GC content where an SV is detected in both Illumina and MinION nanopore data (Welch two sample t-test: p-value = 1.8e-13, 95% CI = 1.0 - 1.8).



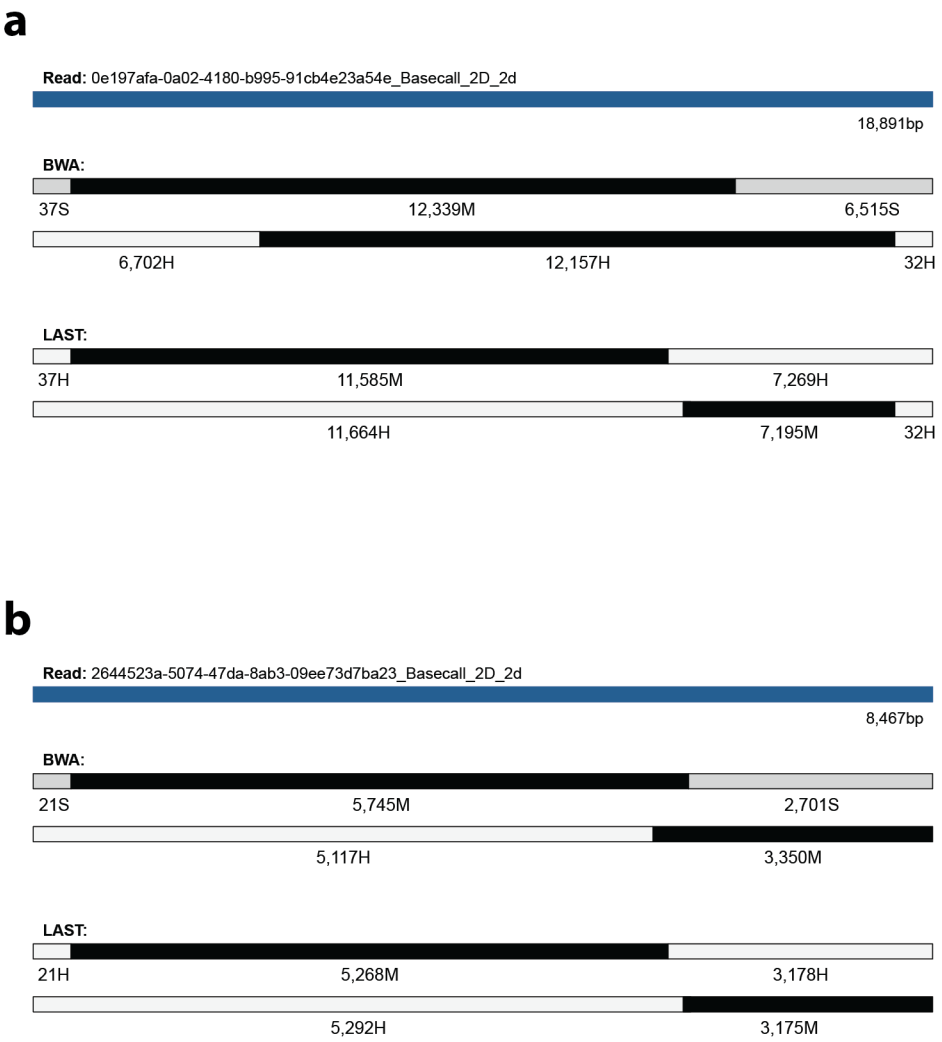
Supplementary Figure 22: Patient1 and Patient2 cumulative distributions of SVs.
We plotted numbers of SV calls across SV types (a and b) and across SV annotations (c and d), after random forest filtering. a shows the histogram of SV type across both patients, subsetted for the “Illumina and nanopore” data and “nanopore” only data. b shows the SV type distribution for the same subsets as a. c shows the annotations distribution, by class, for all deletions detected in both nanopore and Illumina data. d shows the annotations distribution, by class, for all insertions detected in both nanopore and Illumina data.



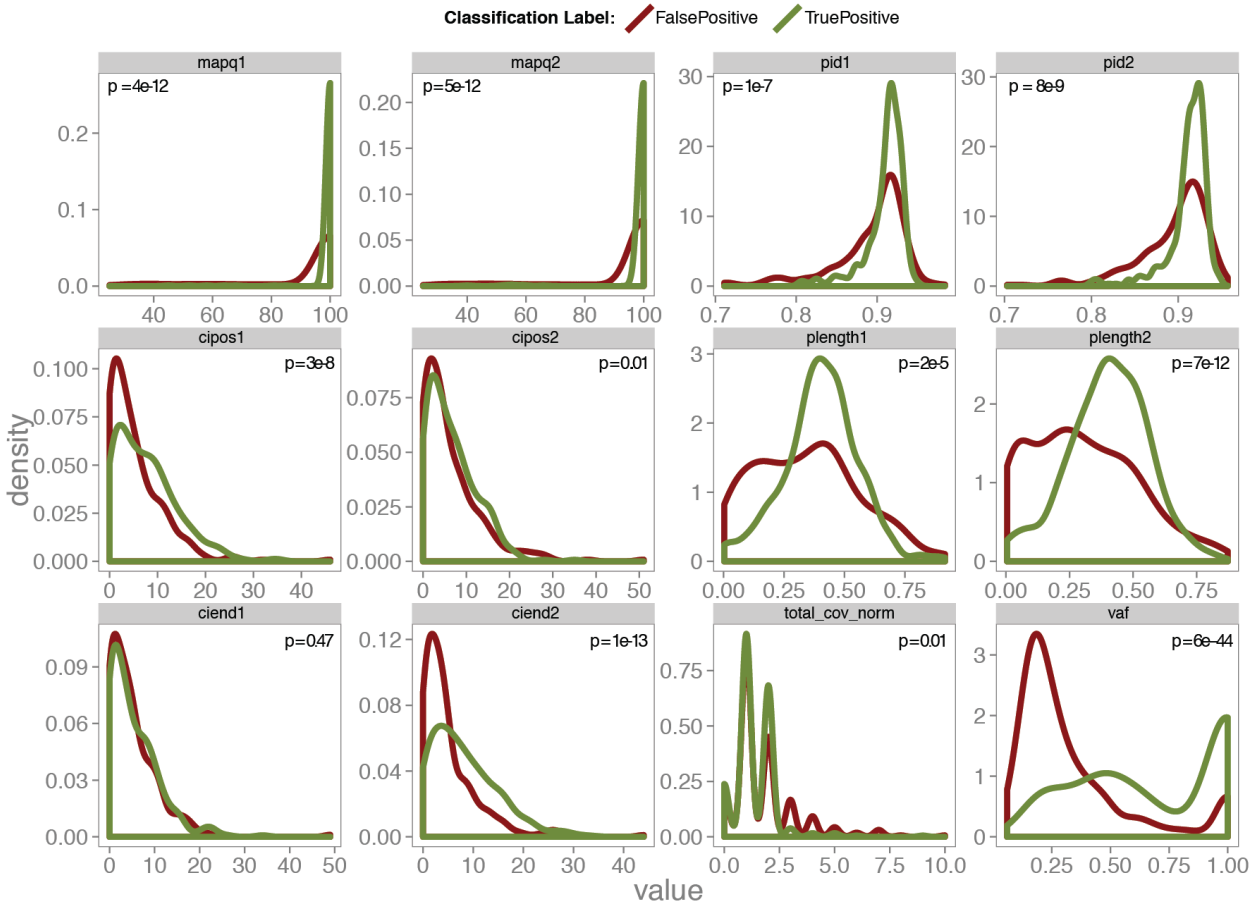
Supplementary Figure 23: Nanopore read phase support.
The plot show the distribution (density) of the percentage p of SNVs per read supporting the read phase of each nanopore read covering at least 20 phase-informative SNVs. The percentage p is defined as SNVsupp/SNVtotal, where SNVsupp is the number of phase-informative SNVs that support the read phase and SNVtotal is the total number of phase-informative SNVs covered by the nanopore read.



