Inherited and *de novo* variation in human genomes

Laurent C. Francioli

Cover art: Michael Eastman

Cover and interior design: Jake Muelle & Laurent Francioli

Printed by: Proefschriftmaken.nl || Uitgeverij BOXPress

Published by: Uitgeverij BOXPress, 's-Hertogenbosch

Inherited and *de novo* variation in human genomes

Erfelijke variatie en spontane mutaties in humane genomen

(met een samenvatting in het Nederlands)

Proefschrift

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
donderdag 9 april 2015 des middags te 4.15 uur

door

Laurent Carlo Francioli
geboren op 23 maart 1983 te Lausanne, Zwitserland

Promotor: Prof. dr. P.I.W. de Bakker

Contents

Chapter 1
Introduction 1

Chapter 2
Whole-genome sequence variation, population structure and demographic history of the Dutch population 19

Chapter 3
A framework for the detection of *de novo* mutations in trio sequencing data 51

Chapter 4
Genome-wide patterns and properties of *de novo* mutations in humans 61

Supplementary information 87

Chapter 5
Characteristics of *de novo* structural changes in the human genome 95

Supplementary information 125

Chapter 6
Summary and discussion 151

Summary 171
Samenvatting 173
List of publications 177
Acknowledgements 181
Curriculum vitae 185

Chapter 1

Introduction

Most human traits, ranging from physical appearance to behavior and disease susceptibility, are in part inherited through genetic material. Although the study of genetics started in the 19th century it took more than a century to be able to read the first entire human genome [1]. This milestone achievement provided the foundation for studying differences between human genomes and has led to the discovery of more than hundred million genetic variants to date [2–4] (Fig. 1).

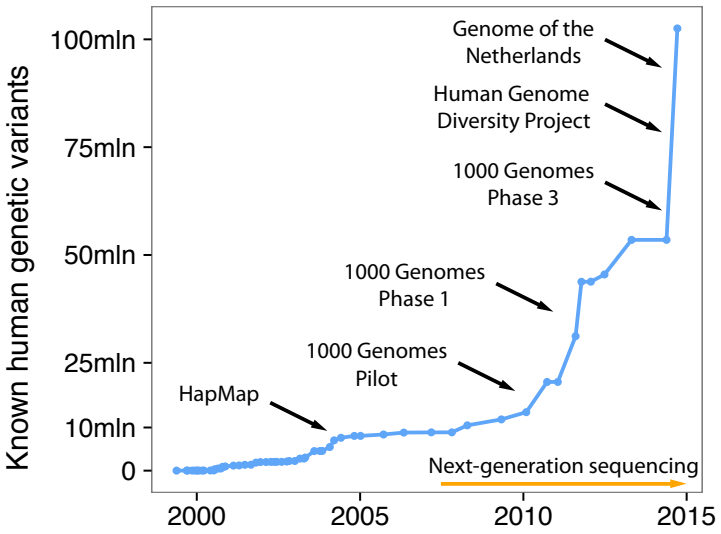


Figure 1 | Growth of catalogue of human genetic variants

This plot shows the chronological increase in the number of entries in dbSNP [4] , the largest database of human genetic variants. With the advent of next-generation sequencing (NGS) technologies, the number of known human genetic variants has grown from less than 10mln to more than 100mln today. Major projects contributing to dbSNP are highlighted using arrows, including the International HapMap Project [2,5,6] , the 1000 Genomes Project [3,7] , the Human Genome Diversity Project [8] and the Genome of the Netherlands Project [9].

Some of these variants are common in the population while others are rare or even private to a single individual. While most variants that exist in the population at large are rare, common variants represent the bulk of the variation within a single genome and account for most of the differences between individuals' genomes [3] . In addition, common variants have arisen many, many generations ago and tend to be shared across different populations while rare variants have emerged more recently and are therefore often specific to a single population [3] .

Although some of these genetic variants have been successfully associated with human traits, the effect of the vast majority of the variants catalogued is unknown. Likewise, the vast majority of the heritability of human traits is currently unexplained and the respective influence of common and rare variation on these traits is under debate [10] .

In this thesis, I will characterize the genetic variation in the genomes of 250 parent-offspring families representative of the Dutch population and show that this resource can help genetic association studies. I will then turn my attention to *de novo* mutations, that is genetic variants present in the offspring but absent from the parents. These mutations are of particular interest because they have not been subjected to natural selection and have been associated with both common and rare diseases [11] . Our current knowledge about *de novo* mutation in humans is mostly indirect [12,13] . This thesis will shed light on the properties of *de novo* mutations through their direct observation.

Heredity

The genetic information underlying heritable human traits is passed from parents to offspring through deoxyribonucleic acid (DNA). DNA represents a linear sequence of nucleotides (or bases), of which four kinds exist: adenine (A), cytosine (C), guanine (G) and thymine (T). The DNA molecule consists of two complementary strands coiled around each so that each base of a strand is paired up with its complement on the other (A with T and C with G). In humans, the genome comprises 3.2 billion bases and is packaged in 22 autosomes (numbered 1 though 22) and two sex chromosomes (X and Y). Humans are diploid, meaning that they carry two homologous copies of each of the autosomes and two sex chromosomes (either XX for females and XY for males).

Homologous chromosomes carry the same genetic information but can carry different versions of each locus called alleles. The genotype at a particular locus refers to the combination of two alleles present at this locus. For example, if a particular location in the genome is known to vary between a guanine (G) allele and a cytosine (C) allele, the possible genotypes would be GG, GC or CC. Following Mendel's laws of inheritance, children inherit one allele from their father and the other allele from their mother. This is achieved through two processes: first, meiosis produces haploid gametes (sperm in males, eggs in females) containing only one copy of each of the 22 autosomes and one of the two sex chromosomes. Second, fertilization during which an egg and a sperm fuse into a diploid zygote will then result in the entire set of 23 pairs of chromosomes.

Genetic variation

Genetic variation is introduced through *de novo* germline mutations, that is changes in the DNA sequence occurring during the formation of egg and sperm. Because these mutations are present in the gametes forming the offspring zygote, they will be present in all resulting cells that emerge during differentiation and development of the human body, possibly influencing phenotype and fitness.

Germline mutations naturally arise mainly from copying errors during DNA replication and from imperfect repair of double-stranded breaks (DSBs). DSBs are a naturally occurring and relatively frequent phenomenon [14] where both strands of the DNA molecule are cut. The repair of DSBs is challenging and can cause the formation of mutations at the repaired breakpoints [14–16] .

The formation of gametes is complex that starts during the embryonic development of the parents and ends at adult age. This process is different in males and females, leading to large differences in the total number of cell divisions required for a mature egg or sperm. In males, the production of sperm requires continuous mitotic cell divisions throughout reproductive life. In females, all eggs are already present at birth and only two meiotic cell divisions occur at adult age. Because the DNA needs to be replicated during each mitotic division, mutations stemming from the replication process show a paternal age trend, where more mutations are found in offspring of older than that of younger fathers. It is estimated that every additional year of paternal age leads to about 23 additional cell divisions [17] and one to two additional *de novo* mutations [18,19] .

It is essential for the organism to minimize the number of (potentially harmful) mutations and therefore repair mechanisms have evolved in order to maintain DNA integrity. The number of mutations passed to offspring thus represents a balance between the number of errors and the efficacy of the various types of repairs. Depending on their formation mechanism and their size, mutations are usually divided into three broad categories (Table 1): single nucleotide variants (SNVs), which replace a single base with another; short insertions and deletions (indels), which delete or insert a few bases (typically <50 bp); and structural variants (SVs), which are larger events affecting tens to thousands of bases.

Single nucleotide variants (SNVs) occur at a rate of about 45-100 per generation [18–20] . They mostly occur during DNA replication, through the incorporation of a base that does not match the corresponding base in the so-called template strand. In case the repair machinery fails to replace the wrongly incorporated base, the mutation will stay in the replicated DNA. Because of the different

Variant Type	Number per genome*	*De novo* rate**	Examples
Single Nucleotide Variants (SNVs)	3mln	61	**Single nucleotide variant** Reference: **CTGGATGAA** Variant: **CTGGGTGAA**
Short insertions and deletions (indels)	360k	3.7	**Deletion** Reference: **CTGGATGAA** Variant: **CTGG–TGAA**
			Insertion Reference: **CTGGA–TGAA** Variant: **CTGGAGTGAA**
Structural variants (SVs)	16k	0.03	**Deletion** Reference: Variant:
			Duplication Reference: Variant:
			Insertion Reference: Variant:

Table 1 | Classification of genetic variation

This table summarizes the classification of the most common types of genetic variation. In the structural variants examples, the black lines represent parts of chromosomes and the blue and red highlight changes in the chromosome containing the genetic variant.

* The numbers of variants are per-genome estimates from the 1000 Genomes Project[7] .

** The *de novo* mutation rates are estimates from Cambpell and Eichler[13] .

chemistry of the bases, not all mismatches are equally likely. Transitions, that is G-T and A-C base mispairing, are the most common and lead to A<->G and C <-> T mutations [12] . In addition, methylated cytosine bases followed by a guanine (called CpG sites) tend to undergo spontaneous deamination to thymine (C>T) [21] . For these reasons, the mutational spectrum of SNVs exhibits about twice as many transitions (A<->G, C<->T) than transversions (A<->C, A<->T, C<->G, G<->T).

The SNV mutation rate varies along the genome both at short and large scale. At short scale, the sequence context has the strongest effect, mainly due to CpG transitions, which are 10 to 18 fold more mutable than non-CpG dinucleotides [12,13,22] . At larger scales, SNV mutation rates correlate with transcription, replication timing, recombination rates, DNase I hypersensitivity sites and nucleosome occupancy [19,22] . The mechanisms behind these correlations are currently unknown and they do not explain entirely the observed variance in mutation rates along the genome [22] . In addition to these patterns, recent studies indicate that about 2% of mutations in an individual are clustered with respect to chromosomal location, suggesting that some of the mutations co-occur [19,23] ,

Short insertions and deletions (indels) are estimated to occur at a rate of 3 to 6 per generation [13] . This is very likely an underestimate of the true *de novo* indel rate, as indel discovery suffers reduced sensitivity in repetitive areas of the genome, where indels could be most abundant [24] . Two mechanisms are thought to explain most of the indel mutations [24] : polymerase slippage and imperfect repair of double-strand DNA breaks. Polymerase slippage occurs in regions of repeats of the same nucleotide (homopolymer) or of repeats of short sequences of a few nucleotides (short tandem repeats). During DNA replication in these regions, the polymerase can mispair the two DNA strands leading to the insertion or deletion of one or more copies of the repeat. The smaller the repeated unit and the longer the repetitive regions, the more it is prone to replication slippage and therefore indel formation. It is currently estimated that polymerase slippage accounts for 50 to 75% of indels and causes a roughly similar number of insertions and deletions [24] . Most of the remaining indels are likely caused by imperfect repair of double-strand DNA breaks (DSBs) where a few nucleotides are inserted or deleted when joining the DNA fragments [15] .

Structural variants (SVs) encompass all larger genetic variations, the most common and well studied being copy number variants (CNVs). CNVs are loosely defined as large insertions or deletions ranging from 1 kilobase to several megabases in size. Most CNVs result from the incorrect repair of the double-strand breaks (DSBs)

where previously distant chromosomal regions are ligated together [15,25] . Their mutation rate is estimated to be 0.03 CNVs per generation for those larger than 500 bp and 0.012 CNVs per generation for those larger than 100 kbp [13] . As these rates have been obtained using high-density microarrays, they likely represent an underestimation of the true mutation rate, especially for smaller CNVs (which are harder to detect).

Rise and fall of mutations

By definition, each genetic variant observed today has at some point emerged *de novo* in the genome of one of our ancestors. Similarly, with each new generation novel mutations will continue to appear. These mutations form the substrate for natural selection and genetic drift, the two processes that lead some variants to rise in frequency in the population and others to disappear.

Natural selection postulates that variants increasing fitness will (ultimately) rise in frequency in the population, as their carriers are more likely to survive and reproduce. For example, a mutation conferring a nutritional advantage by allowing adults to tolerate dietary lactose arose about 5,000 years ago in Europe and is now present in about 90% of modern Europeans [26] . Conversely, variants decreasing fitness will tend to disappear from the population. The strength of natural selection is proportional to the effect of the variant on fitness.

Genetic drift designates the random sampling of variants that are passed from one generation to the next. The magnitude of genetic drift is larger in small populations, where even deleterious variants can quickly reach high frequencies by chance. One such example can be found in the Amish population in Pennsylvania [27] , which is mostly descended from a small group of about 200 German settlers who migrated to the USA in 1744. Two of the founders shared a recessive allele for Ellis-van Creveld syndrome, a rare recessive monogenic disorder causing dwarfism and polydactyly present at about 0.01% in the general population. Due to the small population size and random sampling, the estimated frequency of this variant in the Amish population is currently 6.6%, despite its evidently deleterious effect.

Natural selection and genetic drift always act together and influence the evolutionary trajectories of alleles in populations. The influence of a variant on fitness depends greatly on its environment and therefore some variants only confer a selective advantage in certain situations.

Although new mutations continuously appear in the population, most are quickly lost by selection or drift. A simulation [28] showed that when introducing neutral mutations into a population of 2,000 individuals, 95% of these disappear within 31 generations and not one reaches a frequency of 1% within 2,000 generations.

Genetic linkage

Genetic linkage refers to the fact that nearby alleles are often transmitted together on a single haplotype. This is because the chance of a recombination event occurring in between nearby alleles is low. Conversely, two distant alleles are less likely to be transmitted together because recombination events will frequently occur somewhere in between. If two alleles are located very far from one another or even lie on different chromosomes, then they do not have any linkage and their co-transmission is random. Because the recombination rate is not uniform throughout the genome [29] , genetic linkage is not expressed in terms of bases but in centimorgans: a unit defined as 1% expected average recombination events per generation.

By definition, a novel mutation arises on a single haplotype spanning an entire chromosome. If this haplotype gets passed to future generations (and rises in frequency), recombination events can occur between the mutation and other surrounding alleles. As a result, such recombination events will give rise to new haplotypes (combinations of nearby alleles). Generally, blocks of nearby alleles will tend to be transmitted together more often than would be expected based on the frequencies of the individual alleles alone. That is to say, certain alleles are correlated to other nearby alleles, a phenomenon known as linkage disequilibrium (LD).

The consequence of LD is that the chromosomes for an individual can be thought of a sequence of haplotype "blocks" segregating in a population. The International HapMap Project has characterized the LD patterns for common single nucleotide polymorphisms (SNPs; ≥5% frequency) across the genome in 11 global populations [2] . Because of LD, it has become possible to statistically infer genotypes based on neighboring genotypes (a process called imputation) and to reconstruct chromosomal haplotypes based on genotype information (a process called phasing) [30] . In addition to recombination, the LD structure is dependent on the population history in terms of the ancestral and effective population size, bottlenecks, natural selection and genetic drift [31] . It is thus important to characterize the genetic structure of worldwide populations for disease studies.

Sequencing the human genome

The first DNA sequencing techniques were developed in the 1970s and were laborious and slow. By the 1990s, techniques had evolved to the point where sequencing the human genome became within reach and the Human Genome Project was started. It took three billion dollars and 13 years for the project to reach its goal and produce the first sequence of a human genome [1] . Meanwhile, Next Generation Sequencing (NGS) technologies were developed, allowing for the sequencing of larger number of samples.

The 1000 Genomes Project is the first project to leverage these technologies and aimed to sequence the genomes of more than 2,500 individuals from 26 global populations [3] . By comparing the sequencing data in each of these samples against the human reference sequence and against each other, over 80 million genetic variants have been catalogued and their frequencies estimated. This represents about 90% of all genetic variation present in at least 1% of the individuals in the 26 populations sequenced [3] .

As the cost continues to drop, NGS technologies are opening new doors for exploring common, rare and private genetic variation in humans and their role in health and disease. However, NGS poses many challenges, in part because of the large volumes of data generated, requiring high performance bioinformatics analysis tools [32] .

NGS technologies are based on shearing a DNA sample into small fragments (typically 30 to 500 bp) and sequencing millions of these fragments in parallel. Most sequencers available today support paired-end sequencing, which allows two short fragments physically on the same haplotype and separated by a fixed distance (typically ~500 bp) to be sequenced. The fragments (or fragment pairs) sequenced are random and it is therefore necessary to sequence substantially more DNA than the size of a single genome in order to obtain multiple sequence reads that span most of the genome and represent both haplotypes.

NGS technologies are prone to errors and typically 0.1% to 1% of the bases output by the sequencer are erroneous [33] . Because the DNA fragments are sequenced independently, it is necessary to produce multiple fragments covering each base of the genome in order to distinguish sequencing errors from true genetic variation. Sequencing depth, or coverage, refers to the number of fragments spanning each base of the genome. Importantly, there are systematic technical biases in regional coverage depth across the genome depending on the sequencing technology used and the sequence content of a region [34–36] . As a

result, not all regions of the genome are equally easy to sequence and analyze. In fact, it is estimated that about 15% of the human genome cannot be sequenced properly due to its complexity[7] and is thus currently "inaccessible".

Sequencers output the DNA sequence of each fragment separately and their original location in the genome is lost. The first step in reconstructing the genome sequenced is thus to place these fragments back at their originating location. The most common approach is to use a reference sequence and align each fragment onto that reference[37–39] . Because the genome sequenced differs slightly from the reference (not only due to errors but also true inter-individual genetic variation), alignment programs need to allow for mismatches and gaps in the reported alignments between the sequence reads and the reference sequence. Aligners often output a quantitative confidence score for each read they map reflecting the quality of the alignment[40] . Moreover, it is possible that reads can align to multiple places in the genome that have very similar sequences. In this case, some aligners report all possible alignments[39] , whereas others will randomly choose to align the read amongst the possible locations[37,38] .

Once all reads have been aligned against the human reference sequence, mismatches and inconsistencies should be evaluated as possible genetic variants. Because some of the bases in the sequences reflect sequencing errors or misaligned reads rather than true genetic variants, complex statistical models are necessary to robustly find genetic variation in NGS data[36,41,42] . Moreover, the wide variety of genetic variants, ranging from single nucleotide substitutions to large chromosomal rearrangements requires different algorithms using different information from the sequencing data (Table 2). Although many different models exist for detecting genetic variation, they usually assume that each read is an independent observation of the underlying DNA sequence and use the combined evidence from all reads spanning a certain position in the genome in order to distinguish genetic variants from errors. Coverage therefore plays a key role in the confidence with which variants can be detected.

After genetic variants have been identified, genotypes need to be assigned at each of the variant loci. The combined evidence across all reads spanning a position of region of the genome is evaluated against all possible genotypes at this position to infer the most likely genotype. Most algorithms will output a confidence score associated with the genotype assigned or genotype likelihoods for all possible genotypes. As for variant discovery, deeper coverage allows for more confident and accurate genotyping of the variants.

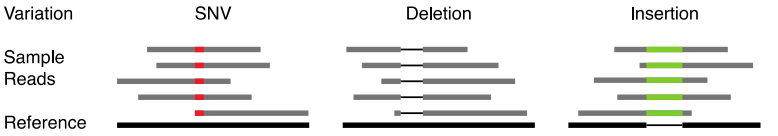
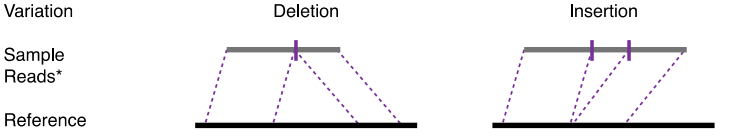
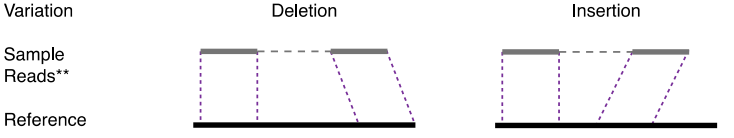
Mismatches in reads	Aligned reads are inspected for mismatches between the sequence of the aligned reads and the reference genome at the aligned position. Base mismatches indicate single nucleotide variants (SNVs), gaps in the alignment indicate deletions and gaps in the reference (to obtain the optimal alignment) indicate insertions. **Variant types**: SNV, insertion, deletion **Variant size**: small (typically smaller then half the read size)	
Split-reads	Reads that could not be aligned are split in smaller fragments. Each fragment is then mapped to the reference genome. When the smaller fragments can be mapped separately, insertions or deletions can be inferred by the distance separating the fragments. **Variant types**: Insertion, deletion **Variant size**: all	 * For clarity, only one read is pictured. The method is repeated for all reads and the joint evidence considered.
Discordant read-pairs	This approach can only be used on paired-end or mate-pair sequencing data, where reads are sequenced in pairs separated by a fixed distance (insert size). Insertions and deletions are then detected as significant difference between the expected and observed insert size. Note that using by the reads direction more complex structural variants can also be detected. **Variant types**: Insertion, duplication, deletion, inversion, evertion, chromosomal rearangment **Variant size**: large (typically greater than 100bp)	 ** For clarity, only one read-pair is pictured. The method is repeated for all pairs and the joint evidence considered.
Read depth	Coverage of reads mapped on the reference genome is evaluated throughout the genome. Deleted regions will display half (if heterozygous) the coveage of the flanking region or no coverage (if homozygous). Duplicated regions will be covered 1.5 times (if heterozygous) or twice (if homozygous) as much than the flanking region. **Variant types**: Deletion, duplication **Variant size**: very large (typically greater than 1,000 bp)	

Table 2 | Variant detection methods in next-generation sequencing data

In addition to the challenges of reconstructing a genome from short sequencing reads, the amount of data generated and analyzed is very large (for example, ~200Gb of raw data per sample for 12x coverage) and thus requires large compute clusters and complex data management [9,43] .

The Genome of the Netherlands Project

The Genome of the Netherlands (GoNL) Project [9] is a large sequencing initiative from the Biobanking and Biomolecular Research Infrastructure - Netherlands (BBMRI-NL) [44] , which aims to characterize the genetic variation in the Dutch population. Five Dutch biobanks (LifeLines Cohort Study, Leiden Longevity Study, Netherlands Twin Registry, Rotterdam Study, Rucphen Study) contributed DNA from 250 parent-offspring families with Dutch ancestry. These families were sampled without phenotypic ascertainment from all provinces of the Netherlands except for the province of Flevoland, which was reclaimed from water during the 20th century.

In **Chapter 2**, I will show that the GoNL project is well powered for common and rare variant detection, including variants that are technically more challenging to find such as indels and structural variants (SVs). The resulting catalogue of genetic variation is then used to investigate the genetic architecture of the Dutch population and the per-individual burden of novel mutations as well as loss-of-function alleles.

One of the aims of the parent-offspring design of the GoNL project is to create an accurate haplotype panel for imputation in existing genotype data sets such as those for genome-wide association studies (GWAS). **Chapter 2** investigates the gain in phasing accuracy of the haplotype panel due to the familial relationships in the GoNL families. The downstream imputation accuracy of the resulting haplotype panel is then compared to that of 1000 Genomes [3] haplotypes for the imputation of Dutch and other European ancestry samples.

In addition to providing a catalogue of the genetic variation in the Netherlands, the GoNL project familial design provides a unique opportunity to observe *de novo* mutations across the genome. Although the GoNL project is the largest set of whole-genome pedigrees to date, finding *de novo* mutations with a sequencing coverage of only 13x is challenging. In **Chapter 3**, I describe a method for robust *de novo* SNV and indel detection in sequencing data. The application of this method on the GoNL data is described in **Chapter 2** and **Chapter 4** for SNV detection and in **Chapter 5** for indel detection.

Because of their rarity, *de novo* mutations are challenging to observe in humans and thus the properties of human mutations have mainly been studied in model organisms, Mendelian diseases or inferred using comparative genomics or population genetics approaches [12,13] . *De novo* copy number variants (CNVs) [45–47] have been observed in pedigrees using microarray technologies and a few recent studies have started to use whole-genome sequencing of pedigrees to observe *de novo* single nucleotide variants (SNVs) [18,19] but were limited in scale and mostly studied disease phenotypes. The GoNL project described in this thesis represents the largest resource to date to study *de novo* mutations across the genome.

Chapter 2 and 4 describe the properties of *de novo* SNVs. **Chapter 2** focuses on the effect of paternal age on the number of mutations. **Chapter 4** presents an in-depth analysis of *de novo* SNV properties, including their mutational spectrum and factors influencing their rates across the genome. The results are compared to those from previous studies using comparative genomics approaches and from somatic cancer mutation studies.

In **Chapter 5,** the properties of *de novo* indels and SVs are investigated, focusing on rates and mechanism of formation. These data are then combined with the *de novo* SNVs to study the relative impact of the different types of mutations on the genome. By comparing inherited and *de novo* mutations by type, I investigate the strength of the selective pressures on the different types of mutations.

Finally, in **Chapter 6** I will describe several lessons learned from the analysis of the GoNL data and discuss their implications for studies using NGS data.

References

1. Lander, ES, Linton, LM, Birren, B, Nusbaum, C & Zody, MC. Initial sequencing and analysis of the human genome. *Nature* (2001).
2. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–8 (2010).
3. McVean, G. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
4. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–11 (2001).
5. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
6. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
7. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–73 (2010).
8. Cann, H. M. *et al.* A human genome diversity cell line panel. *Science* **296,** 261–2 (2002).
9. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* **22,** 221–7 (2014).
10. Gibson, G. Rare and common variants: twenty arguments. *Nature reviews. Genetics* **13,** 135–45 (2011).
11. Veltman, J. & Brunner, H. De novo mutations in human genetic disease. *Nature reviews. Genetics* **13,** 565–75 (2012).
12. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **15,** 47–70 (2014).
13. Campbell, C. & Eichler, E. Properties and rates of germline mutations in humans. *Trends in genetics : TIG* **29,** 575–84 (2013).
14. 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).
15. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10,** 551–64 (2009).
16. Lieber, M. R. The mechanism of human nonhomologous DNA end joining. *J. Biol. Chem.* **283,** 1–5 (2008).
17. Crow, J. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1,** 40–47 (2000).
18. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–5 (2012).

19. Michaelson, J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151,** 1431–42 (2012).

20. Neale, B. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485,** 242–5 (2012).

21. Jabbari, K. & Bernardi, G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* **333,** 143–9 (2004).

22. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12,** 756–66 (2011).

23. Campbell, C. D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44,** 1277–81 (2012).

24. Montgomery, S. B. *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* **23,** 749–61 (2013).

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

26. Swallow, DM. Genetics of lactase persistence and lactose intolerance. *Annual review of genetics* (2003). doi:10.1146/annurev.genet.37.110801.143820

27. McKusick, V. A. Ellis-van Creveld syndrome and the Amish. *Nat. Genet.* **24,** 203–4 (2000).

28. Raychaudhuri, S. Mapping rare and common causal alleles for complex human diseases. *Cell* **147,** 57–69 (2011).

29. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467,** 1099–103 (2010).

30. Browning, S. & Browning, B. Haplotype phasing: existing methods and new developments. *Nature reviews. Genetics* **12,** 703–14 (2011).

31. Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9,** 477–85 (2008).

32. Koboldt, D. C., Ding, L., Mardis, E. R. & Wilson, R. K. Challenges of sequencing human genomes. *Brief. Bioinformatics* **11,** 484–98 (2010).

33. Glenn, T. C. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11,** 759–69 (2011).

34. Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* **5,** 1005–10 (2008).

35. Smith, D. R. *et al.* Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.* **18,** 1638–42 (2008).

36. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

37. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–95 (2010).

38. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

39. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9,** 1185–8 (2012).

40. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18,** 1851–8 (2008).

41. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–93 (2011).

42. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46,** 912–8 (2014).

43. Clarke, L. *et al.* The 1000 Genomes Project: data management and community access. *Nature methods* **9,** 459–62 (2012).

44. Brandsma, M, Baas, F, de Bakker, P. & Beem, EP. How to kickstart a national biobanking infrastructure–experiences and prospects of BBMRI-NL. *Norsk …* (2012).

45. Itsara, A. *et al.* De novo rates and selection of large copy number variation. *Genome Res.* **20,** 1469–81 (2010).

46. Hehir-Kwa, J. *et al.* De novo copy number variants associated with intellectual disability have a paternal origin and age bias. *Journal of medical genetics* **48,** 776–8 (2011).

47. Conrad, D. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–12 (2010).

Chapter 2

Whole-genome sequence variation, population structure and demographic history of the Dutch population

Laurent C. Francioli*, Androniki Menelaou*, Sara L. Pulit*, Freerk van Dijk*, Pier Francesco Palamara, Clara C. Elbers, Pieter B.T. Neerincx, Kai Ye, Victor Guryev, Wigard P. Kloosterman, Patrick Deelen, Abdel Abdellaoui, Elisabeth M. van Leeuwen, Mannis van Oven, Martijn Vermaat, Mingkun Li, Jeroen F.J. Laros, Lennart C. Karssen, Alexandros Kanterakis, Najaf Amin, Jouke Jan Hottenga, Eric-Wubbo Lameijer, Mathijs Kattenberg, Martijn Dijkstra, Heorhiy Byelas, Jessica van Setten, Barbera D.C. van Schaik, Jan Bot, Isaäc J. Nijman, Ivo Renkens, Tobias Marschall, Alexander Schönhuth, Jayne Y. Hehir-Kwa, Robert E. Handsaker, Paz Polak, Mashaal Sohail, Dana Vuzman, Fereydoun Hormozdiari, David van Enckevort, Hailiang Mei, Vyacheslav Koval, Matthijs H. Moed, K. Joeri van der Velde, Fernando Rivadeneira, Karol Estrada, Carolina Medina-Gomez, Aaron Isaacs, Steven A. McCarroll, Marian Beekman, Anton J.M. de Craen, H. Eka D. Suchiman, Albert Hofman, Ben Oostra, André G. Uitterlinden, Gonneke Willemsen, LifeLines Cohort Study, Mathieu Platteel, Jan H. Veldink, Leonard H. van den Berg, Steven J. Pitts, Shobha Potluri, Purnima Sundar, David R. Cox, Shamil R. Sunyaev, Johan T. den Dunnen, Mark Stoneking, Peter de Knijff, Manfred Kayser, Qibin Li, Yingrui Li, Yuanping Du, Ruoyan Chen, Hongzhi Cao, Ning Li, Sujie Cao, Jun Wang, Jasper A. Bovenberg, Itsik Pe'er, P. Eline Slagboom, Cornelia M. van Duijn, Dorret I. Boomsma, Gertjan B. van Ommen, Paul I.W. de Bakker*, Morris A. Swertz*, Cisca Wijmenga*, *Nature Genetics* 46, 818–25 (2014).

Supplementary information: http://www.nature.com/ng/journal/v46/n8/full/ng.3021.html

Abstract

Whole-genome sequencing enables complete characterization of genetic variation, but geographic clustering of rare alleles demands many diverse populations be studied. Here, we describe the Genome of the Netherlands ("GoNL") Project, in which we sequenced whole genomes of 250 Dutch parent-offspring families and constructed a haplotype map of 20.4 million single nucleotide variants and 1.2 million insertions and deletions. The intermediate coverage (~13x) and the trio design enabled extensive characterization of structural variation, including mid-size events (30-500 bp) previously poorly catalogued, and *de novo* mutations. We demonstrate that the quality of the haplotypes significantly boosts imputation accuracy in independent samples, especially for lower frequency alleles. Population genetic analyses demonstrate fine-scale structure across the country and support multiple ancient migrations, consistent with historical sea level changes and flooding. The GoNL Project illustrates how single-population whole-genome sequencing can provide detailed characterization of genetic variation, and may guide the design of future population studies.

Introduction

Although the human genome reference sequence provides a common scaffold for the annotation of genes, regulatory elements and other functional units, it does not contain information about how individuals differ in their DNA sequences.[1] Initial efforts to map such variation across the human genome have successfully catalogued millions of common single-nucleotide polymorphisms (SNPs) in various populations.[2–5] Fueled by the commercial development of microarrays for efficient SNP genotyping, genome-wide association studies (GWAS) have provided a systematic approach to test genetic variants for a role in disease. To date, GWAS have reproducibly identified thousands of loci, providing insight into underlying pathways of disease, in some cases with translational and clinical impact.[6,7] The importance of these discoveries notwithstanding, many questions remain about the allelic architecture of complex traits, especially with regard to the contributions of common versus rare variation.[7–9]

To elucidate the genetic basis of disease, comprehensive sequencing-based approaches are required to interrogate all types of genetic variation, including single nucleotide variants (SNVs), structural variations and *de novo* events.[10–12] The characterization of rare variation poses a major challenge. Since rare alleles have emerged on average relatively recently,[13] they show greater geographic clustering[14] than common variants.[15] It is therefore imperative to study large

samples across multiple populations, even within continental groups, to build a relatively complete catalog of rare variation in the human genome.