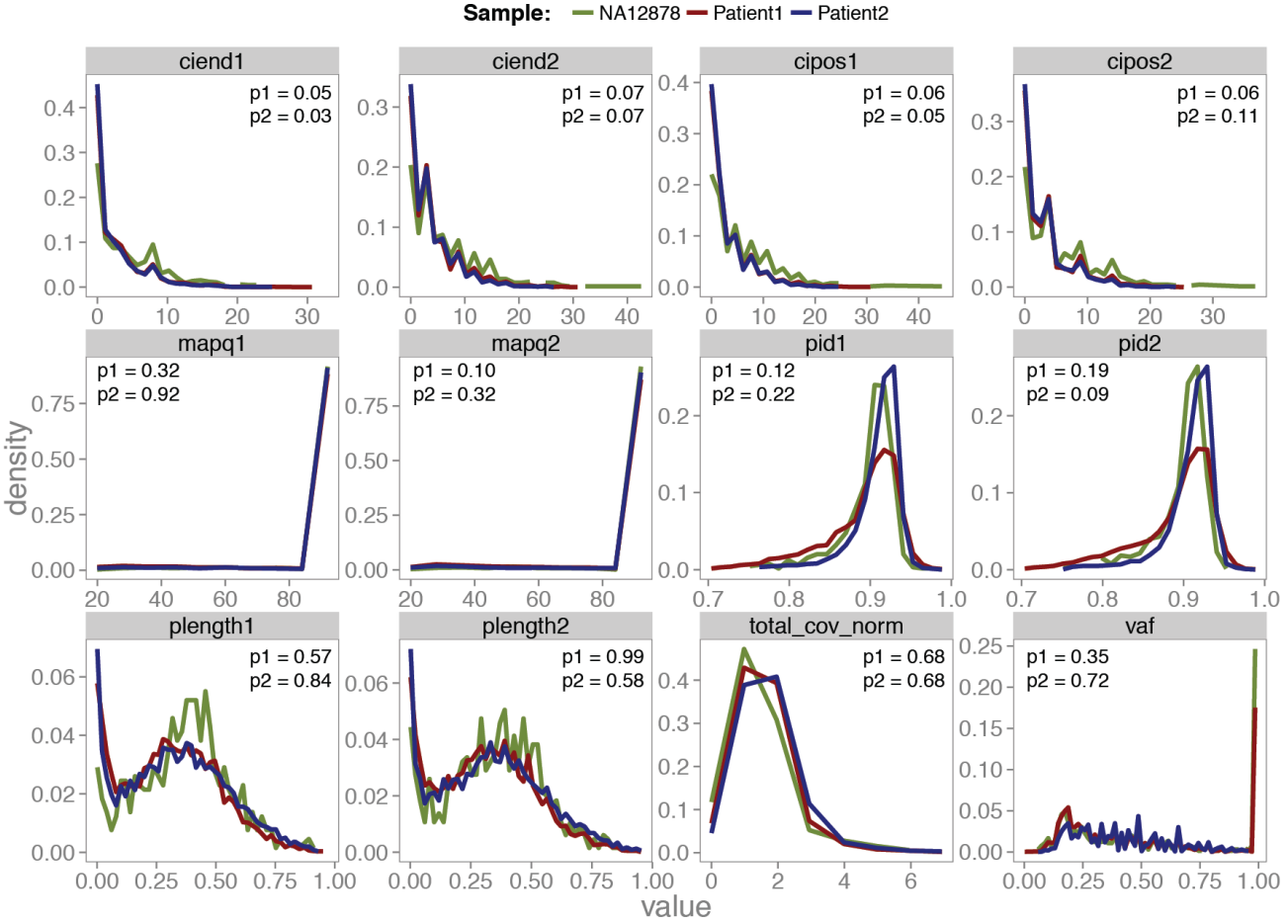
Supplementary Figure 24: Phasing-score distribution for nanopore reads from Patient1.
For each nanopore read a phasing-score S was calculated (x-axis, Methods). The plot shows the distribution of phasing scores (S) for nanopore reads overlapping 1 to 10 phase-informative SNVs. If the phasing score S is positive, the read is assigned to the paternal haplotype, while for a negative value of S the read is assigned to the maternal haplotype.



Supplementary Figure 25: Alignment differences between BWA MEM and LAST.
Two examples (a and b) of how BWA and LAST segment the same read differently at alignment. Each whole read is depicted in blue. For each caller, the two grey/black lines depict how the read is split into two segments at alignment. The black line depicts the part of the read that is aligned and the grey parts depict the clipped parts of the read, for each segment respectively. Both these examples show how bwa splits reads into (at least slightly) overlapping segments, which impair our ability to evaluate candidate breakpoints. Only the read from example b contributes to a non HOM_REF SV call in our dataset.



Supplementary Figure 26: Distribution of random forest feature values in the NA12878 data.
Distribution of random forest feature values, within the NA12878 training data, for true positives (green) and false positives (red) respectively. P-values are derived from a two-sided unpaired wilcoxon test.



Supplementary Figure 27: Distribution of random forest feature values across samples.
Distribution of random forest feature values across all SV calls (after filtering for homopolymers and simple repeats) within NA12878 (green), Patient1 (red) and Patient2 (blue). The feature distribution of the training data (NA12878) is compared to the feature distribution of the two test samples, Patient1 and Patient2 using a wilcoxon paired test and the two p-values are reported in each feature plot; p-value p1 (comparing NA12878 and Patient1 distributions) and p-value p2 (comparing NA12878 and Patient2 distributions).

Supplementary Tables:

See online Supplementary data at:

https://www.nature.com/articles/s41467-017-01343-4

References

1. Yang, C., Chu, J., Warren, R. L. & Birol, I. NanoSim: nanopore sequence read simulator based on statistical characterization. Gigascience (2017). doi:10.1093/gigascience/gix010

2. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15, R84 (2014).

3. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single molecule sequencing. (2017). doi:10.1101/169557

4. Redin, C. et al. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. Nat. Genet. (2016). doi:10.1038/ng.3720

5. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32, 2103–2110 (2016).

6. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81 (2015).

CHAPTER 5

DISCUSSION

Summary

In this discussion, I will focus on a few key aspects where I believe that long read sequencing, nanopore sequencing in particular, can further contribute to our understanding of human genetic variation and eventually to its relation to disease.

I will start by illustrating some of the current advantages and limitations, as they crystallized during our structural variation analysis presented in **chapter 4,** and speculate on the implications and potential follow-up to our analyses.

Further, I will discuss two paths forward, that I believe together will define a new frame for representing and understanding genetic variation, and pave the way for precision medicine initiatives. I show how long read sequencing, on top of established NGS technologies, can facilitate such a transition. Specifically, I will discuss how the use of a human reference genome is shifting towards the use of a population reference of genomes, that should ease the representation of complex variants and increase the sensitivity of detecting population specific, or hard to sequence variants. Furthermore, the integration of different layers of data, beyond the mere linear sequence of the DNA molecule, but intrinsically related to it, is a key factor in the interpretation of (non-coding) genetic variation and can help to both finemap disease risk loci as well as to explain disease etiology. I will show how long read data should facilitate and accelerate the understanding of features such as DNA conformation, gene expression levels and isoforms, and epigenetic markers, producing the resources needed for complex models to explain disease and biological mechanisms.

This discussion focuses on long read nanopore sequencing as being an emerging technology that simplifies and improves a host of analyses, and not as an all-round better solution to the well established NGS technologies. I therefore briefly attempt to scope each technology (i.e.: long read and NGS) to applications to which I believe they are optimal. Finally, I end by highlighting the importance of negative findings and of reporting negative findings in literature.

Advantages and limitations of long read nanopore data

Complex translation of raw nanopore data to a sequence of bases

The relatively lower per-base sequencing accuracy is currently (still) a limitation of long-read nanopore sequencing. Although in **chapter 4** we report good genotyping quality (> 96%) for known SNV sites, applying standard pipelines to call and genotype SNV sites, resulted in twice as many SNVs, leading to a precision of approximately 50% (data not shown). Since its release, MinION per base sequencing accuracy increased from < 60%[1] to the current 90% (**chapter 4**). This has been facilitated by improved base calling algorithms, as I briefly reviewed in the introduction, but the major driver in this quality increase has been the pore used for the electrical DNA measurements. All the biological pores used, including the latest CsgG pore, are continuously artificially mutated, to increase the accuracy of measurements through the pore (i.e.: reduce outside interference)[2]. Improvements include an increased signal to noise ratio, by minimizing environment influences on the measurement and increased precision of the measurements sampled. Thus, while earlier pores measured the convoluted signal of six adjacent DNA bases (i.e.: 6-mer), it is estimated that the signal measurement of the newest CsgG pore is driven mostly by three adjacent bases, with a diffuse influence from the two bases bordering this 3-mer[2]. Such technological improvements are essential for further improvement of per-base sequencing error.

The high short-indel error rates make calling and genotyping indel variation unreliable by current standard tools. Indel errors in nanopore data are non-randomly distributed with respect to sequence content and this introduces systematic biases in indel calling. However, other features of indel errors could be exploited to increase the accuracy of calling such variation. For example, indel error length varies for a given locus (i.e.: there can be an erroneous deletion of one base, as well as an erroneous deletion of 3 bases, spanning the same position) and, typically, the evidence for a deletion is evaluated independently for each base. If instead, the whole indel captured in any read is considered as an allele, a one base-pair deletion and a three base-pair deletion would support different variants. An indel call would then only be made when there is systematic read evidence for the *same* indel, at some given locus. In this way, some upstream errors, stemming from base calling and/or irregular DNA traversal speed would be canceled out as producing random indel errors spanning a locus, whereas real indel variation would be captured with better precision .

Simpler and more precise (structural) variant calling

In **chapter 4** I showed how using long reads produced by nanopore sequencing facilitates an algorithmically simplified and unified approach to identifying and genotyping structural variation in the (diploid) human genome. Evidence for structural variation from long reads does not need to be integrated from multiple sources, such as split-read alignments and discordant mate-pair alignments, anymore. Instead, any variation is directly reflected in the (split-read) alignment of the long reads. The advantage is twofold.