We initiated the Genome of the Netherlands ("GoNL") Project to characterize DNA sequence variation for SNVs, short insertions and deletions (indels), and larger deletions in 769 individuals of Dutch ancestry selected from five biobanks under the auspices of the Dutch hub of the Biobanking and Biomolecular Research Infrastructure (BBMRI-NL).[16,17] Specifically, we sampled 231 trios, 11 quartets with monozygotic twins, and 8 quartets with dizygotic twins, from 11 of the 12 Dutch provinces without ascertaining on phenotype or disease. By whole-genome sequencing these 250 families at ~13x coverage, our aim was to build a resource of 1,000 haploid genomes as representative of a small (41,543 km^2) densely populated country (> 17 million inhabitants) in northwestern Europe (**Supplementary Note**).

Here, we provide the first detailed analysis of the GoNL data after processing and quality control (**Supplementary Fig. 1** and **Supplementary Note)**. To maximize sensitivity, we analyzed all samples jointly[18] and discovered 20.4 million biallelic SNVs, 1.2 million biallelic indels (< 20 bp) and 27,500 larger deletions (> 20 bp). Of the SNVs, 6.2 million are common (MAF > 5%), 4.0 million are low-frequency (MAF 0.5–5%), and 10.2 million are rare (MAF < 0.5%). Based on coverage and mapping metrics, we estimate that 94.1% of the genome could be called reliably (the "accessible" genome) within which 99.2% of SNVs of 1% frequency could be detected (**Supplementary Note**, **Supplementary Table 1**). Indels and large deletions were based on conservative consensus calls from several complementary methods (**Supplementary Note**). We used MVNcall for trio-aware phasing and linkage disequilibrium-based imputation,[19] starting from the genotype likelihoods of SNVs and indels, yielding a phased panel of 998 unique haplotypes. The non-reference genotype concordance for SNVs was 99.4% (compared to genotypes from Complete Genomics sequencing data in 20 overlapping samples) and 99.5% (compared to Illumina Immunochip genotypes collected in all samples). The average coverage of 13.3x coupled with the family-based design allowed us to construct a high-quality whole-genome data set for further analysis, including characterization of structural variation, detection of *de novo* events, imputation, and demographic inference.

Novel variation in GoNL

To determine the number of novel variants, we investigated the overlap between GoNL and existing databases. We detected the majority of sites (98.2%) present in the European sample (CEU) of HapMap Phase 2,[4] and 71.1% of sites in the European subset of the 1000 Genomes Project Phase 1 (1KG-EUR),[20] consistent with the expectation that commonly segregating alleles across European populations should also be detected in GoNL (**Fig. 1a**). Conversely, only 39.0% of rare SNVs observed in GoNL (excluding singletons) were observed in 1KG-EUR, highlighting the value of studying individual populations in greater depth. The contribution of 7.6 million novel SNVs in GoNL represents a 14.6% increase of dbSNP (build 137), although the majority (75.6%) are singletons. Considering that 16.5% of 2.0 million singletons in 1KG-EUR were also observed in GoNL, we expect that a substantial number of the novel GoNL singletons will be encountered again as we continue to sequence larger samples across Europe.

Structural variation could be called confidently across a broad size range, from large deletions to short insertions (**Fig. 1b**). The overall shape of the size frequency distribution shows that larger structural events are less frequent than smaller indels, presumably reflecting their relative deleterious nature. We observed specific peaks in the size frequency spectrum that correspond to microsatellite instability (MSI) around 4 bp, short interspersed elements (SINEs) at 300 bp, and long interspersed elements (LINEs) at 6 kb. In comparison to 1KG, 54.4% of the indels (≤ 20 bp) and 93.3% of the larger deletions (> 20 bp) are novel (**Supplementary Note**). Our analysis thus fills an important gap in the discovery of mid-size deletions (30–500 bp), where 98.4% of the observed variants are novel. The novelty rate for larger deletions (> 500 bp) is still substantial (66.3%). We note that most of the deletions reported here are biased to be common because of stringent filtering (**Supplementary Note**), which allowed us to generate a call set with an overall validation rate of 96.5% (**Supplementary Table 2**). A more complete data set including duplications, inversions, mobile element insertions, and translocations is currently being assembled.

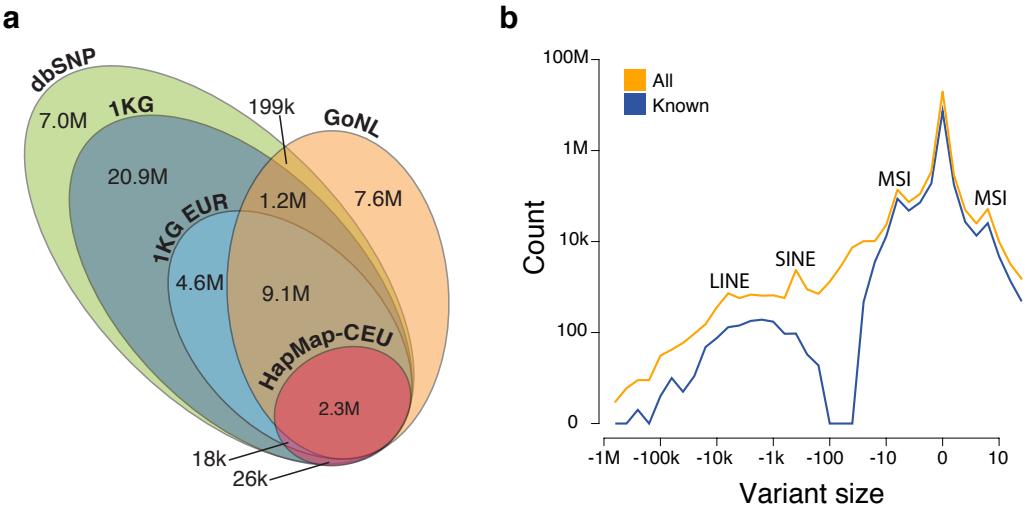


Figure 1 | Discovery of SNVs and structural variation.

a. Venn diagram of all SNVs discovered in GoNL relative to dbSNP (build 137), 1KG Phase 1 and HapMap-CEU. The majority of the 7.6 million novel sites are rare (MAF < 0.5%), including 5.8 million singletons.

b. Size spectrum of structural variation discovered in GoNL. Our detection strategy employed multiple approaches and provided a significant boost in novel SVs in the midsize range (30–500 bp). Peaks corresponding to long interspersed elements (LINEs), short interspersed elements (SINEs) and microsatellite instability (MSIs) are highlighted. The total number of variants called in GoNL are shown in orange, whereas SNVs found in dbSNP (build 137) and short indels and large deletions found in 1KG Phase 1 are shown in blue. For large deletions (> 20 bp), we required at least 80% reciprocal overlap between variants to be considered as similar.

Functional variation

Predicting the biological consequences of variants within a single genome is an ongoing challenge with important implications for using sequencing in a clinical setting. To characterize the burden of loss-of-function (LoF) variants in detail, we classified all such variants in GoNL[21] (**Supplementary Note**). Amongst rare variants, we observed an excess of nonsense variants and frameshift indels, consistent with a model in which such functional variants are subject to purifying selection (**Supplementary Fig. 2**).[22,23] We counted 66 larger LoF deletions (removing the first exon of a gene or more than half of its coding sequence)[21], which showed a relative depletion when compared to all deletions ($p = 0.005$ for 20–100 bp; $p = 2.6 \times 10^{-9}$ for > 100 bp). This effect was amplified when considering only genes listed in the OMIM compendium ($p = 2.4 \times 10^{-27}$), illustrating strong selection against large structural changes in key genes.

The overall patterns and per-individual distributions of LoF SNVs (premature stops or variants interrupting splice sites) and missense variants are consistent with those found in 1KG (**Table 1, Supplementary Fig. 3**). On average, an individual carries 60 LoF SNVs, 69 LoF indels, and 15 LoF large deletions. The bulk of these LoF mutations per individual are common, suggesting that these variants are not subject to strong selective pressure and, though they are protein-truncating mutations, are likely phenotypically benign. This observation emphasizes the need for caution in assigning pathogenicity to variants purely based on their predicted impact on protein structure.

In contrast, when considering rare LoF variants, those more likely to be pathogenic, we found that the average individual in GoNL carries 4 nonsense variants, 2 variants interrupting a splice site, and 2 frameshift indels. Comparing these numbers to synonymous variants (providing a baseline expectation under neutrality), we estimate that each individual carries an excess of 4–5 rare LoF SNVs sufficiently deleterious that they will never reach high frequency in the population (**Supplementary Note**).

We also investigated the number of rare LoF compound heterozygotes events for SNVs, short indels and large deletions. Across all samples, we observed three such events mapping to three genes in three individuals (average 0.01 events per individual). Given their rarity, the phenotypic impact of compound heterozygotes of rare LoF mutations should be considered explicitly in disease studies.

	Non-reference allele frequency		
	Rare (< 0.5%)	**Low-frequency (0.5–5%)**	**Common (> 5%)**
Variant type	Mean [SD]	Mean [SD]	Mean [SD]
All SNVs[1]	28,142 [3009.2]	130,190 [2448.1]	2.90M [10,080.9]
Novel[1,2]	17,751 [1,176.3]	4,354 [346.8]	620 [31.7]
Total conserved	1,892 [187.7]	7,593 [154.5]	106,824 [443.9]
Functional variation			
Synonymous	18 [4.9]	73 [8.9]	990 [19.0]
Nonsynonymous	101 [11.9]	238 [15.6]	2089 [31.8]
Probably damaging	32 [5.8]	58 [7.9]	394 [12.2]
Stop gain[1]	4 [1.9]	5 [2.2]	38 [4.3]
Splice site donor[1]	1 [0.9]	1 [0.9]	4 [1.5]
Splice site acceptor[1]	1 [0.7]	0.5 [0.6]	7 [1.4]
Total LoF[1]	5 [2.2]	6 [2.4]	49 [4.7]
Disease-associated variation			
OMIM	0 [0.6]	2 [1.6]	57 [4.9]
HGMD[3]	2 [1.2]	8 [2.7]	11 [2.3]
Indels (< 20 bp)			
Indel frameshift[1]	2 [1.4]	6 [2.6]	61 [4.8]
Indel non-frameshift[1]	1 [1.1]	6 [2.6]	99 [5.9]
Deletions (> 20 bp)			
Loss of function	0 [0.2]	1 [1.0]	14 [3.3]
Total bases deleted	6.7M bases		

Table 1 | Individual variant load of coding mutations

Only SNV sites at which ancestral state can be assigned with high confidence and that are highly conserved (GERP > 2.0) are reported. Frequency stratifications based on the unrelated samples only. OMIM, Online Mendelian Inheritance in Man.

[1]No conservation filter applied

[2]Not observed in dbSNP build 137 (which includes all SNVs reported in the 1000 Genomes Project Phase I data release)

[3]Frequency stratification and variant counts based on the reported mutation allele

Whereas compound heterozygotes of rare LoF variants are sparse, we expected compound heterozygotes of common LoF variants to be more prevalent, as these variants are less likely to be deleterious. Indeed, we found that the average number of common-LoF compound heterozygotes per individual was 2.89 (range: 0–7). Interestingly, while there are 1,917 common-LoF compound heterozygous events across all samples, they are confined to only 11 genes (**Supplementary Table 3**). All but one of these genes have extreme Residual Variation Intolerance Scores[24] (all >84th percentile across 16,956 genes), which is unlikely to occur by chance ($p = 1.41 \times 10^{-5}$, **Supplementary Fig. 4**) and suggests these genes are more tolerant of disruptive mutations.

Because disease mutation databases are often employed to identify potential variants of interest, we annotated variants in GoNL listed as disease-causing ("DM") in the Human Gene Mutation Database (HGMD).[25] We observed that a sample carries, on average, 20 "DM" variants (range: 9–33) (**Table 1**). Since all samples were derived from population-based cohorts, the impact of these alleles is unclear. One possibility is that the presence of modifier alleles induces incomplete penetrance or variable expressivity of "DM" variants depending on the carrier's genetic background.[26] An alternative explanation is that HGMD contains a considerable number of false-positive disease-causing mutations.[27] Of the 1,093 "DM" mutations occurring in GoNL, 32% have a frequency >1%, higher than the prevalence of many of the diseases described in HGMD. Given the inheritance patterns of the diseases conferred by these variants, many individuals in GoNL should have been affected by diseases with profound physical or even lethal manifestations (**Table 2, Supplementary Table 4**). In fact, one of these variants (chr14:94847262, an alpha 1 antitrypsin deficiency variant) was recently implicated as a pathogenic incidental finding in a set of 1,000 exomes.[28] The prevalence of alpha 1 antitrypsin deficiency (OMIM: 613490), an autosomal recessive disease, is estimated to be 0.02-0.06%, yet two unrelated GoNL individuals are homozygous carriers of this variant (prevalence = 0.4%, ~10x higher than the disease prevalence). Further, the typical age of onset of alpha 1 antitrypsin deficiency is 20–50 years old, whereas the two homozygous carriers in GoNL were ages 60 and 63 at ascertainment. These results highlight the potential pitfalls of employing such databases in disease studies and the challenge of interpreting personal genomes.

Chr	Pos	Gene	Mut-ation allele	Refer-ence allele	Disease in HGMD	Disease prevalence	Inheritance pattern	Affected individuals[3]	Phenotypic manifestations[1]	OMIM ID(s)	Mutation allele frequency GoNL[4]	Mutation allele frequency 1KG – CEU
4	6302519	WFS1	A	G	Wolfram syndrome	0.0002%[1]	AR	257	Hyperglycemia, vision and hearing loss	604928, 222300	0.728	0.759
13	52515354	ATP7B	G	A	Wilson disease	0.003%[1]	AR	167	Liver disease, neuropsychiatric problems	277900	0.574	0.582
16	3304463	MEFV	T	C	Familial Mediterranean fever	0.10% in Mediterranean populations; rarer elsewhere[1]	AR	36	Recurrent fevers, inflammation of the abdomen, chest, joints	249100, 134610	0.277	0.224
11	6415463	SMPD1	A	G	Niemann-Pick disease	0.0004%[1]	AR	37	Nervous system deterioration, failure to thrive, fatal in infancy or early childhood (type A)	257200, 607616, 257220, 607625,	0.230	0.230
20	61463522	COL9A3	A	C	Pseudo-achondroplasia	0.003%[1]	AD	177	Short stature, joint pain	177170	0.197	0.200
10	13340236	PHYH	A	G	Refsum disease	Unknown, current estimate 0.0001%[1]	AR	18	Anosmia, progressive blindness, deafness, hand/feet bone abnormalities, arrhythmia	266500	0.188	0.153
15	52643564	MYO5A	A	G	Griscelli syndrome	<0.0001%[2]	AR	10	Albinism (all types), intellectual disability (type 1), recurrent infection (type 2)	214450, 607624, 609227	0.159	0.141
19	36339247	NPHS1	T	C	Congenital nephrotic syndrome (Finnish type)	0.01% in Finland; rarer elsewhere[2]	AR	2	Proteinuria, rapid progression to renal failure	256300	0.082	0.082 (0.110)[5]
14	94847262	SERPINA1	A	T	Alpha 1 antitrypsin deficiency	0.02-0.06%[1]	AR	2	Lung disease, liver disease	613490	0.039	0.053

Table 2 | HGMD disease-causing mutations in the GoNL samples

Acronyms are: HGMD (Human Gene Mutation Database); AR (autosomal recessive); AD (autosomal dominant); OMIM (Online Mendelian Inheritance in Man). [1]National Institutes of Health, Genetics Home Reference – USA. [2]National Institute of Health and Medical Research – France. [3]Unrelated individuals in GoNL carrying two copies of the mutation allele (for autosomal recessive diseases) or at least one copy of the mutation allele (for autosomal dominant diseases). [4]Calculated from unrelated individuals. [5]Frequency in 1KG Phase I samples from Finland

De novo mutations

A distinct advantage of the family-based study design was the ability to call *de novo* events in genomic regions with sufficient coverage in a trio. To this end, we developed the PhaseByTransmission (PBT) module in the Genome Analysis Toolkit (GATK).[29] From an initial 4.5 million Mendelian violations in the original calls made in the 258 independent offspring, we prioritized 29,162 autosomal *de novo* mutation (DNM) candidates at non-polymorphic sites with PBT (**Supplementary Note**). Given that the average number of *de novo* mutations per offspring is still higher than expected (~63.2 mutations per offspring[30]), we evaluated to what extent sequencing features could help increase the DNM prediction accuracy and reduce false positives. We validated 592 candidate sites as true DNMs (**Supplementary Note**) and classified another 1,674 candidates as false positives (on the basis of validation experiments and Complete Genomics genotype data). We trained a random forest classifier on various features using 70% of the validation results (**Supplementary Note**), and obtained a model with an estimated classification accuracy of 92.2% using the remaining 30% of the data (**Fig. 2a**). This illustrates that the joint assessment of raw trio data and sequencing context can greatly boost prediction accuracy. We applied the classifier to our initial candidates and determined 11,020 high-confidence DNMs (18–74 DNMs per offspring) for downstream analyses. Due to regional coverage fluctuations, we expect a substantial fraction of genuine DNMs to be missed. We also note that early embryonic somatic mutations would be indistinguishable from germline mutations.