Firstly, long read data enables a more transparent and straightforward variant calling pipeline. High quality NGS SV call sets are typically a consensus of up to a dozen different tools, where each tool is often engineered to have increased sensitivity and accuracy across specific SV classes or sizes. Furthermore, obtaining such high quality consensus sets involves additional engineering, which often requires arbitrary or ad-hoc quality thresholds. As a result, a set of good guidelines for SV detection is typically applied, rather than an end-to-end, standardized, algorithmic solution. Using a single algorithm and a machine learning post-calling filtering step on the long-read data, we were able to reach state-of-the art precision (> 95%) in SV detection genome wide. Our approach recovered ~70% of the SVs described in the high quality NGS consensus dataset of a well characterized Genome In A Bottle (GIAB) sample, which is more than the typical, estimated contribution of any NGS algorithm when consensus sets of SVs are derived from short read data[3].

Secondly, precise BP estimation is very important for accurate interpretation of the functional effects that such SVs may entail. Knowing exactly what coding parts of two genes have been fused enables us to predict whether the resulting fusion-gene is translated into a protein or not, as well as to estimate the structure and properties of the putatively new gene product. In relation to disease this helps to delineate between a fusion-gene's contribution by lack of expression of the two original genes, or by modified structure and binding properties respectively. Similarly, precise BP identification will offer more sensitive estimates of how SVs affect genomic functions. For instance, if an SV inserts, deletes or transposes (part of) a gene, it can result in a dysregulated processing of the respective gene. If the same SV additionally captures an enhancer region, the result could imply dysregulation of more genes[4].

Split reads allow for direct and base-pair-precise estimation of the SV breakpoints. Estimating breakpoint locations from discordant mate-pair alignments is approximate because it is impossible to know exactly where in the unsequenced part of the DNA fragment (i.e.: the insert), the breakpoint occurred. By aggregating information from many such discordant mate-pair alignments that span the SV and, potentially, split-reads, the breakpoint position is approximated to a genomic interval. In the case of long-read SV detection, only the direct, split-read information is available. Imperfect local mapping, stemming from sequencing errors or sequence homology between the two ends of a breakpoint junction still results in non exact BP junction coordinates, but the approximated region is typically much more precise. As depicted in **Figure 1**, in the patient data that we used for **chapter 4**, at the SV sites where NGS data only offers discordant mate-pair alignment data, the SV BP position is approximated by an average interval of 570 base-pairs and 580 base-pair (i.e.: by Manta or Delly respectively), whereas NanoSV is able to pinpoint the same BPs to a region spanning, on average, only 10 base-pairs (based on nanopore data). When split-read information is also available for a specific BP, its position is estimated more precisely from NGS data and narrowed down to an interval of 40 and 260 base-pairs, respectively (i.e.: Manta and Delly).

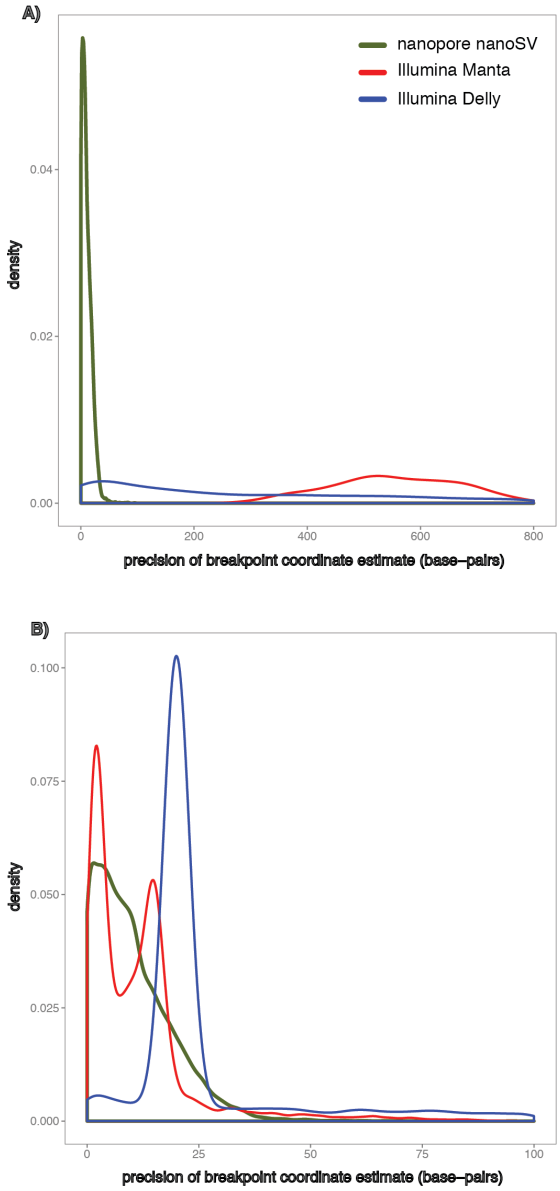


Figure1: Precision of estimating breakpoint junctions in nanopore data and illumina data respectively, in the Patient1 data analyzed in chapter 4.
A) the length of the interval to which a breakpoint is estimated, for SVs detected in both nanopore and illumina data (by Manta or Delly), for which the illumina data only contains discordant mate-pair information. B) the length of the interval to which a breakpoint is estimated, for all SVs detected in both nanopore and illumina data (by Manta or Delly).

Complex structural variation

Long read sequencing also serves to improve sensitivity in detecting complex genomic events. This is illustrated in **chapter 4,** where we are able to fine-map large scale chromosomal rearrangements and detect shorter translocated segments of a few hundred nucleobases that were missed in the original reconstruction of the respective patient's genome. While the initial short read investigation required the use of a special library preparation, that produced insert sizes in the order of kilobases in order to detect the long range structure of the chromothripsis, we detected this additional structure using standard long-read nanopore sequencing and a much lower sequencing coverage. Furthermore, we were able to describe the long range haplotype structure of many adjacent chromothripsis breakpoints through assembly, but also through individual reads that spanned multiple distant BP junctions (**chapter 4 - Supplementary Figure 16**). Lastly, many structural variants, such as SINE or LINE retrotransposons , are accompanied by adjacent (shorter) duplications, introduced during the repair of DNA double strand breaks. These convoluted events are hard to reconstruct from NGS data, and only one of the two connected structural variants is typically reported, but are readily detectable in the structure of the long reads. The manier and the longer the reads, the better and the more straightforward we will be able to reconstruct such complex structural variants.

Towards personal genomes

Consequences of reference based detection of genetic variation

As more genetic variation in general, and more complex structural variation in particular is identified between individuals, limitations of how all this variation is captured and represented became apparent. Since the first assembly of the human genome in 2001[5], it has been under continuous improvement and it is by far the best assembled mammalian genome, with only very few gaps and unsequenced regions left, estimated to represent ~5% of the whole genome[6]. The identification and characterization of genetic variation in the human population relies heavily on the use of the reference genome to compare the sequencing data of individuals' genomes against it. Although this enabled the identification of a tremendous degree of genetic variation, there is increasing concern about the biases that using a (single) genome as reference induces. First and foremost, the "reference allele bias" is the subtle, but systematic tendency to miss true genetic variation, where an individual carries an allele that is different than the reference allele, for some locus. This bias is introduced primarily during the read mapping step, when true variation in the reads, similar to sequencing errors, may impair alignment in genomic regions where an individual's data is divergent from the reference genome. Indeed, comparing whole genome sequencing results to the human reference genome has revealed a systematic bias in genotyping the highly polymorphic regions of the HLA genes[7]. Specifically, the allele frequency estimates resulting from these genotypes differ from the gold standard estimates of Sanger sequencing by more than 10% for about 25% of the SNPs within these genes. This effect is attributed to read mapping failures as,

predominantly, the allele frequency of the human reference genome allele is being overestimated. Furthermore, the human reference genome is a mosaic assembly of five individuals of caucasian descent and is representative of European ancestry[5]. It is therefore expected that alleles that are absent or rare in the European ancestry but potentially more common in other populations (i.e.: Asian, African, etc.) would be systematically missed, making the human reference genome a suboptimal lens to accurately describe all sequenced individuals.

Genome assembly as an alternative to read mapping

De novo genome assemblies are employed, as an alternative to read mapping altogether, in order to reconstruct the genome of an individual in an unbiased manner. Assembly methods reconstruct a sequenced genome by evaluating sequencing read overlaps, and extending individual reads to contigs (i.e.: haplotypes) that far exceed the length of the initial reads (i.e.: megabase long contigs can be obtained). The assembled contigs may then be aligned against the human reference genome to identify genetic variation. Beyond allowing for the incremental extension of contigs, read overlaps are used to build a higher confidence consensus sequence, by averaging the content of reads at overlapping positions, a feature that is most useful for erroneous long read data[8]. Furthermore, reads spanning heterozygous genetic variants help to locally delineate the two haplotypes of a diploid individual, provided that the sequencing data is of sufficient quality. *De novo* assemblies of African Yoruban and Asian individuals showed that while such approaches are feasible for detecting genetic variation, the resulting assemblies fall short of the human reference assembly and contain many gaps and uncaptured regions[9,10]. Recent approaches, relying on various long read technologies, have instead produced near complete assemblies[8,11] and showed extensive differences from the human reference genome, including many novel coding regions[11].

While assembly methods were successfully used to accurately reconstruct even the most polymorphic regions of the HLA region[12,13], a major limitation is their high computational requirements, making the approach unfeasible for the routine analysis of whole genomes. A hybrid approach can be proposed, where reads that align unequivocally to the human reference genome are used in a standard read-mapping manner, and unmapped or poorly mapped reads (i.e.: typically much fewer) and discordantly mapped reads are assembled separately and subsequently anchored to the human reference through alignment. A further limitation is that many assembly algorithms, particularly the ones used on long read data, do not delineate between the two haplotypes of an individual, but instead produce an arbitrary (high accuracy) consensus sequence of the two alleles of each locus. Genotypes can then be obtained by using the assembly result, as a personalized reference genome of the respective individual, to align all reads against, and call variants[8].

Population reference genomes

A solution to minimize the bias of using a reference to detect genetic variation, while making full use of the extensive knowledge generated thus far, is using a population reference instead of a single human reference genome (i.e.: which is built from five different genomes).

Models of population reference genomes start with the assembled human reference, onto which they can integrate (potentially all) genetic variation thus far discovered, including SNPs, short indels and structural variants. Early efforts to augment the canonical human reference with known, common variation include the introduction of 7 MHC alternative scaffolds in the Genome Reference Consortium (GRC) build37 of the reference genome, along with hundreds of other alternative scaffolds for various genomic loci[14]. Tools were developed that incorporate genetic variation information upon read mapping[15] or variant calling[16].

More systematic approaches use graphs to represent the structure of a population reference, where the primary assembly is used as a backbone, and every variation is represented as a different path[17]. This representation allows the integration of any event, including complex structural variation that breaks the linearity of the canonical reference genome. A genomic locus can therefore no longer be identified by a position on a specific chromosome. Because such graphs can become arbitrarily complex and there can be SNP or indel variation within large structural variation, a locus is defined by the genetic sequence context around it and localized hierarchically from structural alleles to base-pair substitutions[18]. Utilizing such complex genome representations will require a paradigm shift and the development of new tools, for alignment and variant calling that can handle and make full use of the new data structures[19,20] as well as new means of visualizing the results. The increased structural variant sensitivity and precision of long read sequencing can provide valuable sets of variants for describing population variation, that were previously hard to detect and/or integrate in the canonical reference genome.

Beyond the linear sequence: annotation with more layers of data

Although the complete representation of an individual's genome is a first prerequisite to enable understanding of genetic variations, biological interpretation of such variation requires much more data, such as gene expression, epigenetic data and genome folding.

Genome wide association studies (GWAS) were an early statistical means to relate genetic variation to a host of diseases and human traits. They did not necessarily require, nor explicitly produce, knowledge on the mechanisms through which the associated variation influences the observable outcome, but they offered a lens through which specific regions of the genome could be focused, for further experiments that would produce biological knowledge or explain disease etiology.

As genome sequencing evolved, became affordable and throughput increased, protocols were developed for interrogating features of the DNA molecule that influence the way it is functioning in a cell, such as 3D conformation of the molecule, base modifications such as methylation, as well as direct products of DNA, such as ribonucleic acids (RNA) and, sub-

sequently, proteins. These new fields produced a wealth of knowledge (and data), and it became apparent that the genome determines a cell's and organism's phenotype through a complex and dynamic system involving all these elements[21]. As the DNA sequence itself is (or could be) identical in all the cells of the body, different genes are expressed in different relative quantities across different tissues and cell states. Gene expression can be modulated by non-coding, sometimes distal regions of the genome (i.e.: enhancer regions), that attract or inhibit the transcription machinery. Dysregulation in the gene expression patterns of cells and tissues is a hallmark of cancer, as well as other diseases[22,23]. Genes and enhancers are organized into independent, local clusters of topologically associated domains (TADs), that may span several mega-bases. Complex interactions between multiple genes and multiple enhancers inside each TAD regulate the relative expression of each gene; these interactions can in turn help to interpret and to explain the contribution of non-coding variation to disease.The 3D conformation of the DNA molecule allows for distal genomic regions to be connected and plays a key role in defining TADs and modulation of expression. Disruption of the 3D folding of a DNA molecule, by short deletions in boundary regions of TADs, results in changed expression patterns that can also lead to cancerous behaviour of cells[24].

Genome conformation

Chromosome conformation capturing methods (3C) were first formulated and developed in the early 2000'[25]. DNA regions that may be distal on a chromosome, but in close 3d proximity when the molecule is folded, are artificially ligated together. The resulting fragments are a mosaic of different regions and after sequencing, genetic interactions manifest in the read data similarly to breakpoint junctions[26]. Since then, many variations, building on the same basic principle have emerged, where differences between each protocol include the type of genetic interactions that they are able to capture[27]. Most protocols zoom in on a target genomic region (i.e.: "viewpoint") and quantify long range interactions, of this viewpoint region, to a specific other region (3C), or to single but many/any other regions (4C) or attempt to fully characterise the pairwise interactions between a set of regions (multiple viewpoints - 5C). The Hi-C protocol can be used agnostically, without a predefined viewpoint to detect any and all interactions genome-wide and it was applied to produce a genome 3D interaction matrix at a resolution of ~1MB[28].

Recently, the Multi Contact 4C (MC-4C) protocol was used in conjunction with nanopore long read sequencing to characterise genetic interactions at three loci in the mouse genome[29]. The long nanopore reads captured on average 3-4, but up to 10 genomic interactions per read. Using this data, the authors were adequately powered to distinguish between cooperative 3-way interactions, where two other loci were co-interacting with the chosen viewpoint, such as genes coding for different components of a protein complex. Competing 3-way interactions, where two other loci were interacting with the viewpoint but not at the same time (i.e.: not in the same cell) were characterised between a gene and one of more possible enhancers, as well as random 3-way interactions[29]. Interestingly, arbitrarily complex genomic interactions could theoretically be characterized in this manner, involving many loci simulta-

neously. However, the number of possible multi contact interactions grows exponentially with the number of loci considered so being adequately powered to accurately describe multi-locus genetic interactions would require an increasingly higher depth of coverage.

RNA SEQUENCING AND EXPRESSION

Characterizing gene expression patterns offers a wealth of information, regarding the relative expression levels of genes across different tissues, but also functionally relevant splicing variation, exons that are skipped or retained introns, fusion genes or antisense transcription[30,31]. Current high throughput technologies involve specialized protocols to measure gene expression, such as RNAseq[30], followed by short read NGS sequencing. They have enabled the routine interrogation of the transcriptomics but present a couple of shortcomings, that are, to some extent, mitigated through complex downstream analysis. First of all, the library preparation for sequencing RNA is more complex, involving bio-chemical manipulation of RNA extracted from cells such as reverse transcription, which is susceptible to errors. Furthermore, PCR may introduce amplification biases that are particularly relevant as quantification of relative amounts of different transcripts is usually pursued, as well as selection biases (i.e.: underrepresentation of GC rich sequences). Lastly, while the short reads may be directly used to quantify abundance of transcripts, detecting splice variation and/or fusion genes is much harder, because of the short transcript snippet that the read captures.

Long read nanopore sequencing has the potential to enable both simplified protocols and more sensitive results with direct RNA sequencing. Because the RNA molecule can pass the nanopore just like the single stranded DNA molecule, the only difference in sequencing the two molecules is properly calibrating the subsequent base calling algorithms. Library preparation biases, such as reverse transcription, may be biased altogether. Similarly to DNA library preparation, it only requires isolating the desired RNA and adding sequencing adaptors and motor proteins. A PCR can be performed but it is not necessary if sufficient RNA is available, thus enabling a more sensitive estimation of the relative quantity of sequenced transcripts. Nanopore transcriptome sequencing in mice[32] and yeast[33] cells shows a high correlation to traditional NGS approaches (~0.8). Furthermore, despite the lower per-base sequencing accuracy, RNA transcripts of highly similar, paralogous genes are correctly discriminated[33].

Because the long reads tend to span entire transcripts[33], isomorphic transcripts, as well as gene-fusions can be readily detected by mapping to the reference genome (i.e.: or transcriptome), similarly to structural variants[32,33], although attempting to reconstruct these transcripts through assembly of the raw read data produced much less accurate results [33]. Read mapping however, identified thousands of novel splice variants as well as complex transcript isoforms in a small population of just seven cells[32]. With increased sensitivity to detect and quantify differentially transcribed genes, the relation between genetic content around a gene and transcriptional variation could be quantified, similarly to expression quantitative trait loci (eQTL) analyses that have related genetic variation to variation in gene expression levels,

thus further aiding the interpretation of non-coding genetic variants.

Base modifications

Base modifications are yet another layer of data, important to functionally understand the genome, that could now be directly retrieved through nanopore sequencing. Accurate detection of 5-methylcytosine (5-mC) in DNA molecules is now possible[32]. The potential to detect other base modifications such as DNA 5-hydroxymethylcytosine (5-hmC)[34] and RNA N^6-methyladenosine (m^6A) and 5-methylcytosine (5-mC)[33] has also been suggested. DNA bases can suffer one of many such epigenetic modifications, or markers, which can influence how DNA is processed (e.g.: transcribed)[35]. As any chemical modification of a nucleobase would change its physical and electrical properties, theoretically, any modified base would modulate the ionic current measurement through the pore slightly differently, as it traverses the pore. On the other hand, the task of translating the electrical trace to a nucleotide sequence (i.e.: basecalling) will become increasingly computationally intensive, as one or more variations of each base are considered. I believe that detecting an arbitrary number of base modifications will turn out computationally demanding and it remains to be seen if the ionic current measurements are sensitive enough to allow for such simultaneous detection. Alternatively, since base-calling is a modular, stand-alone step with nanopore sequencing, one could build separate models that detect various base modifications. Each model can be run on the same raw data and a union and consensus of the calls from different models can be obtained.