We observed a significant positive correlation ($r^2 = 0.47$, $p < 2.2 \times 10^{-16}$) between father's age at conception and number of DNMs in the offspring (**Fig. 2b**), providing a third, independent estimate based on a larger sample size.[30,31] Accounting for a Poisson-distributed background mutation rate and conditioning on coverage, we estimate that each additional year of father's age causes a 2.5% increase in the mean number of DNMs in the offspring. While parents' ages are highly correlated ($r^2 = 0.66$), comparing models based on father's and mother's age at conception suggests that the age-related increase in DNMs is a predominantly paternal effect (**Supplementary Note**). Interpolating from the paternal model, we expect on average 75.4% of the DNMs in the GoNL offspring to originate from the father (assuming a linear increase in DNMs from puberty). Using read-pair information, we were able to assign parental origin to 2,613 DNMs, and found that indeed 76.0% were paternal. When considering only mutations for which parental origin could be determined, the correlation with father's age remained significant ($r^2 = 0.11$, $p = 2.0 \times 10^{-6}$, **Supplementary Fig. 5**) but did not for mother's age ($p = 0.94$), highlighting the relative impact of paternal and maternal mutations.

Within a single family, we attempted to discover *de novo* indels and large deletions. Using strict filtering criteria for Mendelian violations followed by PCR-based Sanger sequencing (**Supplementary Note**), we confirmed 6 intergenic *de novo* indels (1–2 bp) and a large 113 kb *de novo* deletion located in an intron of the *SUMF1* gene (which seems unlikely to have a significant impact on gene function). These results illustrate that our predictions of indels and structural changes are a valuable source for both commonly segregating alleles and *de novo* events. Further work is needed to assess the frequency of such *de novo* events in the general population.

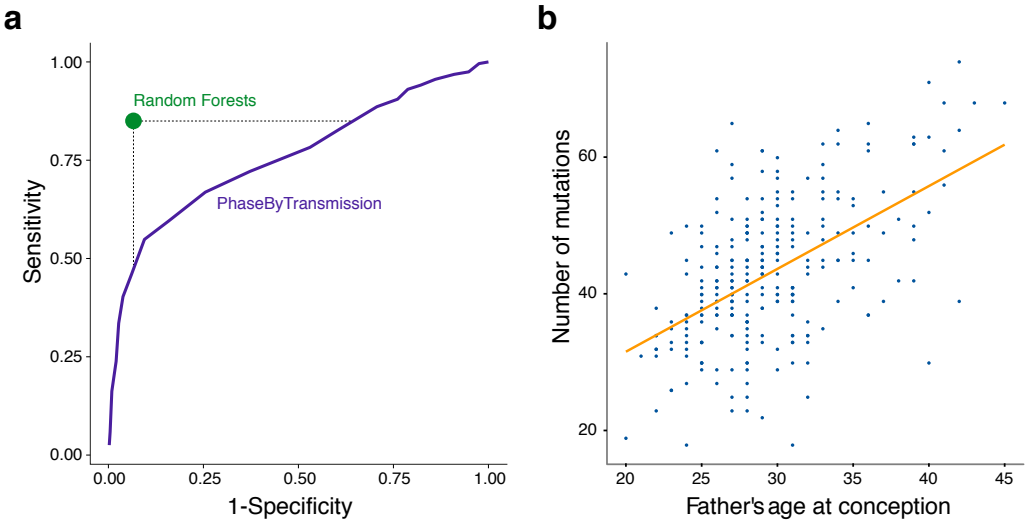


Figure 2 | *De novo* mutation detection.

a. Receiver-operating-characteristics (ROC) curve to predict *de novo* mutations using PhaseByTransmission only (purple line, 2,199 sites) or using PhaseByTransmission followed by Random Forests classification trained on 70% of the validation data (green line, evaluation subset only, 657 sites). The highlighted circle is the cutoff we used for our analyses with an estimated 84.5% sensitivity and 94.6% specificity.

b. The number of *de novo* mutations in each of the 258 independent offspring is plotted (in blue) as a function of paternal age at conception. Linear regression of mutational load on paternal age is significant (Pearson's correlation = 0.47, $p < 2.2 \times 10^{-16}$), with the least-squares fit plotted in orange.

Imputation

One of the goals of the GoNL Project was to provide a community resource for downstream imputation into GWAS samples. To evaluate the performance of the GoNL panel we used independent Complete Genomics sequence data collected in 81 individuals of Dutch ancestry (the "NTL" data set). In these NTL samples, we masked all genotypes at SNVs not present on the Illumina Human-1M array, imputed these masked SNVs from the 1M-genotyped SNVs, and then compared the imputed and known genotypes (**Supplementary Note**). The aggregate mean r^2 was 0.99 for common SNVs, 0.86 for low-frequency SNVs and 0.63 for rare SNVs, indicating a good overall imputation quality (**Fig. 3**). We repeated this evaluation based on the SNV content of other microarrays and obtained similar imputation performance for common SNVs but observed significant differences for lower frequency alleles (**Supplementary Fig. 6**). To directly measure the impact of trio-based phasing, we constructed a panel based on the unrelated parents alone, and re-evaluated the imputation quality in the NTL samples. The imputation accuracy dropped to a mean r^2 of 0.47 for rare variants (0.85 and 0.98 for low-frequency and common SNVs, respectively), indicating that the trio-based phasing contributed significantly to the imputation quality of rare variants.

In comparison to 1KG as a reference panel, we observed better imputation accuracy with the GoNL panel for SNVs up to 10% frequency despite the larger sample size of 1KG (**Fig. 3**). To investigate the basis for the improved imputation accuracy with the GoNL panel, we constructed three reference panels based on 1KG-CEU (Northern Europeans from Utah), 1KG-TSI (Tuscans from Italy), and GoNL, all with 85 individuals. With each of these reference panels, we imputed in independent CEU, TSI and NTL samples with Complete Genomics data (**Supplementary Note**). Of the three panels, GoNL gave the highest imputation accuracy (especially for rare variants) not only for the NTL samples but also for the CEU samples, indicating that the improved performance of the GoNL panel was not simply due to shared ancestry of the GoNL and NTL samples (**Supplementary Fig. 7**). Differentiation between northern and southern European populations may explain why the 1KG-CEU and GoNL panels gave roughly equivalent performance for TSI (but certainly worse than 1KG-TSI). Overall, these results suggest that the GoNL trios enabled accurate reconstruction of long-range haplotypes with marked improvement in imputation of rare alleles.

To assess the potential value of larger reference panels, we combined the 1KG and GoNL panels with IMPUTE2,[32] and evaluated imputation accuracy in the NTL samples. Here we obtained an additional gain in accuracy over the GoNL panel alone, reaching a mean r^2 of 0.70 for rare SNVs and 0.88 for low-frequency SNVs (**Fig. 3**). Increasing the sample size of the reference panel will likely continue to improve imputation performance (especially for lower-frequency alleles), motivating a community-wide effort to create a unified reference panel across diverse populations.

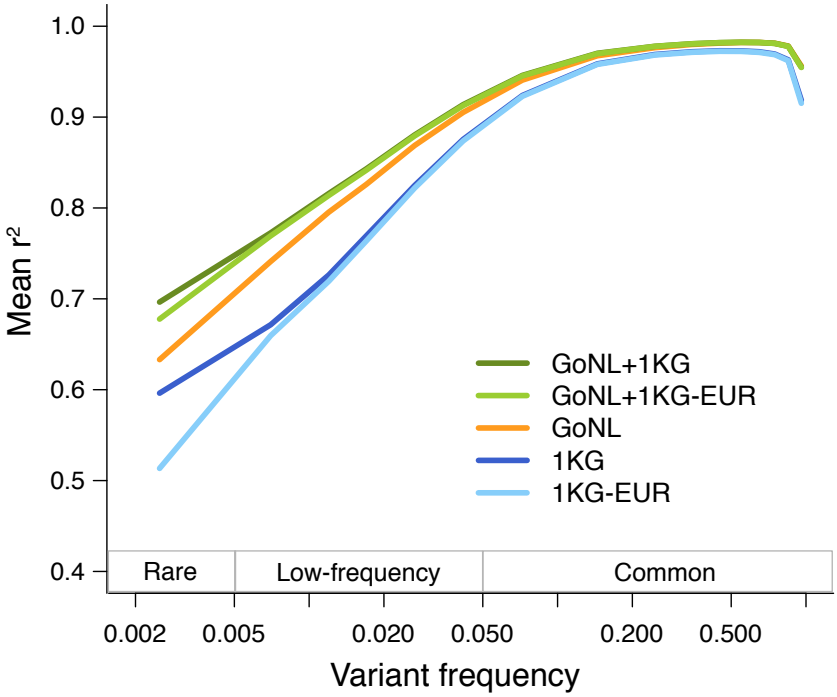


Figure 3 | Imputation accuracy. The aggregate r^2 between imputed and gold-standard genotype dosages is plotted as a function of allele frequency. We used genotypes from 81 Dutch samples (independent from GoNL) all sequenced with Complete Genomics as the gold-standard truth. The GoNL panel consistently outperforms the 1KG panels, especially at lower allele frequencies. A combined GoNL+1KG panel provides the best performance.

Population structure and demographic inference

Although it is well understood that extensive migration and gene flow occurred amongst European populations,[33–35] we focused on creating a unified picture of Dutch demography in recent millennia. Because of unbiased ascertainment and inclusion of rare variation, whole-genome sequence data can potentially offer greater resolution for demographic inference than SNP array data.

First, we explored global relationships, analyzing both common and rare variants to elucidate ancient and recent population differentiation. We calculated Hudson's F_{ST} between the Dutch and the 14 populations represented in 1KG and found that F_{ST} patterns were consistent with continental clustering in principal component analysis (PCA) and with previous estimates (**Supplementary Table 5, Supplementary Fig. 8)**.[36] To investigate more recent population connections, we focused on so-called f_2 variants, mutations appearing exactly twice (in two heterozygote carriers) in the joint data of GoNL and 1KG (**Supplementary Note**). As was observed in 1KG, within-population sharing accounts for the majority (50.8%) of all f_2 alleles (**Supplementary Fig. 9**), but f_2 sharing revealed cross-population connections as well. For example, a Dutch sample sharing an f_2 variant with a non-Dutch individual was far more likely to share that variant with another individual from Europe (71.6%) or the Americas (21.0%, due to European admixture) than with an individual from Africa (6.2%) or East Asia (1.3%). These results underscore the high degree of geographic clustering of recent mutations. Analysis of mitochondrial DNA revealed that the major haplogroups (H: 39.4%, U: 25.2%, J: 10.4%, and T: 10.8%) and minor haplogroups are in agreement with previous observations in other European populations (**Supplementary Note**).[37]

Within the Netherlands (**Fig. 4a**), PCA revealed subtle substructure along a North–South gradient (**Fig. 4b** and **Supplementary Note**), consistent with previous findings.[38,39] Because PCA cannot elucidate demography (particularly migration patterns),[40] we also performed an independent analysis of identity-by-descent (IBD) sharing that revealed subtle signals of migration (**Supplementary Note**).[41] From the length distributions of the IBD segments,[42] we inferred demographic models and estimated effective population sizes of the Dutch provinces at different time scales, reflecting historical demographic changes (**Supplementary Note**).

Analysis of IBD segments of 1–2 cM, corresponding to an estimated time-to-most-recent-common-ancestor ≈ 4,200 years, revealed homogeneous effective population sizes across the 11 provinces, consistent with common genetic origins (**Supplementary Fig. 10**). Additionally, we observed a smooth south-to-north gradient of decreasing ancestral population size and increased homozygosity

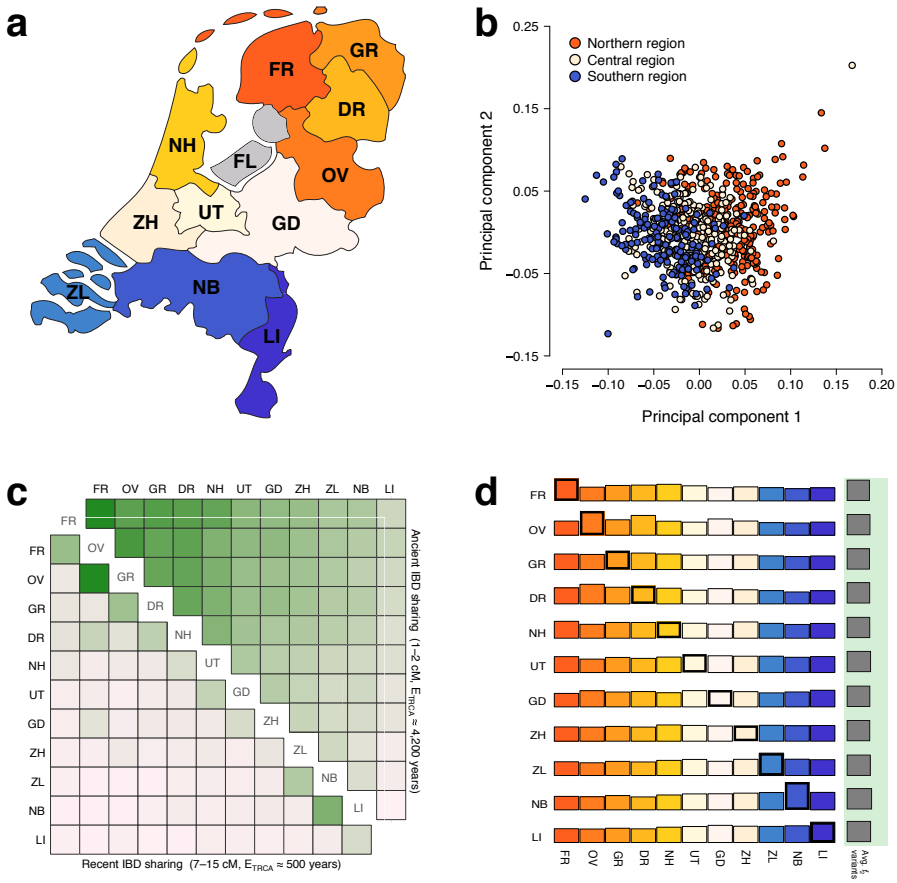


Figure 4 | Population genetic analyses in the Dutch population.

a. Map of the Netherlands with its 12 provinces. We selected 769 individuals from five BBMRI-NL biobanks across all provinces except Flevoland.

b. Principal component analysis. Individuals are projected onto the two dominant principal components, revealing subtle substructure along a North-South axis within the Netherlands.

c. Heat map of IBD segment sharing within and across provinces. The upper half represents ancient IBD sharing (1-2 cM), the bottom half represents recent IBD sharing (7-15 cM). Strikingly, all GoNL individuals, regardless of current residence, share more short IBD segments with individuals from the northern provinces than with other individuals from their own province. Long IBD segment patterns are consistent with restricted geographic movement in recent times.

d. Sharing of rare doubleton (f_2) variants within and across provinces. The level of within-province sharing of f_2 variants exceeds that of across-province sharing, reflecting strong geographically localized clustering of these recent variants. The degree of f_2 sharing amongst northern or southern provinces is statistically significant compared to central provinces ($p < 10^{-200}$).

in the northern provinces (average within-province IBD sharing and latitude correlation: r = 0.923, p = 5 x 10^{-5}; **Supplementary Fig. 10**, **Supplementary Fig. 11, Supplementary Note**). Traditionally, this observation has been explained by a serial founder effect characterized by migration from the south to the north.[38] Interestingly, GoNL samples, regardless of place of birth, tend to share more IBD segments with other individuals from the north than with individuals from the same region. Although within-province IBD sharing is strong, excess sharing with the northern provinces is evident (average between-province IBD sharing and average province latitude correlation: r = 0.934, p < 1 x 10^{-5}; **Fig. 4c, Supplementary Table 6**). This pattern indicates that a simple south-to-north serial founder model may not be sufficient to explain the observed IBD sharing (**Supplementary Note**, **Supplementary Fig. 12**). Instead, different demographic scenarios remain plausible, but all support a model of substantial regional migration. Assuming ancient serial migrations towards the North are causing the observed gradient of increasing homozygosity, a possible explanation for these results is that additional migratory events out of the North took place after initial settlements. These subsequent migratory events are consistent with the dynamic nature of the Netherlands, particularly in the northern coastal regions, between 5000 B.C.E. and 50 C.E. (**Supplementary Fig. 13**). A series of abandonments and resettlements were likely prompted by ocean level shifts and flooding that changed once-habitable land into dunes and marshes or buried regions under water entirely. We emphasize that other more complex demographic models may yield similar patterns of IBD sharing; additional analyses are required to assess alternative scenarios.