Short read or long read sequencing?

After the completion of the Human Genome Project[5], high throughput Next Generation Sequencing technologies have spiraled into a revolution of human (and other organisms) genetics[36,37]. They facilitated the sequencing of an unprecedented number of genomes and exomes[38–41], that became an invaluable resource to common and rare disease studies and to understanding cellular processes. The need for analysis and interpretation of the vast amounts of data generated from NGS created and established the field of bioinformatics and created a new paradigm for biology research, heavily reliant on interdisciplinary groups of researchers. Furthermore, there is increasing adoption of NGS testing in the clinic[42–44] and clinical geneticist is already a medical specialization in many countries.

NGS technologies encompass a mature and reliable set of tools that remain unchallenged in many critical applications. Accurately detecting *de novo* SNPs and indels is a challenging task even within the relatively more accurate NGS data, as I show in **chapter 2**, as these events occur at a much lower rate than the sequencing error rates of NGS data. Furthermore, for the identification of short variation such as SNPs and short indels, the per-base sequencing quality is crucial and the relatively higher error rates of long read data currently prohibit these applications. This makes NGS methods the prefered choice, for general applications when accurately detecting *most* of the genetic variants of an individual is of

interest. NGS data is also the currently preferred method, for most applications that require high quality SNP and indel calling, such as clinical applications, where confidence in the set of variants which facilitate diagnosis is crucial, or MHC sequencing, where the polymorphic sites are extremely dense and miss genotyped variants easily introduce ambiguities.

If, instead, there is reasonable a priori knowledge to assume that structural variation plays an important role in the question or disease of interest, such as autism spectrum disorders, neurodevelopmental disorders or severe congenital abnormalities, long read nanopore sequencing could offer a more informative insight into potentially causal (structural) variants. While long read data does not currently allow accurate genome-wide interrogation for SNP variation (**chapter 4**), a high genotyping accuracy can be achieved if SNV sites are a priori known/chosen. The existing knowledge about population variation can be thus leveraged, by using long reads to genotype an individual only at positions that are known to be variant in the general population (i.e.: 1KG sites, GoNL sites, etc.), thus minimizing the rate of false positive SNVs are extracting the most information possible from the long reads.

In the context of cancer genome sequencing, SNP mutational signatures were found to be informative for underlying mechanisms[45] and to differ between different types of cancer [45,46]. Detection of variants in these cases is further polluted by genetic heterogeneity within a population of cancerous cells and the need to estimate tumour variants from a mixture of cancerous and healthy cells. Pending further increases in sequencing accuracy, it seems unlikely that long read sequencing can yet produce sensible sets of (SNP) variants for cancer genetics. For cancer types that are associated with high genomic instability, such as prostate cancer, long read data can be used to characterize SVs and complex chromosomal rearrangements. Furthermore, long read data could be a solution to expression analysis of cancer genomes and the identification of fusion genes, as described above.

I believe however, that, on a shorter term, long read sequencing will prove very valuable for a much more comprehensive characterization of genetic structural variation and will also facilitate the transition towards personal genomes. We and others have benchmarked the power and accuracy of nanopore sequencing against state of the art datasets of the Genome In A Bottle (GIAB) consortium, and further nanopore sequencing of GIAB samples is being performed. Furthermore, they could be instrumental for a deeper understanding of the consequences of genetic variation and of how gene expression patterns and genomic 3D conformation influences cellular processes or disease.

Reporting negative findings

Finally, I wish to briefly emphasize the need for transparent research an the value of negative research results, best illustrated by **chapter 3**, in the context of an ever increasing number of association studies being published. The number of genetic and epidemiologic association studies published increased (dramatically) over the past decades. Subsequent replication of initially reported associations however, came to divergent and even contradictory conclusions, to a proportion that is way beyond the expected 5% false discovery rate (FDR) which is typically aimed for. Up to 16% of associations that were estimated to have had a high impact in the medical field, were contradicted by subsequent studies, and another 16% reported effect sizes that were not matched during replication[47]. Similarly, in genetics, up to 72% of initially reported associations are subsequently found to be overestimates[48]. A review of the causes and consequences pertaining to these facts is beyond the purpose and work presented in this thesis. Beyond cases of provable scientific misconduct or analysis mistakes, there are limitations and biases arising from statistical sampling or adequate power, but also a bias from towards only publishing positive results (i.e.: meaning claims of association, as opposed to claims of no association), from both scientists and journals, which together may lead to a grey area of interpreting and valuing results[49,50]. Some of the suggested solutions to overcoming biases are rather ad-hoc and serve to perform "damage control" rather than improve the outcome directly (i.e.: such as trusting results on repeatedly interrogated questions less)[51]. I believe that reporting negative findings can be extremely valuable in this context. Negative replication findings are obviously valuable to update our knowledge base, especially when they have increased power to detect a putative signal and novel data, as show in **chapter 3**. Novel, or secondary negative findings are however, also very valuable, especially as more data dimensions (i.e.: expression, conformation) are added to the picture and the search space expands exponentially. Many more questions can and will be investigated, and proper reporting of any answer to a relevant question will help us understand the data better. Together with positive findings, negative findings form the complete picture of what we know at a certain point in time. Furthermore, as the main means of research and producing knowledge are currently statistical evaluations of hypothesis, the value of the research lies more in asking informative questions and employing the correct methodologies, on the right data to answer these questions, rather than solely in a positive answer. By definition, it cannot be a priori known whether a hypothesis is true or false (i.e.: hence the testing). Generating a sensible hypothesis to be tested, thus requires the use of a certain degree of intuition. This intuition is best served when all the possibly relevant facts are evaluated, including what is known to be true and what is known to be false.

References

1. Laver, T. *et al.* Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif* **3,** 1–8 (2015).

2. Carter, J.-M. & Hussain, S. Robust long-read native DNA sequencing using the ONT CsgG Nanopore system. *Wellcome Open Res* **2,** 23 (2017).

3. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv* 193144 (2017). doi:10.1101/193144

4. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14,** 125–138 (2013).

5. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

6. Altemose, N., Miga, K. H., Maggioni, M. & Willard, H. F. Genomic Characterization of Large Heterochromatic Gaps in the Human Genome Assembly. *PLoS Comput. Biol.* **10,** e1003628 (2014).

7. Brandt, D. Y. C. *et al.* Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3* **5,** 931–941 (2015).

8. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **14,** S18 (2018).

9. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–272 (2010).

10. Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8,** 61–65 (2011).

11. Cao, H. *et al.* De novo assembly of a haplotype-resolved human genome. *Nat. Biotechnol.* **33,** 617–622 (2015).

12. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44,** 226–232 (2012).

13. Jensen, J. M. *et al.* Assembly and analysis of 100 full MHC haplotypes from the Danish population. *Genome Res.* **27,** 1597–1607 (2017).

14. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol.* **9,** e1001091 (2011).

15. Huang, L., Popic, V. & Batzoglou, S. Short read alignment with populations of genomes. *Bioinformatics* **29,** i361–70 (2013).

16. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R. & McVean, G. Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* **47,** 682–688 (2015).

17. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of

genome inference. *Genome Res.* **27,** 665–676 (2017).

18. Rand, K. D. *et al.* Coordinates and intervals in graph-based reference genomes. *BMC Bioinformatics* **18,** 263 (2017).

19. Novak, A. M. *et al.* Genome Graphs. *bioRxiv* 101378 (2017). doi:10.1101/101378

20. Paten, B., Novak, A. M., Garrison, E. & Hickey, G. Superbubbles, Ultrabubbles and Cacti. in *Research in Computational Molecular Biology* 173–189 (Springer, Cham, 2017).

21. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).

22. Dermitzakis, E. T. From gene expression to disease risk. *Nat. Genet.* **40,** 492–493 (2008).

23. Cancer gene expression signatures – The rise and fall? *Eur. J. Cancer* **49,** 2000–2009 (2013).

24. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351,** 1454–1458 (2016).

25. Dekker, J. Capturing Chromosome Conformation. *Science* **295,** 1306–1311 (2002).

26. 4C Technology: Protocols and Data Analysis. in *Methods in Enzymology* **513,** 89–112 (Academic Press, 2012).

27. Denker, A. & de Laat, W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev.* **30,** 1357–1382 (2016).

28. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).

29. Allahyar, A. *et al.* Locus-Specific Enhancer Hubs And Architectural Loop Collisions Uncovered From Single Allele DNA Topologies. *bioRxiv* 206094 (2017). doi:10.1101/206094

30. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

31. Wu, J. Q. *et al.* Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome. *Genome Biol.* **9,** R3 (2008).

32. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8,** 16027 (2017).

33. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* (2018). doi:10.1038/nmeth.4577

34. Laszlo, A. H. *et al.* Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 18904–18909 (2013).

35. Viner, C. *et al.* Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet. *bioRxiv* 043794 (2016). doi:10.1101/043794

36. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-gener-

ation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–351 (2016).

37. Cao, Y., Fanning, S., Proos, S., Jordan, K. & Srikumar, S. A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies. *Front. Microbiol.* **8,** 1829 (2017).

38. Maretty, L. *et al.* Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548,** 87–91 (2017).

39. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

40. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–825 (2014).

41. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

42. Yohe, S. & Thyagarajan, B. Review of Clinical Next-Generation Sequencing. *Arch. Pathol. Lab. Med.* **141,** 1544–1557 (2017).

43. de Ligt, J. *et al.* Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *N. Engl. J. Med.* **367,** 1921–1929 (2012).

44. Yohe, S. *et al.* Clinical validation of targeted next-generation sequencing for inherited disorders. *Arch. Pathol. Lab. Med.* **139,** 204–210 (2015).

45. Jager, M. *et al.* Deficiency of global genome nucleotide excision repair explains mutational signature observed in cancer. *bioRxiv* 221168 (2017). doi:10.1101/221168

46. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24,** 52–60 (2014).

47. Ioannidis, J. P. A. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* **294,** 218–228 (2005).

48. Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nat. Genet.* **29,** 306–309 (2001).

49. Goodman, S. & Greenland, S. Why Most Published Research Findings Are False: Problems in the Analysis. *PLoS Med.* **4,** e168 (2007).

50. Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Med.* **2,** e124 (2005).

51. Ioannidis, J. P. A. Why Most Published Research Findings Are False: Author's Reply to Goodman and Greenland. *PLoS Med.* **4,** e215 (2007).

English Summary

The study of human genetics was greatly facilitated by the sequencing of the first human genome in 2001. A race to develop and perfectionize DNA sequencing technologies and data analysis followed this milestone project, that has enabled the sequencing of thousands of human genomes since.

Based on the sequencing data from many human genomes, gathered through consortia such as the thousand genome project and the Genome of the Netherlands, an average human genome was found to vary at a few million loci compared to the genome of an unrelated human individual. Currently, roughly ~100 million genetic variations have been found so far, but new variation is discovered with every sequenced genome. Thousands of genetic variants were associated to common and/or rare disease. The processes through which genetic variation results in disease are sometimes linked directly to altering one of the ~20,000 known genes' product content or abundance and have even enabled new therapies. In many cases however, the functional consequences of genetic variation were hard to identify precisely. These functional effects could be further explained by relating the genetic variation to more distal regions that interact with a gene or by affecting DNA organization and conformation.

While information about the sequence content as well as about many other relevant DNA features (such as conformation and regulation) may be retrieved through sequencing, the type of different sequencing technology eventually used can have a significant impact on results. Thus, current sequencing technologies that produce short, but highly accurate read-outs of the genome are successfully employed to determine the genetic content of most loci in the genome. Analyzing more complex structural variation within a genome, or reconstructing regions of a genome however, requires long-range information that is cumbersome, to obtain from the short read-outs. Alternative technologies have emerged, that are able to produce very large read-outs of our genome and can offer the information necessary to reconstruct complex regions. These longer read-outs are currently, relatively more erroneous, making the analysis of short genetic variation very hard.

My work in this thesis concerns the development of appropriate methodologies to accurately extract and value all the information that state of the art sequencing technologies produce, and I show how different sequencing technologies are best suited for interrogating the human genome for different types of variation and information.

In chapter 2, we show how highly accurate short read sequencing technologies can be leveraged to accurately identify de novo mutations. These mutations, that are present in low numbers in any individual, but not in his or her parents, are the rarest form of genetic varia-

tion and represent the fuel for subsequent natural selection and genetic drift and are shown to cause many genetics diseases, such as intellectual disability or autism. We developed an algorithm to identify de novo mutations from the sequencing of family trio's, i.e. a father, mother and their child. We apply our tool to accurately identify all de novo mutations on chromosome X in the offspring of 250 Dutch trios, that were sequenced as part of the Genome of The Netherlands project.

We further use the genetic data of the same 250 Dutch trios in chapter 3 to answer the long-standing hypothesis of preferential mating. Humans as well as other animals show an odor preference towards individuals that have different genetic content than their own across a region that contains many genes involved in the immune response. Mating between such diverse individuals would then produce children that are genetically better equipped to respond to disease. Previous tests of the hypothesis of preferential mating based on immune-related gene diversity were inconsistent between studies, also due to only partly sequencing, or inferring the genetic content of the region of interest. I show how direct sequencing of all the relevant genetic variation in the region of interest and appropriate quality controls enable an unequivocal rejection of this hypothesis in Dutch ancestry humans.

In chapter 4, We focused on new, emerging nanopore sequencing technologies. This technology measures the electrical signature of a DNA molecule passing through a pore and is able to read much larger snippets of an individual's DNA than before. We show how this novel technology can be leveraged to reconstruct complex genomic re-arrangements in patients with genetic disease. Furthermore, we demonstrate that with an optimized, but simple methodology, we are now able to identify large genetic variations (i.e.: structural variation) throughout the genome more easily and more accurately than using the standard approaches. We find that nanopore sequencing enables the interrogation of previously less accessible regions of the genome, such as regions containing a high proportion of C and G bases, thus enabling the discovery of yet novel variation. Furthermore, we find that medium sized structural variation, that is known to be under-represented in human variation catalogues, is now readily accessible. We accurately map medium and large insertions of repeated DNA elements that were previously hard to accurately place in the human genome, but that are estimated to have historically contributed to ~40% of human genetic variation.

Overall, this thesis illustrates how using the appropriate methodology and technology is key for reaching accurate and clear conclusions from large amounts of genetic data useful both in a research and in a diagnostic setting. Short-read accurate sequencing technologies are a benchmark for small and/or rare genetic variation, whereas emerging long-read technologies are perfectly suited for larger, structural variation. Furthermore, by reading longer stretches of DNA, nanopore sequencing may be instrumental for understanding functional consequences of genetic variation and facilitate data integration and a paradigm shift towards analyzing an individual's genome in its entirety.

SAMENVATTING

Het sequensen van het eerste humane genoom in 2001 was van groot belang voor de studie van de humane genetica. Op deze mijlpaal volgde een race om de DNA sequencing technologie en data-analyse verder te ontwikkelen en vervolmaken, wat ons in staat heeft gesteld om vele duizenden humane genomen te sequensen.

Op basis van de sequentie-data van het genoom van vele mensen, verzameld door middel van consortia als het Thousand Genome Project en het Genome of The Netherlands, blijkt een gemiddeld genoom op een paar miljoen plekken te verschillen met het genoom van een onverwant persoon. Tot nu toe zijn er circa 100 miljoen variërende posities gevonden, en met elk nieuw onderzocht genoom worden er nieuwe gevonden. Duizenden van deze variaties worden geassocieerd met veelvoorkomende en/of zeldzame ziektes. De manier waarop genoom-variatie leidt tot ziektes kan soms direct terug worden gebracht tot een verandering in een van de ~20.000 genen, en dit heeft al geleid tot het ontstaan van nieuwe therapieën. Echter, in veel gevallen is het lastig om de consequenties van genoom-variatie aan te duiden. De functionele effecten zouden beter begrepen kunnen worden door ver weg gelegen genoom-regio's, die genen reguleren, erbij te betrekken. Ook rekening houden met de driedimensionale structuur van DNA zou ons meer kunnen leren.

Informatie over de sequentie en andere DNA-kenmerken (zoals de driedimensionale structuur) kunnen worden verkregen uit het sequensen van het genoom, maar de precieze technologie die wordt gebruikt kan een groot effect hebben op de resultaten. De huidige sequensers, die korte DNA-fragmenten lezen maar dit zeer nauwkeurig doen, worden succesvol gebruikt om de DNA-sequentie op de meeste plekken in het genoom te lezen. Complexere structurele genoom-variatie vraagt om langere DNA-fragmenten die in een keer gelezen worden. Alternatieve technologieën zijn in staat om heel lange DNA-fragmenten te lezen, en bieden de informatie die nodig is om complexe regio's te reconstrueren. Deze langere fragmenten zijn echter vaker onjuist afgelezen, wat de analyse lastig maakt.

Mijn werk in dit proefschrift gaat over het ontwikkelen van passende methodes die op accurate wijze alle informatie kunnen halen uit de data geproduceerd door de nieuwste sequencing technologieën. Ik laat zien hoe verschillende sequencing technologiën het best in staat zijn verschillende types informatie te bemachtigen uit het humane genoom.

In hoofdstuk 2 laat ik zien hoe heel nauwkeurige sequencing technologieën, die korte DNA-fragmenten lezen, gebruikt kunnen worden om accuraat de novo mutaties te identificeren. Deze mutaties, die in elk individu op lage frequentie voorkomen maar niet in het genoom van de ouders, zijn de meest zeldzame vorm van genetische variatie en vormen de

basis voor natuurlijke selectie en genetische drift. Ook is er gevonden dat ze de oorzaak zijn van vele genetische aandoeningen, zoals verstandelijke handicaps en autisme. Wij hebben een algoritme ontwikkeld dat de novo mutaties ontdekt in de DNA-sequenties van familie-trio's (een vader, een moeder en hun kind). We passen onze methode toe op de nakomelingen van 250 Nederlandse trio's, onderdeel van het project Genoom van Nederland, om op accurate wijze alle X-chromosomale de novo mutaties te identificeren.