In recent centuries, the advent of water defense technologies (beginning in the 13th century) increased land stability, allowing for other forces to influence demography. An analysis of f_2 variants revealed non-random sharing within and across provinces. Though the proportion of within-province f_2 sharing comprises only 12% of all f_2 alleles, consistent with homogenous ancestry, it is significantly larger than expected under the null hypothesis of uniform allele sharing (**Fig. 4d**). This geographic localization of rare variants is suggestive of limited migration in recent centuries, consistent with current demographic studies. Notably, Noord-Brabant and Overijssel show significantly stronger within-province f_2 sharing in comparison to the other provinces (p = 1.2 x 10^{-151}, p < 10^{-200}, respectively), consistent with smaller effective population sizes in these two provinces inferred from long IBD segment sharing (**Fig 4c**, **Supplementary Table 6**, **Supplementary Fig. 9**). Further, we found that within-region sharing in the northern and southern regions was substantially stronger when compared to the central region (p < 10^{-200}, both comparisons). Altogether, these results

suggest increased migration in the central region (as compared to the northern and southern regions), consistent with recent urbanization in the wealthier central provinces.

Discussion

The results presented here reflect the enormous wealth of knowledge that can be gleaned from whole-genome sequencing data, and illustrate how intermediate-coverage sequencing within a single country complements cosmopolitan, low-coverage efforts[20]. The observed proportion of novel variation (in particular for structural variation) underlines the added value of in-depth population studies such as GoNL. Combining sequencing data sets within and across populations will not only maximize sensitivity and resolution for discovery of all types of DNA variation, but also enable population genetic analyses that can shed more light on global shared ancestry.

In spite of the intermediate coverage, we were able to reliably call *de novo* point mutations and confirm the relationship between paternal age and mutation load. We showed that we could also identify larger *de novo* events; these calls will have to be validated empirically and their properties studied across the entire cohort. The methods we developed for DNM discovery should be broadly applicable for disease studies where DNMs are suspected to play a role.[12] DNM represents an important class of DNA sequence variation that can further elucidate fundamental processes of mutagenesis.[43] In cancer, for example, accounting for the genome-wide heterogeneity of mutation rates may be necessary to accurately pinpoint driver mutations against a background of random mutation.[44] Our results suggest that trio-based sequencing of large samples at intermediate coverage may be a cost-effective way to ascertain genome-wide variation in mutation rates and establish a "null expectation" for the general population against which mutations in cases can be compared.

As long as the cost of genotyping continues to be competitive with whole-genome sequencing, imputation will remain important. The consolidation of available whole-genome data sets into a single cosmopolitan panel including low-frequency, structural and other complex types of variation[45,46] should therefore be considered a top priority. Through more complete interrogation of genetic variation, studies of large, well-phenotyped samples will continue to increase opportunities for development of diagnostic tools, prevention measures and therapeutics for human disease.

URLs

The Genome of the Netherlands Project, http://www.nlgenome.nl; European Genome-phenome Archive, http://www.ebi.ac.uk/ega/; The Groningen Center for Information Technology, http://www.rug.nl/cit; Target project, http://www.rug.nl/target; BiG Grid, http://biggrid.nl; MOLGENIS, http://www.molgenis.org

Data Access

Sequence data, variant calls, inferred genotypes, and phased haplotypes have been deposited at the European Genome-phenome Archive (EGA, see URLs), which is hosted by the European Bioinformatics Institute (EBI), under accession number EGAS00001000644.

Acknowledgements

We wish to dedicate this work to the memory of David R. Cox, an enthusiastic supporter of human genetic research in the Netherlands for many years. The GoNL Project is funded by the BBMRI-NL, a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007). We acknowledge additional financial support from eBioGrid, CTMM/TraIT, the Ubbo Emmius Fund, the Netherlands Bioinformatics Center (NBIC), and EU-BioSHARE. We thank the individual participants of the biobanks; Mark Depristo, Eric Banks, Ryan Poplin, Guillermo del Angel from the Broad Institute for expert advice on setting up our alignment and calling pipeline; Kiran Garimella for the initial implementation of PhaseByTransmission; Ger Strikwerda, Wietze Albers, Robin Teeninga, Hans Gankema, and Haije Wind of the Groningen Center for Information Technology (see URLs) for support of the compute cluster and Target storage; Edwin Valentyn and Reese Williams of the Target project (see URLs) for hosting project data on IBM GPFS storage; Tom Visser and Irene Nooren of BiG Grid (see URLs) and SURFsara for providing backup storage, additional compute capacity and expert advice; the MOLGENIS (see URLs) team for software development support; Hélène Lauvenberg for handling data access requests; Konrad Zych for the GoNL logo design; Lude Franke, Harm-Jan Westra, Javier Gutierrez-Achury for useful discussions, and Soumya Raychaudhuri and Ben Neale for their critical reading of the manuscript. Target is supported by Samenwerkingsverband Noord Nederland, European Fund for Regional Development, Dutch Ministry of Economic Affairs, Pieken in de Delta, and the Provinces of Groningen and Drenthe. Target operates under the auspices of Sensor Universe. BiG Grid and the Life Science Grid are financially supported by

the Netherlands Organization for Scientific Research (NWO). AA is funded by the Center for Medical Systems Biology-2 and DIB by the European Research Council (ERC 230374). AS and PIWdB are recipients of a VIDI Award (NWO projects 016.138.318 and 016.126.354, respectively).

Methods

Sample collection

Five Dutch biobanks (LifeLines Cohort Study, Leiden Longevity Study, Netherlands Twin Registry, Rotterdam Study, Rucphen Study) contributed samples of Dutch ancestry (with parents born in the Netherlands). A total of 250 parent-offspring families (231 trios and 19 quartets, of which 11 with monozygotic twins and 8 dizygotic twins) comprising 769 individuals were selected without phenotype ascertainment across 11 of the 12 provinces of the Netherlands (**Supplementary Fig. 4**, **Supplementary Table 7**). Limited phenotype data was available: age, sex, height, BMI, total cholesterol, HDL-cholesterol, LDL-cholesterol and triglycerides levels (**Supplementary Fig. 14**).

Data generation and processing

Samples were sequenced on Illumina HiSeq 2000 (91 bp paired-end reads, 500 bp insert size) and reads aligned on the UCSC human reference genome build 37 using BWA 0.5.9-r16[47,48] Samples were also genotyped on the Illumina Immunochip as well as at least one other genotyping chip (**Supplementary Table 8**).

Single nucleotide variants (SNV) calling

SNV calling was performed on all samples jointly using GATK UnifiedGenotyper v1.6. The calls were filtered using GATK VariantQualityScoreRecalibration (**Supplementary Note**, **Supplementary Fig. 15**) and quality metrics were evaluated (**Supplementary Note**, **Supplementary Table 9**, **Supplementary Fig. 16**). We defined the accessible genome using the same methodology as the 1000 Genomes Project[20], (**Supplementary Note**, **Supplementary Fig. 17**). We assessed the robustness of our pipeline with respect to its stochastic components, its parameters and the version of the tools by re-processing one trio under different conditions (**Supplementary Note**, **Supplementary Table 10**, **Supplementary Table 11**).

Indels and structural variants (SVs)

To create reliable indel and SV call sets, we used a combination of 10 algorithms (GATK UnifiedGenotyper, PINDEL[49], 1-2-3SV (see URLs), Breakdancer[50], DWAC (see URLs), CNVnator[51], FACADE[52], MATE-CLEVER[53], GenomeSTRiP[54], SOAPdenovo[55]). These algorithms are based on 6 approaches: (i) gapped reads, (ii) split-read, (iii) read pair, (iv) read depth, (v) combined approaches, and (vi) *de novo* assembly (**Supplementary Table 12**). Each of the 10 tools was run and their calls were filtered separately (**Supplementary Note**). The variants were divided into three groups according to their size (1-20 bp, 20-100 bp, > 100 bp) and different merging and filtering criteria were applied to obtain the final set (**Supplementary Note**).

Mitochondrial DNA

Unmapped reads were re-mapped to an appended version of the revised Cambridge Reference Sequence (rCRS)[56,57]. Consensus sequences were called using GATK and used for phylogenetic analyses. All sequences were assigned to haplogroups according to the human mtDNA phylogeny.[58] Analysis of molecular variance (AMOVA) based on provincial haplogroup frequencies was performed using Arlequin[59] v3.5.1.2 (**Supplementary Note**, **Supplementary Fig. 18**).

Validation of de novo variants

A total of 1,133 *de novo* mutations (DNMs) in 54 families were assayed using 3 sequencing technologies (**Supplementary Note**). Variants called using Sanger were analyzed manually using Phred.[60,61]. MiSeq and IonTorrent data were aligned to the reference genome using BWA and TMAP[62] aligner, respectively, and genotyped with GATK UnifiedGenotyper. Putative *de novo* indels and SVs in one trio were selected for validation. *De novo* indel candidates were sequenced using MiSeq, reads were aligned to the reference genome using BWA, and candidates were genotyped with the GATK HaplotypeCaller. *De novo* SV candidates were sequenced with Sanger and traces analyzed with NCBI BLAST[63].

Validation of polymorphisms

We randomly selected 433 deletions and 407 insertions for validation in one family considering novelty (compared to 1KG), allele frequency (rare: <0.5%, low-frequency: 0.5-5%, common: >5%), size (short: ≤ 10 bp, long: > 10 bp). Candidates were sequenced using MiSeq and reads were aligned to the reference genome with additional non-reference allele contigs for all candidates larger than (1 kb

padding). Indels < 6 bp were genotyped using GATK HaplotypeCaller, those larger were genotyped using read counts mapping to the reference and non-reference allele contigs (**Supplementary Note**).

A random set of 96 medium (20-100 bp) and 48 large (> 100 bp) deletions were assayed in one sample by Sanger sequencing (**Supplementary Note, Supplementary Table 2**). Sanger data was called using Phred[60,61] and aligned to reference genome with NCBI BLAST[63]. Medium deletions were also sequenced on MiSeq and reads aligned to the reference genome with additional non-reference allele contigs. Genotyping was based on read counts mapping to the reference and non-reference allele contigs.

Variant annotation

We functionally annotated SNVs with the Variant Annotation Tool[64] (VAT) and SnpEff[65] (**Supplementary Note**), keeping only concordant annotations in coding regions and VAT annotations in non-coding regions (SnpEff provides only coding annotations). Non-synonymous SNVs were annotated with Polymorphism Phenotyping v2 (Polyphen-2)[66] . SNVs in OMIM and "disease-causing" HGMD were annotated. SNVs were also annotated with Genomic Evolutionary Rate Profiling[67] (GERP) scores and ancestral and derived allele status. Indels were annotated using indelMapper in VAT with Gencode v16 annotations (**Supplementary Note**). SVs were annotated based on RefSeq[68] annotations and loss-of-function (LoF) annotations were defined using MacArthur et al[21] (**Supplementary Note).** Overall and per-sample variant counts stratified by novelty and functional impact were computed for all variants (**Supplementary Note**, **Table 1**, **Supplementary Table 13**).

Loss-of-Function (LoF) analyses

We computed excess LoF mutations per genome based on the expected number LoF/Synonymous ratio of common SNPs (**Supplementary Note**)[20]. To identify purifying selection of LoF variation, we tabulated counts of LoF vs. non-LoF variation, stratified by frequency. We considered LoF SNVs vs. synonymous SNVs, LoF indels vs. nonframeshift indels, and LoF SVs vs. SVs only removing intronic regions (**Supplementary Fig. 2**). Compound heterozygote LoF events were extracted and stratified by frequency (**Supplementary Table 3**). Residual Variation Intolerance Scores (RVIS)[24] were extracted for genes with LoF variants in GoNL (**Supplementary Table 3**).

De novo mutation analyses

De novo mutations (DNMs) were called using GATK PhaseByTransmission and filtered using a random forest machine-learning algorithm (**Supplementary Note**). We fit both a linear model and a log-linear model (assuming a Poisson distribution of the residuals) to the number of DNMs in the offspring given the father's age at conception, conditioning on the depth of coverage of each trio. We also tested the effect of father's age at the number of DNMs in the offspring, conditioning on mother's age at conception. Using read-phase information, parent of origin was determined using GATK ReadBackedPhasing and analyses repeated on the paternal and maternal DNMs separately (**Supplementary Note , Supplementary Fig 5**).

Integrated phased panel construction

SNV genotype likelihoods (PLs) from GATK UnifiedGenotyper were used as input into BEAGLE[69] treating all samples as unrelated to produce a first set of haplotypes. A subset of SNPs (based on the Omni2.5M array) was extracted from the BEAGLE output to construct a phased scaffold using SHAPEIT2[70] using trio information. This scaffold was used by MVNcall[19] to phase the remaining SNVs. For chromosome X, we truncated the males PLs in non-pseudo autosomal regions, yielding a negligible heterozygous genotype likelihood.

SNP discovery power and genotype concordance

SNP discovery power was estimated by comparing called sites with SNPs genotyped on the Immunochip (**Supplementary Note**, **Supplementary Fig. 19**). Genotype concordance was assessed by comparing called genotypes against (i) Immunochip genotypes in all samples and (ii) genotypes called by Complete Genomics[71] (CG) in 20 overlapping samples.

Imputation in Dutch samples

To evaluate imputation accuracy in Dutch samples, we used a set of 81 Dutch samples sequenced at ~40x coverage using Complete Genomics (CG) from the population based ALS study in the Netherlands (PAN)[72] We used CG genotypes at sites overlapping with Illumina Human-1M, 1KG and GoNL (702,253 SNPs) in these samples to impute the other CG genotypes using IMPUTE2[32,73] with 5 imputation panels: (1) GoNL, (2) GoNL + 1KG, (3) GoNL + 1KG-EUR, (4) 1KG, (5) 1KG-EUR.

Imputation accuracy was measured using the aggregate Pearson's r^2 correlation between the CG genotypes and the imputed dosages (**Supplementary Note, Fig. 3**).

Imputation with country-specific reference panels

We created 3 country-specific imputation panels of equal size (n = 85) using TSI and CEU samples from 1KG and GoNL samples. Using the different panels, we imputed into non-overlapping samples sequenced with Complete Genomics (n = 1 for TSI; n = 3 for CEU; n = 5 for NTL) (**Supplementary Note**, **Supplementary Fig. 7**).

Comparison of imputation accuracy using different chips

Using NTL samples, we evaluated GoNL imputation accuracy of 5 different chips (Illumina-Human1M, HumanExomeCore, HumanOmniExpress, Affymetrix 500k, Affymetrix 6.0) by masking and imputing all variants not on the chip (**Supplementary Note**, **Supplementary Fig. 6**).

Principal components analysis

We performed three sets of principal component analyses (PCA) using EIGENSTRAT[74]: (i) across GoNL and all 14 populations of 1KG, (ii) GoNL and 1KG-EUR, and (iii) GoNL only (**Supplementary Fig. 8**). We computed PCA in (i) on SNPs included on the Omni2.5M chip and with frequency > 5% in each individual population. We removed SNPs with missingness > 0.1% and LD-pruned the remaining SNPs ($r^2 < 0.3$). We computed PCA in (ii) following the same procedure except that we extracted sites included on the Omni1M chip. PCA in (iii) (**Fig. 4b**) was computed using SNPs in phased haplotypes with frequency > 10%, no missingness and LD-pruned ($r^2 < 0.3$). PCs were calculated in unrelated individuals only and offspring projected onto these PCs. We checked for PC significance (Tracy-Widom) and for spousal correlation along the top 10 PCs (**Supplementary Note**, **Supplementary Table 14, Supplementary Fig. 20**).

IBD-based demographic inference

We used phased SNPs (MAF >1%, phasing posterior =1.0) and kept regions >45 cM (**Supplementary Table 15**) with IBD sharing within 5 SD from the mean. IBD sharing was inferred using GERMLINE and FastIBD[41,75]. Ancestral population sizes were inferred using the average fraction *f* of genome spanned by segments of

length 1-2 cM for a pair of individuals. Recent effective size was inferred using segments >7 cM, through the estimator $\widehat{N} = 50\left(1 - f + \sqrt{1-f}\right)/(uf)$, where *u* represents the minimum centimorgan length[76]. The average time to a common ancestor was estimated using DoRIS[76]. Populations were grouped into North, Center and South using hierarchical clustering[77] based on sharing of IBD segments > 1cM. Demographic models involving a single or multiple populations with migration were analyzed using DoRIS[76,78].

Runs of Homozygosity

We used PLINK[79] to find runs of homozygous genotypes (SNPs with MAF > 5%, run length > 500 kb with at most 1 heterozygote genotype) using sliding windows of 5 Mb in unrelated GoNL and 1KG samples **(Supplementary Fig. 21**). We performed analysis of variance (ANOVA) to compare means between Dutch regions (North, Center, South) based on 1,000 bootstrap samples (**Supplementary Note**).

Population differentiation (F_{ST})

We computed Hudson's F_{ST} and Weir and Cockerham's (WC) F_{ST} between GoNL and each 1KG population. For WC estimate, we calculated F_{ST} from allele frequency data using correction for small sample size[80]. We calculated Hudson's F_{ST} estimate on two different sets of SNPs: (a) using the 1KG YRI population as an "outgroup" population, (b) sites polymorphic in the YRI population and in both populations for which the F_{ST} is calculated (**Supplementary Note**, **Supplementary Table 5**). We also computed the Hudson's F_{ST} between (i) the 11 Dutch provinces in GoNL, and (ii) between the three regions (northern, central, southern) used in the IBD and f_2 analyses (**Supplementary Note**, **Supplementary Table 16**).

Rare variant f_2 sharing analysis

We performed an inter-population f_2 analysis by merging 1KG and 88 random samples from GoNL (matching 1KG European populations size) evenly distributed amongst the 11 provinces (**Supplementary Note**, **Supplementary Fig. 9**). We performed a f_2 analysis using 330 GoNL samples selected evenly amongst the 11 provinces (**Fig. 4**). Provinces were then grouped in 3 regions (North, Center, South) and excess between provinces and regions was tested using a proportion test (**Supplementary Note**).

Singleton analysis

A filtered set of singletons was extracted from the SNPs. To account for sequencing biases, we computed the residuals of the following generalized linear regression (GLM) model: singletons per individual ~ sequencing batch + Biobank + depth of coverage + transmitted singletons. To investigate for possible geographic differences in genic singletons, we computed the Pearson's correlation between the PCs and (a) the singleton counts and (b) the residuals of the GLM above (**Supplementary Fig. 22**).

Impact of indels and SVs

We used 1,000 permutations to compute differences in the distribution of indels and SVs with respect to intergenic, intronic, exonic, OMIM and LoF annotations (**Supplementary Table 17**).

References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307,** 1072–1079 (2005).
3. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–320 (2005).
4. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
5. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
6. Manolio, T. Bringing genome-wide association findings into clinical use. *Nat. Rev. Genet.* **14,** 549–58 (2013).
7. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90,** 7–24 (2012).
8. McClellan, J. & King, M.-C. Genetic heterogeneity in human disease. *Cell* **141,** 210–217 (2010).
9. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13,** 135–145 (2012).
10. Goldstein, D. B. *et al.* Sequencing studies in human genetics: design and interpretation. *Nat. Rev. Genet.* **14,** 460–70 (2013).
11. 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).
12. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13,** 565–575 (2012).
13. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493,** 216–20 (2013).
14. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 11983–11988 (2011).
15. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44,** 243–246 (2012).
16. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* **22,** 221–7 (2014).
17. Brandsma, M. *et al.* How to kickstart a national biobanking infrastructure – experiences and prospects of BBMRI-NL. *Nor. Epidemiol.* **21,** (2012).
18. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).

19. Menelaou, A. & Marchini, J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29,** 84–91 (2013).

20. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

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

22. Tennessen, J. A. *et al.* Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* **337,** 64–69 (2012).

23. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* **44,** 623–630 (2013).

24. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet.* **9,** e1003709 (2013).

25. Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med* **1,** 13 (2009).

26. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* **132,** 1077–130 (2013).

27. Cassa, C. a, Tong, M. Y. & Jordan, D. M. Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum. Mutat.* **34,** 1216–20 (2013).

28. Dorschner, M. O. *et al.* Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am. J. Hum. Genet.* **93,** 631–640 (2013).

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

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

31. Michaelson, J. J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151,** 1431–1442 (2012).

32. 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).

33. Lao, O. *et al.* Correlation between Genetic and Geographic Structure in Europe. *Curr. Biol.* **18,** 1241–1248 (2008).

34. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).

35. Ralph, P. & Coop, G. The Geography of Recent Genetic Ancestry across Europe. *PLoS Biol.* **11,** 1001555 (2013).

36. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting F_{ST}: The impact of rare variants. *Genome Res.* **23,** 1514–21 (2013).

37. Zheng, H.-X., Yan, S., Qin, Z.-D. & Jin, L. MtDNA analysis of global populations support that major population expansions began before Neolithic Time. *Sci. Rep.* **2,** (2012).

38. Abdellaoui, A. *et al.* Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.* **21,** 1277–85 (2013).

39. Lao, O. *et al.* Clinal distribution of human genomic diversity across the Netherlands despite archaeological evidence for genetic discontinuities in Dutch population history. *Investig. Genet.* **4,** 9 (2013).

40. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40,** 646–9 (2008).

41. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19,** 318–326 (2009).

42. Palamara, P. F., Lencz, T., Darvasi, A. & Pe'er, I. Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.* **91,** 809–822 (2012).

43. Gratten, J., Visscher, P. M., Mowry, B. J. & Wray, N. R. Interpreting the role of *de novo* protein-coding mutations in neuropsychiatric disease. *Nat. Genet.* **45,** 234–8 (2013).

44. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–8 (2013).

45. Boettger, L. M., Handsaker, R. E., Zody, M. C. & McCarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat. Genet.* **44,** 881–885 (2012).

46. Jia, X. *et al.* Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLoS One* **8,** e64683 (2013).

47. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

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

49. 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).

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

51. 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).

52. Coe, B. P., Chari, R., MacAulay, C. & Lam, W. L. FACADE: a fast and sensitive algorithm for the segmentation and calling of high resolution array CGH data. *Nucleic Acids Res.* **38,** e157 (2010).

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

54. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43,** 269–276 (2011).

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

56. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290,** 457–465 (1981).

57. Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23,** 147 (1999).

58. Van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30,** E386–E394 (2009).

59. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10,** 564–567 (2010).

60. Ewing, B., Hillier, L., Wendl, M. & Green, P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.* 175–185 (1998).

61. Ewing, B. & Green, P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).

62. Wijaya, E., Frith, M. C., Suzuki, Y. & Horton, P. Recount: expectation maximization based error correction tool for next generation sequencing data. *Genome Inform.* **23,** 189–201 (2009).

63. 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).

64. Habegger, L. *et al.* VAT: A computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* **28,** 2267–9 (2012).

65. Reumers, J. *et al.* SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res.* **33,** D527–D532 (2005).

66. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7,** Unit7.20 (2013).

67. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).

68. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42,** D756–63 (2014).

69. Browning, B. L. & Yu, Z. Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies. *Am. J. Hum. Genet.* **85,** 847–861 (2009).

70. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9,** 179–181 (2011).

71. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2010).

72. Huisman, M. H. B. *et al.* Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. *J. Neurol. Neurosurg. Psychiatry* **82,** 1165–1170 (2011).

73. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 Genes|Genomes|Genetics* **1,** 457–70 (2011).

74. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).

75. Browning, B. L. & Browning, S. R. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* **88,** 173–82 (2011).

76. Palamara, P. F. & Pe'er, I. Inference of historical migration rates via haplotype sharing. *Bioinformatics* **29,** i180–8 (2013).

77. Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58,** 236–244 (1963).

78. Palamara, P. F., Lencz, T., Darvasi, A. & Pe'er, I. Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.* **91,** 809–822 (2012).

79. 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).

80. Cockerham, C. C. & Weir, B. S. Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics* **40,** 157–164 (1984).

Chapter 3

A framework for the detection of *de novo* mutations in trio sequencing data

Laurent C. Francioli*, Mircea Cretu-Stancu*, Kiran V. Garimella, Kaitlin E. Samocha, Benjamin M. Neale, Mark J. Daly, Eric Banks, Mark A. DePristo and Paul I.W. de Bakker,
Manuscript in preparation

Abstract

De novo mutation detection from human sequence data is challenging due to the rarity of such mutations relative to the error rates in sequencing technologies and the uneven coverage of the genome. We developed PhaseByTransission, an automated tool to identify *de novo* single nucleotide variants (SNVs) and short insertions and deletions (indels) from sequence data collected in 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 are reported along with their posterior probability, providing a systematic and automated way to prioritize them for validation. Using simulated data, we show that our approach outperforms existing tools, especially at lower coverage. We also present results on deep-coverage whole-exome data collected in 104 trios and show a specificity of 93% based on experimental validation.

Introduction

De novo mutation between generations is a key mechanism in evolution. In humans, the mutation rate is estimated between 1×10^{-8} and 3×10^{-8} per base per generation from direct observations [1–4] and from species comparisons [5] . These estimates are averages across the entire genomes of multiple individuals and *de novo* mutation rates have been shown to vary locally [2,6] and across families [2–4] . *De novo* mutations, while mostly selectively neutral, may have severe phenotypic consequences when affecting functional elements in the genome [7] and are thus of critical interest for medical genetics.

Next generation sequencing (NGS) technologies applied to whole genomes or exomes in pedigrees enable systematic discovery and analysis of *de novo* mutations. Because of the coverage fluctuations and error rates inherent to NGS technologies [8] , detecting *de novo* mutations 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 either parents or offspring may lead to a false positive or false negative result. Variant callers such as the UnifiedGenotyper in the Genome Analysis Toolkit (GATK) [9,10] emit genotype likelihoods for each possible genotype to incorporate the uncertainty from the NGS read-level data.

We developed a GATK module called PhaseByTransmission 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 a prior probability for the mutation rate. As output, a list of all candidate *de novo* sites ranked by their posterior probability is generated. PhaseByTransmission works on bi-allelic single nucleotide variants (SNVs) and short insertions and deletions (indels), and supports autosomes and the X-chromosome.

Model implementation

As input, Phase By Transmission takes individual genotype likelihoods, defined as the likelihood *L* of the bases *D* observed at a site given each bi-allelic genotype *G*: $L(D|G)$. These likelihoods can be computed from the sequence data using different genotype calling algorithms, such as the GATK UnifiedGenotyper, GATK HaplotypeCaller or Samtools [11] .

Given a trio and a site in the genome, we can enumerate all possible genotype combinations in the trio. For bi-allelic autosomal sites, there are 27 possible genotype combinations within a trio: 15 consistent with Mendelian inheritance, 10 implying 1 *de novo* mutation and 2 implying 2 *de novo* mutations. For bi-allelic sites on the X chromosome of a female offspring, there are only 18 possible genotype combinations because the father is haploid: 8 consistent with Mendelian inheritance, 8 implying 1 *de novo* mutation and 2 implying 2 *de novo* mutations. Because male offspring are haploid on the X chromosome and inherit their X chromosome from their mothers, there are only 6 relevant mother-offspring genotype combinations: 4 consistent with Mendelian inheritance and 2 implying one *de novo* mutation.

Given a mutation rate μ, *a* genotype combinations implying one *de novo* mutation and *b* genotype combinations implying 2 *de novo* mutations, we define the following genotype combination prior:

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

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

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

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 (denoted x here) to the trio at this site and compute its posterior probability as:

$$P(D|G_M^x,G_F^x,G_C^x) = \frac{P(D|G_M^x,G_F^x,G_C^x)}{\sum_{i=1}^{N} P(D|G_M^i,G_F^i,G_C^i)} \quad (3)$$

where i corresponds to the i^{th} genotype combination amongst the N possible ones. All sites and trios assigned a genotype combination violating Mendel's laws are reported as putative *de novo* mutations and the posterior probability assigned to each of them reflects the confidence of the call.

In addition to the familial relationship of samples, population information in the form of allele frequencies can be used as a prior in our model. Because one of the most common error modes when detecting *de novo* mutations is to miss the non-reference allele in (one of) the parents, informing the model about allele frequencies in the population allows to quantify the probability of missing the non-reference allele in each parents. When adding allele frequency priors, equation (2) becomes:

$$P(D|G_M,G_F,G_C) = P_C \cdot P_{AF}^{G_M} \cdot P(D|G_M) \cdot P_{AF}^{G_F} \cdot P(D|G_F) \cdot P(D|G_C) \quad (4)$$

where G_M, G_F and G_C are the genotypes of the mother, father and child, $P_{AF}^{G_M}$ and $P_{AF}^{G_F}$ 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 from an external source (as a separate VCF file) or computed from the genotype likelihoods of the input samples when multiple samples from a single population are studied. In this case, we allele frequencies are estimated as P_{AF}^{G} for each genotype G following Hardy-Weinberg expectations:

$$P_{AF}^{G} = \begin{cases} p^2 & \text{if the genotype } G \text{ is homozygous reference} \\ 2pq, & \text{if the genotype } G \text{ is heterozygous} \\ q^2 & \text{, if the genotype } G \text{ is homozygous alternative} \end{cases} \quad (5)$$

where p and q are the estimated allele dosage for the reference and alternate alleles, respectively, in the parents (founders).

Evaluation on simulated data

We simulated sequencing data for 10 parent-offspring trios, 5 with a male offspring and 5 with a female offspring and evaluated the sensitivity and specificity of the calls made by PhaseByTransmission as well as two recently described *de novo* mutation callers: TrioDeNovo [12] and DeNovoGear [13] . We first created the parents haplotype using real haplotypes from 10 unrelated samples from the Genome of the Netherlands (GoNL) Project [4] . We then created haplotypes for the children by randomly selecting one haplotype from each of the parents and adding 10k autosomal *de novo* single nucleotide variants (SNVs) and 2k *de novo* SNVs on the X chromosome of these haplotypes. In order to place *de novo* mutations realistically, we used the GoNL mutation map (*manuscript submitted*) to derive probabilities for each of the bases to be mutated. This mutation map covers 75% of the human genome and provides mutation rate estimates at the mega base scale for each substitution types, as well as for C>T transitions in a CpG context. Because *de novo* SNVs exhibit a pronounced paternal bias [2–4] , we assigned 70% of them to the paternal haplotype and 30% of them to the maternal haplotype. We used SimSeq [14] to simulate 100bp Illumina paired-end reads with an insert size of 250bp for all 30 samples in 10kb regions centered on each simulated *de novo* SNV (5kb upstream and 5kb downstream) with an average coverage of 60x. The reads were aligned to the UCSC human reference sequence build 37 using BWA [15] to produce aligned BAM files. To evaluate the effect of coverage on *de novo* SNV detection, the generated BAM files were downsampled such that we obtain variant call sets for average depths of coverage of 60x, 30x and 15x for each sample..The For each of the coverage depths, SNVs were called using the GATK UnifiedGenotyper on each trio separately to the genotype likelihoods used as input to PhaseByTransmission.