Ik gebruik de genetische data van dezelfde 250 Nederlandse trio's in hoofdstuk 3 om de al lang bestaande hypothese over preferentiële partnerkeuze te beantwoorden. Mensen, net als andere dieren, laten een voorkeur zien voor de geur van individuen die een andere genetische sequentie hebben dan zijzelf (in een genoom-regio die betrokken is bij het immuunsysteem). Gemeenschap tussen twee van zulke diverse individuen zou kinderen voortbrengen die beter in staat zijn om te reageren op ziektes. Eerdere testen van de hypothese van preferentiële partnerkeuze gebaseerd op immuun-gerelateerde gen diversiteit gaven inconsistente resultaten, ook doordat de genetische inhoud van de regio van belang maar deels gesequensed of afgeleid werd. Ik laat zien hoe het direct sequensen van alle relevante genetische variatie in de regio van belang en het gebruiken van geschikte kwaliteitscontroles ons in staat stellen deze hypothese zonder twijfel van de hand te wijzen in mensen van Nederlandse herkomst.

In hoofdstuk 4 focussen we op de nieuwe, opkomende nanopore sequencing technologieën. Deze technologie meet de elektrische handtekening van een DNA-molecuul terwijl het zich door een porie beweegt en is in staat veel langere stukken van iemands DNA te lezen dan voorheen. We laten zien hoe deze nieuwe technologie gebruikt kan worden om complexe genomische herschikkingen te reconstrueren in patiënten met een genetische ziekte. Verder laten we ook zien dat met een geoptimaliseerde, maar simpele, methodologie we nu in staat zijn grote genetische variaties (ofwel: structurele variaties) door het hele genoom makkelijker en accurater te identificeren dan met de standaard aanpak. We constateren dat nanopore sequencing ons in staat stelt om voorheen ontoegankelijke regio's van het genoom te onderzoeken, zoals regio's die een hoge proportie van C en G basen bevatten. Hierdoor kunnen we nog meer nieuwe variatie ontdekken. Ook de structurele variaties van gemiddelde grootte, waarvan we weten dat ze ondervertegenwoordigd zijn in catalogussen van humane variatie, zijn nu toegankelijk. We kunnen op accurate wijze gemiddelde en grote inserties van herhaalde DNA-elementen plaatsen. Voorheen was het lastig deze op de accurate plek in het humane genoom te plaatsen, maar het wordt geschat dat deze historisch gezien voor ~40% hebben bijgedragen aan de humane genetische variatie.

Dit proefschrift laat over het geheel zien dat het gebruik van de gepaste methodologie en technologie de sleutel is om accurate en duidelijke conclusies te kunnen trekken over de grote hoeveelheid genetische data, zowel in de context van onderzoek als in de context van diagnostiek. Technologieën die korte DNA-fragmenten lezen zijn een ijkpunt voor kleine of zeldzame genetische variaties, terwijl de opkomende technologieën die lange DNA-frag-

menten lezen perfect geschikt zijn voor de grotere, structurele variatie. Door het lezen van langere stukken DNA, kan nanopore sequencing bovendien van groot belang zijn voor het begrijpen van de functionele consequenties van genetische variatie. Het kan data integratie faciliteren en het kan leiden tot een paradigmaverschuiving naar het in het geheel analyseren van het genoom van een individu.

List of publications

A framework for the detection of de novo mutations in family-based sequencing data

Laurent C Francioli*, Mircea Cretu-Stancu*, Kiran V Garimella, Menachem Fromer, Wigard P Kloosterman, Genome of the Netherlands Consortium, Kaitlin E Samocha, Benjamin M Neale, Mark J Daly, Eric Banks, Mark A DePristo, Paul IW de Bakker

European Journal of Human Genetics volume 25, pages 227–233 (2017), doi:10.1038/ejhg.2016.147

Mapping and phasing of structural variation in patient genomes using nanopore sequencing

Mircea Cretu Stancu*, Markus J. van Roosmalen*, Ivo Renkens, Marleen M. Nieboer, Sjors Middelkamp, Joep de Ligt, Giulia Pregno, Daniela Giachino, Giorgia Mandrile, Jose Espejo Valle-Inclan, Jerome Korzelius, Ewart de Bruijn, Edwin Cuppen, Michael E. Talkowski, Tobias Marschall, Jeroen de Ridder, Wigard P. Kloosterman

Nature Communicationsvolume 8, Article number: 1326 (2017) doi:10.1038/s41467-017-01343-4

Lower frequency of the HLA-G UTR-4 haplotype in women with unexplained recurrent miscarriage – Journal of Reproductive Immunology

T. Meuleman,, J. Drabbels, J.M.M. van Lith, O.M. Dekkers, E.Rozemuller , M. Cretu-Stancu, F.H.J. Claas, K.W.M. Bloemenkamp, M. Eikmans

Journal of Reproductive Immunology, Volume 126, April 2018, Pages 46-52, https://doi.org/10.1016/j.jri.2018.02.002

Functionally distinct ERAP1 and ERAP2 are a hallmark of HLA-A29-(Birdshot) Uveitis.

Kuiper JJW, Setten J, Devall M, Cretu-Stancu M, Hiddingh S, Ophoff RA, Rothova A, Missotten TOAR, van Velthoven M, Den Hollander AI, Hoyng CB, James E, Reeves E, Martín J, Koeleman BPC, de Boer JH, Pulit SL, Martinez A, Radstake TRDJ

Mnuscript submitted to Proceedings of the National Academy of Sciences of the United States of America

No evidence that mate choice in humans is dependent on the MHC

Mircea Cretu-Stancu, Wigard P. Kloosterman, Sara L. Pulit

Manuscript submitted to PLoS Genetics

** Authors contributed equally to the work*

Acknowledgements

I have a lot of people to be grateful to, for guiding and supporting (and baring with) me in producing this thesis. The last 5 years, since I started as a master student in Paul's group and now, graduating as a PhD student in Wigard's group, were a roller coaster (at times steep) into a new academic field and on all levels. Trying to look back on it, it sometimes feels like 2 weeks have passed and other times it feels like 20 years, so I will try to piece together my experience by thanking all the people that generously shared with me a little of themselves along this path.

First off, thank you to my promotor, Prof. **Edwin Cuppen** and my copromotor, Dr. **Wigard Kloosterman** for supervision and guidance and to Prof. **Paul de Bakker** for supervision during the first half of my PhD and financial support throughout.

Thank you to the members of my reading committee, Prof. **Dick de Ridder**, Prof. **Alex Schoenhuth**, Prof. **Berend Snel**, Prof. **Frank Holstege** and Prof. **Wouter de Laat** for reading and evaluating my work and Prof. **Frank Holstege** and Prof. **Wouter de Laat** for their critical feedback and support, as part of my supervisory committee, especially towards the end of my PhD.

I would first like to thank Dr. **Ad Feelders** for finding "some other kinds" of projects, for my otherwise Computer Science, Masters' thesis, and thus sending me off to the UMC.

To Dr. **Laurent Francioli** thank you for your supervision during my master project and for your always honest (..and frank) and well taken advice as a colleague and friend. I basically learned all about Next Generation Sequencing data analysis from you, in any case much more than was needed for my project. I learned how to adapt abstract models to obtain reproducible and defensible results and how to make my work and algorithms useful for the community. But I also learned to put the necessary time into presenting my results clearly and making a good presentation, and relevant details such as not being 10 minutes late for the Thursday department meeting every other time (…so my attendance halved for a short while after that). You 'Mr. Myagy'-ed a set of very useful research skills, teaching by showing, rather than by telling. Furthermore, you fitted my supervision into your schedule of constantly travelling and commuting between the Netherlands and the UK, which was a perfect preparation for Paul's schedule.

To Prof. **Paul de Bakker**, thank you for giving me the chance to learn so much and grow, as a PhD candidate in your group, and for being a resourceful and truly inspiring supervisor. Every meeting we were having was different and surprising, and many times It felt like I was

meeting you again all over. I remember being frustrated after a few of our regular meetings that I would not find a common language to get feedback for my HLA-typing algorithm; after organizing my half-baked results into a figure you reverse engineered all of my algorithm offering more answers and suggestions than I would have hoped for. I learned how to be clear (and precise) and how to question everything, without losing sight of the goal. Whether we were talking about a direct next step or the larger scope of my project, despite building no expectations about the outcome, I always left motivated and feeling clear about what needs to happen next. Furthermore, you always knew the best person to talk to, for any particular problem, and this enabled me to talk to very interesting people and get in depth understanding of many genetics and statistics topics. Thank you so much for looking after my PhD even after you had left the UMC.