After SNV calling, each trio comprised on average 161k inherited SNVs and 8.3k *de novo* SNVs (72% of the *de novo* mutations simulated). PhaseByTransmission, TrioDeNovo and DeNovoGear were then all run on the input VCF files. PhaseByTransmission and DeNovoGear use a mutation prior as a parameter; this

prior influences the sensitivity and specificity of the *de novo* calls. We therefore ran these tools using mutation priors of $1.5x10^{-8}$, 10^{-7}, 10^{-6}, 10^{-5} and 10^{-4}. We ran PhaseByTransmission with and without allele frequency priors. When using allele frequency priors, we obtained the allele frequency estimates from 1000 Genomes Phase 3 CEU data. After running all tools, we extracted the *de novo* mutations (DNMs) called by each of the tools and computed the sensitivity as $\frac{\#\text{DNMs called as DNM}}{\#\text{DNMs in input file}}$ and the specificity as $\frac{\#\text{inherited SNVs called as inherited}}{\#\text{inherited SNVs in input file}}$. It should be noted that the numbers presented here reflect the sensitivity and specificity with respect to the simulated mutations detected by GATK UnifiedGenotyper, so the true detection rate from the data is on average 38% lower. Figure 1 shows the receiving operator characteristic (ROC) curves for the autosomes and the X chromosome with different coverage depths. Each of the points on these ROC curves represents the result of one of the tools using one of the 5 mutation priors.

Results show that the mutation prior for both PhaseByTransmission and DeNovoGear has a substantial influence on the sensitivity at 15x coverage, whereas at higher coverage its influence on sensitivity gradually becomes marginal. The mutation prior has an influence on the specificity at all coverage depths with larger number of false positive results for higher mutation priors. PhaseByTransmission outperforms both TrioDeNovo and DeNovoGear in sensitivity and specificity at 15x coverage when the mutation prior is set properly. For higher coverages, PhaseByTransmission and TrioDeNovo show similar performance while DeNovoGear constantly exhibit larger numbers of false positives.

Results on whole-exome data

We evaluated the software on whole exome data in a cohort of 104 trios sequenced at 60x depth on the Illumina HiSeq platform for an independent autism study [16] . The sequence data were aligned to the human reference hg19 using BWA, duplicate reads removed, re-alignment performed around insertions/deletions, and base quality scores recalibrated. Variant discovery and genotyping was performed using the GATK Unified Genotyper across all samples simultaneously, and calls were subsequently filtered using Variant Quality Score Recalibration (VQSR).

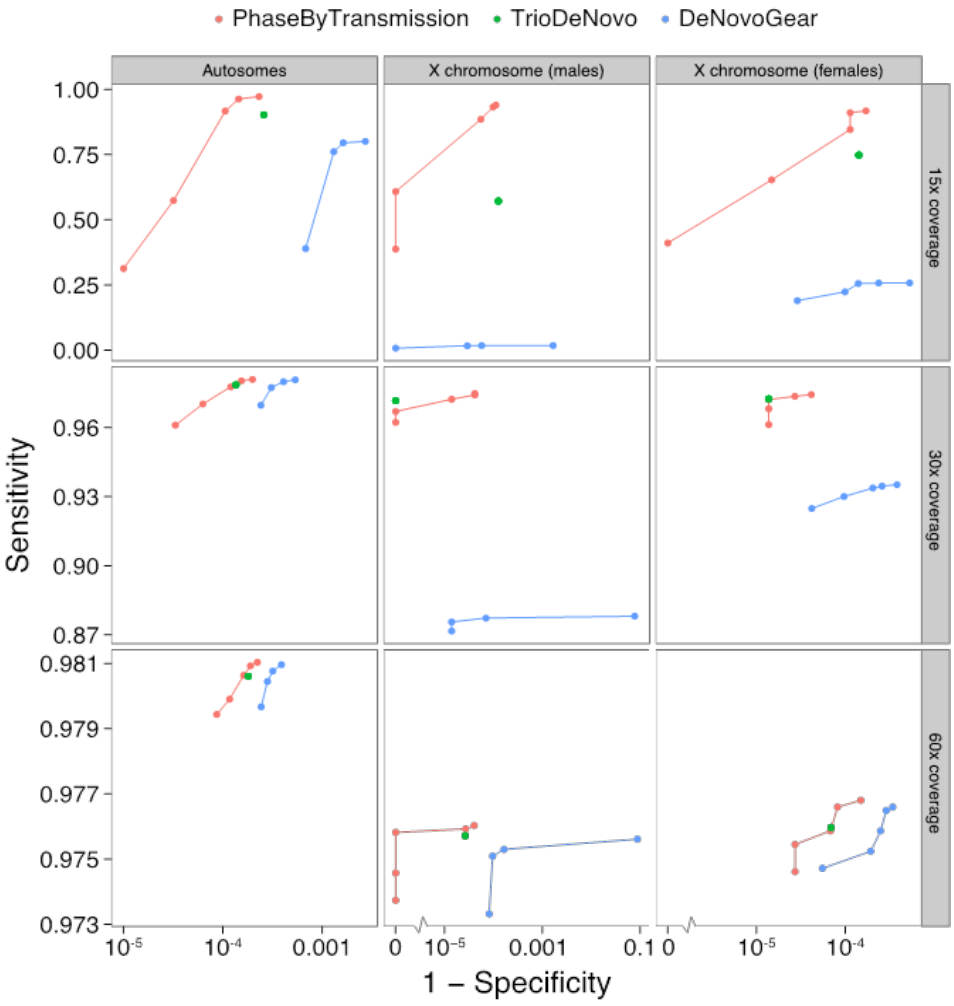


Figure 1 | Evaluation on simulated data

Receiving operator characteristic (ROC) curves evaluating PhaseByTransmission (PBT, pink), TrioDeNovo (TDN, green) and DeNovoGear (DNG, blue). The different points for PBT and DNG represent different mutation priors parameters. The data used is simulation data for 10 trios (5 with male and 5 with female offspring), each with 10k simulated *de novo* mutations in the autosomes and 2k simulated *de* novo mutations on the X chromosome. At 15x coverage, PBT outperforms both TDN and DNG when the mutation prior is set appropriately. At higher coverage depths, results are less sensitive to the mutation prior and PBT and TDN show similar performance, both substantially better than DNG.

PhaseByTransmission (PBT) was run using a mutation prior of 10^{-7} to maximize sensitivity and using allele frequency prior computed from the data. In each case it took about 90 minutes on one core to process the entire dataset. In total, PBT called 148 putative *de novo* single nucleotide variant candidates. All candidates were subjected to experimental validation using Sequenom, and 115 (77.8%) could be assayed successfully. From these, 107 (93%) mutations were validated as *de novo* in the offspring. 5 (4.7%) were monomorphic in all samples and 3 (2.8%) were inherited mutations.

Conclusion

PhaseByTransmission is an accurate and automated *de novo* mutation caller using a probabilistic model to estimate the probability of a *de novo* mutation at each site. Because PhaseByTransmission accepts VCF files as input, it can easily be integrated in an existing NGS analysis pipeline. It is computationally inexpensive, and scales linearly with the number of sites and trios. When compared to other *de novo* mutation callers, it performs better, especially at lower coverage depth. Results on whole-exome sequencing data show excellent specificity and sensitivity on SNVs. Finally, PhaseByTransmission use has been demonstrated in whole-genome intermediate coverage data on both SNVs [4] and indels (*manuscript submitted*).

PhaseByTransmission is available as part of the Genome Analysis Toolkit (GATK) as a precompiled Java package as well as source code at http://www.broadinstitute.org/gatk/download.

References

1. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43,** 712–4 (2011).
2. Michaelson, J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151,** 1431–42 (2012).
3. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–5 (2012).
4. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–25 (2014).
5. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304 (2000).
6. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12,** 756–66 (2011).
7. Veltman, J. & Brunner, H. De novo mutations in human genetic disease. *Nature reviews. Genetics* **13,** 565–75 (2012).
8. Glenn, T. C. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11,** 759–69 (2011).
9. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297–303 (2010).
10. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).
11. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–93 (2011).
12. Wei, Q. *et al.* A Bayesian framework for de novo mutation calling in parents-offspring trios. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu839
13. Ramu, A. *et al.* DeNovoGear: de novo indel and point mutation discovery and phasing. *Nature methods* **10,** 985–7 (2013).
14. Earl, D. *et al.* Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome research* **21,** 2224–2241 (2011).
15. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–95 (2010).
16. Neale, B. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485,** 242–5 (2012).

Chapter 4

Genome-wide patterns and properties of *de novo* mutations in humans

Laurent C. Francioli*, Paz P. Polak*, Amnon Koren*, Androniki Menelaou, Sung Chun, Ivo Renkens, Genome of the Netherlands Consortium, Cornelia M. van Duijn, Morris Swertz, Cisca Wijmenga, Gertjan van Ommen, P. Eline Slagboom, Dorret I. Boomsma, Kai Ye, Victor Guryev, Peter F. Arndt, Wigard P. Kloosterman, Paul I. W. de Bakker*, Shamil R. Sunyaev*, *Accepted at Nature Genetics*

Mutations create variation in the population, fuel evolution, and are the source of genetic disease. Our current knowledge on genome-wide patterns of *de novo* mutations and underlying biological mechanisms is incomplete and mostly derived from indirect sources [1–10] . Recent pedigree sequencing studies have begun characterizing properties of *de novo* mutations [11–13] and confirmed that mutation rate increases with father's age. Here, we analyze 11,020 *de novo* mutations from whole-genome sequences of 250 Dutch families and reveal novel insights into human germline mutagenesis. First, the age-related increase in paternal mutations is accompanied by a change in the distribution of their chromosomal location. That is, mutations in offspring of younger fathers occur more frequently in late-replicating, gene-poor regions, while a higher fraction of mutations in children of older fathers fall in early-replicating, genic regions. Second, mutation rates in functional regions are elevated due to sequence context influences and exhibit pronounced strand asymmetry compatible with the action of transcription-coupled repair. Third, mutation clusters within individuals account for 1.5% of all mutations and exhibit a unique mutational spectrum suggesting their formation via a novel mechanism. Fourth, mutation and recombination rates are independently associated with nucleotide diversity within populations, and genomic variation in the rate of evolution is only partly explained by heterogeneity in mutation rate. Finally, we provide the first empirical genome-wide human mutation rate map as a resource for medical and population genetics. Our results point to a novel link between epigenomic properties and mutation rates, underscore the importance of father's age in disease susceptibility, and refine long-standing hypotheses about mutation properties during evolution.

Understanding rates and patterns of human mutations is important for analyzing relationships among species and populations [1,2] , for detecting natural selection [3,4] and for mapping genes underlying complex traits [5] . The properties of mutations have traditionally been studied using model organisms [6] , fully penetrant dominant Mendelian diseases [7,8] , and comparative genomics and population genetic approaches [9,10] . However, these approaches are limited in scope, indirect, and influenced by other factors such as natural selection. Using high-throughput sequencing technologies, recent pedigree sequencing studies have provided whole-genome observations of germline *de novo* mutations and revealed that mutation rate increases with paternal age [11–13] , varies along the genome in weak correlation with various epigenetic properties and is higher in conserved genomic regions including exons [11] .

We identified *de novo* mutations of 250 parent-offspring families from the Netherlands (231 trios, 11 families with monozygotic twins, 8 families with dizygotic twins) by whole-genome sequencing of DNA from blood at an average 13-fold coverage. We considered dizygotic twins to be genetically distinct and included one twin at random from each monozygotic twin pair, resulting in 258 children. Comparison of variant calls between parents and offspring resulted in the identification of 11,020 *de novo* mutations, with an estimated sensitivity of 68.9% and specificity of 94.6% [13] . By comparing 350 validated novel mutations in monozygotic twins, we estimate that ~97% of the mutations in our data are germline and ~3% are somatic. To account for mutation calling biases inherent to sequencing data, we generated a simulated set of *de novo* mutations taking into account the sequence coverage fluctuations (Methods). We used this simulated set as a "null" baseline against which we compared the observed *de novo* mutations to characterize their patterns and properties.

Paternal age explains about 95% of the variation in global mutation rate in the human population [12] . Specifically, there is an increase of one to two mutations per year of paternal age [11–13] , which is thought to stem from continuous cell divisions in the paternal germ line, beginning in the embryonic development of primordial germ cells and continuing in spermatogenesis throughout a man's life. A key question is whether changes in the global mutation rate are accompanied by a shift in the mechanisms of spontaneous mutagenesis. If so, this might be reflected by the genomic distribution of *de novo* mutations. Because previous studies suggested that the epigenetic landscape varies with age [14] , we used a linear regression model to investigate whether paternal age was associated with the location of *de novo* mutations with respect to various epigenetic variables (Methods). We found that the replication timing of *de novo* mutations showed a significant association with paternal age ($p = 0.0022$, **Fig. 1a**), while chromatin accessibility, chromatin modifications and recombination rate did not ($p > 0.098$, **Supplementary Fig. 1**). Mutations in the offspring of younger fathers (< 28 years old) were strongly enriched in late-replicating genomic regions ($p = 4.9 \times 10^{-4}$; **Fig. 1b**), while there was no significant replication timing bias for mutations in the offspring of older fathers (≥28 years old; $p = 0.68$; **Fig. 1b**). The age groups were chosen to maximize the difference in replication timing between the mutations in offspring of younger and older fathers ($p = 5.7 \times 10^{-4}$, 23 tests, Supplementary **Fig. 2**). The age-dependent association between the distribution of mutations and DNA replication timing was not specific to the replication timing dataset we analyzed nor the cell type in which it was measured (**Supplementary Fig. 3**).

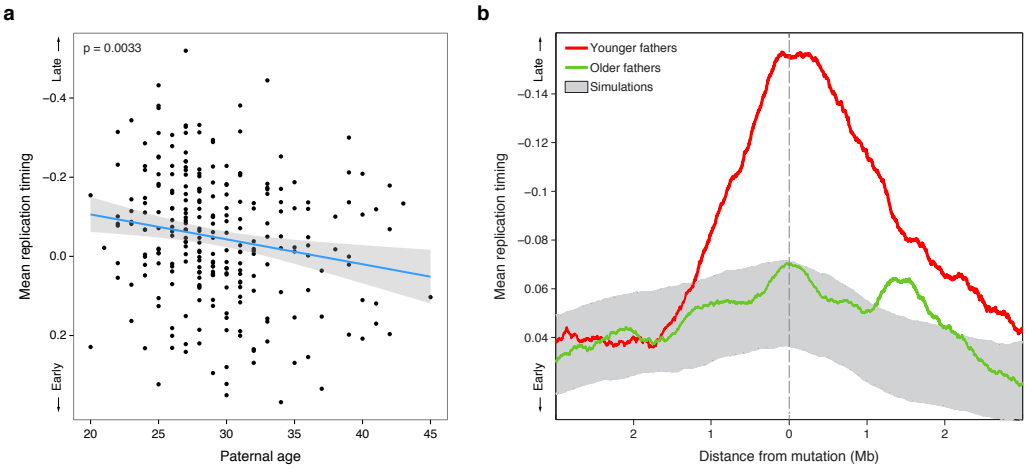


Figure 1 | Mutations in offspring of younger fathers are biased towards later replicating regions

a. Mean replication timing of *de novo* mutations in each of the 258 offspring as a function of their father's age. The blue line shows the least-square regression line (p = 0.0033) and the grey area the 95% confidence interval. The downward slope of the regression line indicates a shift of mutations towards earlier replicating regions with advancing paternal age.

b. The mean replication timing profile around *de novo* mutations, stratified by paternal age (red: under the age of 28, N = 3,697, green: aged 28 or older, N = 7323). The grey area shows the null expectation based on simulations (mean ±1 standard deviation). The age of the split between younger and older fathers was chosen to maximize the difference between the groups ($p = 5.7 \times 10^{-4}$, 23 tests). Mutations in younger fathers tend to be located in large (~2Mb) regions of late-replicating DNA. In contrast, the replication timing distribution of mutations in older fathers is similar to that of simulated mutations. Together, this shows that *de novo* mutations in offspring of younger fathers are biased towards late-replicating regions, while those in offspring of older fathers aren't.

To confirm that our results are due to paternal age rather than maternal age (which are highly correlated with each other within families; r = 0.81), we restricted our analysis to mutations with unambiguous paternal or maternal origins (Methods). Consistent with the results above, paternal age was significantly associated with replication timing (N = 1,991, p = 0.032) but maternal age was not (N = 630, p = 0.26). The difference between paternal and maternal age was significant (p = 0.0019), controlling for the different numbers of mutations assigned to the maternal or paternal lines (**Supplementary Fig. 4**).