To Dr. **Wigard Kloosterman**, thank you for being a tireless supervisor, for showing me how to truly enjoy the research that I do and for offering me a new direction, when I myself didn't fully realize that I was up in the air. I think the first time I met you was shortly into my PhD, when you asked whether we can look for de novo mutations in one of your samples' data. Excited about a first "collaboration" with a different PI, I dropped the rest of the day and produced the desired mutations, along with some suggestions on how they could associate to chromothripsis (and solve it forever). At the end of the day, after glimpsing at my results and seeing that the mutations were not near the region of interest, you decided it was not relevant to pursue this further, and I was hoping no-one noticed that I just wasted half a day for nothing. By far the most inspired waste of half a day yet !! Thank you so much for adopting me into your group and for almost immediately treating me like I have been working with you since the beginning. You have a been a great supervisor, going along with my ideas of how we could best make sense of our data and being pragmatic and micro-managing where I would stubbornly go into too many different directions. You were very open and invested time and thought to help me to continue or explore previous projects (even travelled to the Sanger to see how we can make it work). Thank you for the chance to work with new sequencing technologies and on cool projects, which proved very interesting and fruitful. I have learned a lot from you, about structural variation, and many other different types of questioned that sequencing data analysis can answer! Thank you for helping me scope the work of my thesis, for all the support and guidance in writing it and for not (completely) dropping fate whenever I was coming (razor-blade) close to deadlines.

To Dr. **Sara Pulit**, thank you for your guidance and mentorship! You showed me how to be rigorous and clear in my analyses and how to avoid being sloppy both in research and outside of it. Thank you for all the 'tough love' (never felt the toughness!), for watching over me and making me take time management seriously!

To Prof. **Edwin Cuppen**, thank you accepting to be my promotor at a complicated and confusing time, especially after having talked, outside of our department meetings, roughly twice. Thank you for always making yourself approachable to me!

To the whole Wizard family, the Kloosterman group, thank you for combining work and pleasure so seamlessly and productively. After many educative lab-meetings I feel I can reason (or at least not feel like a martian) when I hear a molecular focused talk. To the double-trouble team, **Chris**, **Christina**, **Jose**, thank you for your friendship and the many interesting PhD events where we went together, you helped me structure and deal with aspects of a PhD that I was comfortably (or not) being ignorant to; and for many more events outside the PhD! Chris, I find your enthusiasm contagious and productive. Christina, thank you for being a voice (and sword) of justice and Jose thank you for always being ready to brainstorm about cool nanopore ideas and for covering for me when I forget to pack random stuff for travelling. **Mark**, thank you for your collaboration and enthusiasm, and for making me doubt my coding skills; where I would think and model 5 times before I implement something, you would have a working script in no-time. **Ivo**, thank you for making me feel I understand what happens before I get my fasta file from the sequencer, and also for testing my HLA primers. **Alessio**, I owe you every pixel on my cover (on a 5-minute notice !!) and many insightful explanations of the immune system and biological engineering. **Glen** thank you for the well-timed 'frosties' and good advice. **Ellen**, thank you for all the thesis suggestions (and warnings) and **Mirjam**, **Joline**, **Carl** and **Mark**, thank you for the interesting conversations about related, or far away fields of science. To **Marloes** and **Yasmine** thank you for contributing to my projects through your internships and to **Anne**, **Camilla**, **Tamara**, **Dide**, **Aleks, Irene**, **Marijam**, **Ewout** and many more wizard students (probably 90% of which supervised by Alessio) thank you for the gezelligheid that you brought to our group. I enjoyed our fun rotating dinners and I think we are ready to climb any mountain together (hopefully including Alpe d'Huzes)!

To the former de Bakker lab, **Sara**, **Androniki**, **Laurent**, **Vinicious**, **Jessica**, **Jytte**, **Daiane**, **Daniel**, **Sander**, **Balder**, **Maarten**, **Kristel**, and **Clara**, thank you for introducing me to the field of genetics and teaching me so much about genetics and statistics in and outside our lab meetings. It was a stimulating and ambitious environment and you were always open and enthusiastic to talk about research and to think along! I enjoyed our lab meetings as well as our retreats into the woods. Even after we spread in many directions, you were always available, academically and personally and I felt I had my back covered. **Jessica**, thank you for referring many GoNL questions my way, although I was never on the GoNL project itself and for tips on my thesis layout. **Daiane** and **Daniel**, thank you for bearing together through the pain of uncertain projects.

To Prof. **Alex Schoenhuth** and **Jasmijn Baaijens**, thank you for our long collaboration, I hope we can have a new HLA-typing algorithm soon.

To **Jonas**, thank you for complex and captivating talks about genetic interactions and the HLA in general.

To Dr. **Jeroen de Ridder** and the de Ridder group, thank you for making bioinformatics a stand-alone topic in the Genetics department as well as for many interesting algorithmic and

machine learning talks. I hadn't talked about these topics so long that I almost forgot I had actually studied them. To **Joske** and **Johanna** (should I say from the Pulit group) thank you for translating my thesis summary, with great care for proper Dutch and to **Joep**, thank you for throwing at me interestingly cryptic ideas about cover design. To **Joske**, **Amin** and **Roy**, thank you for very interesting 'quick' algorithmic chats or debates that would quickly claim two hours of our day.

To the Cuppen group, thank you for interesting talks about whole genome sequencing and structural variation. To Dr. **Booby Koeleman** and Dr. **Gijs van Haaften**, thank you for the teaching experience and to the Koeleman and van Haaften groups thank you for very interesting lunch breaks and nice genetics retreats.

Big thank you to **Monique**, for calmly putting up with me and not letting me become sure prey to any of the forms and requirements that I had to meet and to **Cristina Arpesella** for vigilently looking after my CMM duties.

To **Renee** for very productive brainstorming on paper and thesis titles and to **Jesse** for good distractive breaks and for providing me a steady supply of coffee while I was writing my thesis.

To **Kirsten**, **Nayia**, **Terry**, **Sakshi**, **Wout** and **Marijn**, thank you for many interesting and surprising talks about research and not only. To **Ellen Carbo**, thank you for the short breaks and for explaining to me how the UMC is organized (many times over).

To everyone in the Genetics department, thank you for always stepping up to help, whenever intention was not backed up by actual ability or surprise problems would try to confiscate the day. **Flip** thank you for helping with the most random computer problems (changing my password after the deadline always got me somehow) and **Sanne** thank you for the extremely efficient (and, again, short noticed) crash-course into movie editing. To **Martin**, thank you for introducing me to very interesting people at the ASHG. **Mark** and **Henk** thank you for solving three more computer problems everytime I come to you with one.

To everyone in the CMM, thank you for very interesting and diverse Friday meeting talks and discussions, as well as for the nice PhD events like master classes and PhD retreats.

To **Alex Babeanu**, thank you for letting me explain to you for 2 hours how genetics got PC analysis so very wrong; by the end of the phone call I had actually understood it and saw that it can work.

To my paranymphs, **Christina** and **Daniel**, thank you for your friendship and for your support in finishing off my thesis. Christina, thank you for a rigorous plan, that I did not follow...and Daniel, thank you for making plan B and C and D…until I decided I really want to stick to the initial one!

Finally, to my last authors, my parents **Gheorghe** and **Ioana Cretu Stancu**, thank you so dearly for the values and the education that you gave me, for always being inspiring and for supporting me endlessly and tirelessly.

Curriculum Vitae

Mircea Cretu Stancu was born on the 23rd of July in the city of Calarasi, Romania and underwent his pre-university education in Constanta, Romania. There, he participated in, and won NASA engineering competitions, in teams of students coordinated by Bararu Ion. He went on to do his Bachelor studies at Jacobs University in Bremen, Germany. He pursued a Bachelor of Science degree in Electrical Engineering and Computer Science. During his study in Bremen he worked as a research assistant for the 'Deutsches Forschungzentrum fur Kunstliche Intelligenz' where he implemented robot communication protocols for space exploration robots. He also held a position as teaching assistant for the programming course of his study, under the supervision of Dr. Stammer Johans. Mircea graduated with a Bachelor thesis in Machine Learning, using Recurrent Neural Networks to recognize handwritten text, invariant of an individual's style, under the supervision of Prof. Herbert Jaeger, in June 2011. The summer after his Bachelor studies he spent in Bucharest working as a research assistant within the Research and Development department of BitDefender, Softwin. There, he worked on a project aiming to authenticate handwritten signatures. He went on to start his Masters in Computing Science, at Utrecht Universiteit in the fall of 2011 with an interest in artificial intelligence and complexity theory. During his Masters course he did an internship for the IT Risk and Advisory, Financial Services Office of 'Ernst & Young LLP' and worked as a part-time programmer for 'Global Orange'. For his Master thesis he moved to the Genetics Department of UMC Utrecht, in Prof Paul de Bakker's lab, under the supervision of Dr. Laurent Franciolli. He graduated his masters in 2014 with a Bioinformatics thesis about analysis of whole genome sequencing data and went on to pursue a PhD degree, with Prof. Paul de Bakker as his promotor, in Human Genetics. Following Prof. Paul de Bakker's leave of the institution, he changed supervision to Prof. Edwin Cuppen as promotor, in the group of Dr. Wigard Kloosterman, who was his copromotor. He defends his PhD thesis in May 2018 and is currently living in Utrecht, the Netherlands.