Replication timing itself correlates with chromatin structure and gene activity: early-replicating regions of the genome have a higher gene density and elevated gene expression levels compared to late-replicating genomic regions [15] . Therefore, the paternal age effects described above are likely to have functional consequences. Indeed, we found that the proportion of *de novo* mutations in genic regions increased by 0.26% with each additional year of paternal age (p = 0.0085; **Fig. 2**). On average, offspring born to 40 year-old fathers harbored twice as many genic mutations than offspring of 20 year-old fathers (19.06 vs 9.63 mutations), but only 55% more intergenic mutations (35.24 vs 22.68). An important implication of this result is that mutations in older fathers are not only more numerous, they are also individually more likely to be functional.

Taken together, these observations suggest that the increase in the number of *de novo* mutations with paternal age is accompanied by a change in their mechanism of formation related to DNA replication timing, and consequently, their chromosomal distribution and functional impact. While the source of these differences is currently unclear, they could reflect variations in replication timing or mutagenesis between the symmetrical mitoses that occur during the generation of paternal germ cells, and the asymmetrical mitoses of the spermatogonia during spermatogenesis.

Irrespective of paternal age, mutation rates are higher in functional regions of the genome [11]. Indeed, 1.22% of *de novo* mutations were exonic, which represents a 28.7% enrichment compared to our simulated null baseline (p = 0.008). Similarly, mutation rates in regulatory regions marked by DNase I hypersensitive sites (DHS) were elevated (p = 0.005). The elevated mutation rate for both exons and DHS appeared to be driven by CpG dinucleotides, since after excluding CpGs we observed no significant difference from the null expectation. Methylated CpGs represent highly mutable sequences in humans due to the spontaneous deamination of cytosine bases. The increased mutation rates at CpG sites are thought to have evolved recently (around the time of mammalian radiation) [16] . Thus, while sequences of neutrally evolving regions of the genome have had sufficient time to equilibrate with respect to dinucleotide contexts, purifying selection has maintained the hypermutable CpG context in functional regions of mammalian genomes [17,18] .

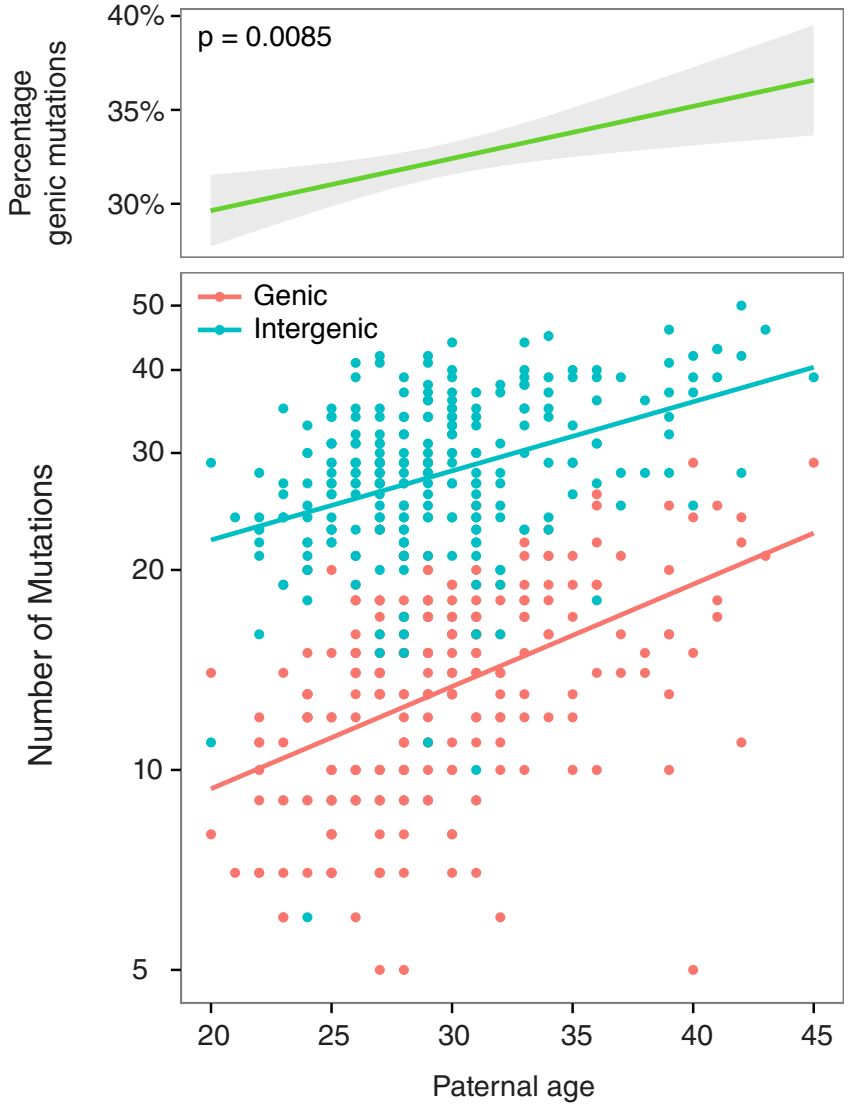


Figure 2 | Offspring of older fathers harbor a higher percentage of *de novo* mutations in genes.

Top panel: the percentage of *de novo* mutations within genic regions as a function of paternal age at conception ($p = 0.0085$, slope = 0.26% per year of paternal age).

Bottom panel: the number of genic (red) and intergenic (blue) *de novo* mutations in offspring (on a logarithmic scale) as a function of paternal age. The red line shows the least-square regression for genic mutations ($p < 2 \times 10^{-16}$), the blue line for intergenic mutations ($p = 3.7 \times 10^{-14}$). The steeper slope of the regression line for genic mutations indicates a faster relative increase in genic than intergenic mutations with paternal age.

The above result is discordant with the reduction of cancer somatic mutation density in transcribed regions and in regulatory regions marked by DHS [19,20] and with comparative genomics studies[9] that reported a reduction of sequence divergence between humans and non-human primates within DHS after correction for sequence context. While our study may lack sufficient power to detect this effect, it is also possible that comparative genomics observations reflect the effect of purifying selection on regulatory sequences[9] .

The distribution of the *de novo* mutations along the genome was non-random, both within and across individuals (**Fig. 3a**) beyond correlations with epigenetic variables and functional elements. At the extreme, we observed clusters of nearby mutations in an individual. This clustering was particularly strong for distances of up to 20kb ($p < 1 \times 10^{-6}$), at which there were a total of 78 clusters of 2-3 mutations. These observations are consistent with, and expand on, previous studies based on more limited data [11,21] . We did not find a significant difference between the 161 clustered mutations and the 10,859 non-clustered mutations with respect to recombination rates ($p = 0.52$) or replication timing ($p = 0.059$). Interestingly, mutations within clusters exhibit a unique spectrum ($p = 9.7 \times 10^{-16}$), with reduced transitions and strongly elevated C→G nucleotide changes (**Fig. 3b**), suggesting a specific underlying mechanism. This is a distinct signature from the previously observed same-strand TCW→TTW or TCW→TGW mutations (where W corresponds to either A or T) reminiscent of the activity of the APOBEC cytosine deaminases that leads to clustered mutations in cancer cells [22,23] . Although not caused by APOBEC activity, C→G mutations may result from deaminated cytosines in single-stranded DNA that would be converted to apurinic sites by base-excision repair DNA glycosylases and subsequently subjected to error-prone translesion DNA synthesis [24] .

Comparative genomics studies have predicted that mutation rate is variable at the megabase scale [9,25] . However, the extent to which the mutation rates and patterns predicted from comparative genomic studies reflect the true underlying properties of germline mutations is unknown. Here, we sought to separate intrinsic properties of mutational processes from other population processes that lead to allele transmission biases, such as background selection, hitchhiking, and biased gene conversion.

Previous studies have shown that nucleotide diversity within populations (π) is correlated with local recombination rate but it is unclear whether this is due to a mutagenic effect of recombination [26] or to background selection and hitchhiking mechanisms [27,28] . In our study, local recombination rates [29] are significantly associated with *de novo* mutation rates ($p = 0.0015$), when controlling for CpG sites and GC content. Despite this association, we found that rates of both

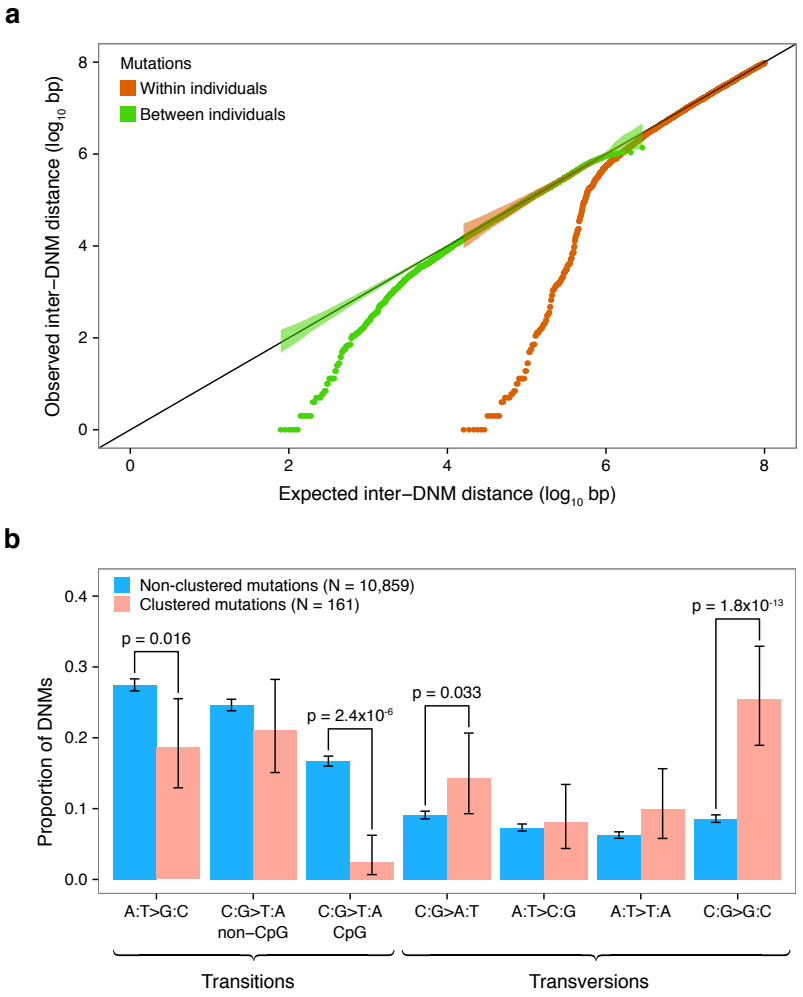


Figure 3 | Mutation clusters exhibit a unique mutational spectrum

a. The distances between adjacent *de novo* mutations (observed) compared to a uniform distribution of mutations across the genome (expected). Closely spaced mutations are enriched both across individuals (brown) and within individuals (green). The strength of this effect is strongest within individuals, where 78 mutation clusters of up to 20kb in size are observed. In fact, 1.5% of all *de novo* mutations in our study are in such clusters. Shaded areas represent the 95% confidence intervals.

b. Comparison of mutation spectra between clustered (pink) and non-clustered (blue) *de novo* mutations. We defined mutation clusters as regions with two or more mutations within 20kb in the same individual. Mutations within clusters show a significantly reduced number of transitions ($p = 1.2 \times 10^{-12}$ for all transitions, $p = 4.1 \times 10^{-6}$ when excluding C>T transitions at CpG sites) and a strongly elevated number of C→G transversions ($p = 1.8 \times 10^{-13}$), indicating a novel mutational mechanism.

mutation ($p < 2 \times 10^{-16}$) and recombination ($p < 2 \times 10^{-16}$) independently contribute to nucleotide diversity. Thus, recombination appears to influence nucleotide diversity above and beyond its mutagenic effect.

Next, we estimated the extent to which human-chimpanzee sequence divergence is influenced by mutation rates and recombination rates (**Fig. 4**). The correlation between substitution rates from a human-chimpanzee comparative genomics (HCCG) model [30] and observed *de novo* mutations was significant ($r = 0.18$, $p = 1.3 \times 10^{-15}$). When compared to mutation rates based on sampling the HCCG itself for the same number of mutations (mean $r = 0.33$), we found that *de novo* mutation rates explained about a third of the human-chimpanzee sequence divergence along the genome (Methods). However, observed mutation rates adjusted for local recombination rates [31] are more strongly correlated with the HCCG model ($r = 0.37$) than observed mutation rates alone (**Fig. 4**). This illustrates that the comparative genomic model captures both variation in mutation rate and other, orthogonal evolutionary forces associated with recombination rate, as has been suggested by others [27,32] .

In contrast to the large-scale regional variation, we found that the influence of sequence context (flanking nucleotides) on *de novo* mutations was in excellent agreement with results based on comparative genomics [33] ($r^2 = 0.993$; **Supplementary Fig. 5**), suggesting that the mutation spectrum has been relatively constant in recent evolution. We also observed a previously predicted[26–28] strand asymmetry for mutations in transcribed regions (**Supplementary Fig. 6**), especially for A→G mutations ($p = 5.9 \times 10^{-5}$). This is likely a byproduct of the action of transcription-coupled repair. We found a modest 2.8% depletion of mutations in transcribed regions relative to intergenic regions ($p = 0.047$). This is in sharp contrast with somatic cancer mutations where a similar strand asymmetry was accompanied by a strong reduction of mutations in transcribed regions [20] .

Having a well-calibrated mutation model is essential for evaluating the significance of *de novo* mutation patterns observed in pedigree sequencing studies (especially in the absence of appropriate controls in disease studies) [34] . Previous mutation models have been based on comparative genomics, but, as shown above, these models are not representative of germline mutation rates alone as they also incorporate other evolutionary forces. To bridge this gap, we used the empirical distribution of *de novo* mutations along the genome to refine a mutation model based on human-chimpanzee divergence rates, considering flanking sequence context, local recombination rates, mutation type and transcriptional strand in coding regions (Methods). In addition to the

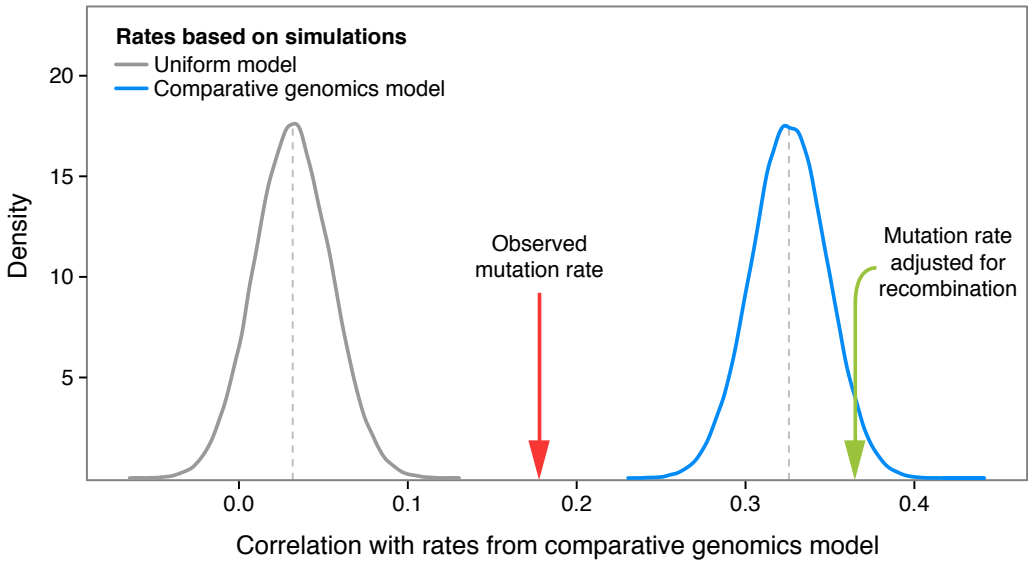


Figure 4 | The influence of mutation and recombination rates on human-chimpanzee divergence.

The correlation to substitution rates computed from a human-chimpanzee comparative genomics (HCCG) model is plotted for mutation rates inferred from a uniform mutation rate model (grey distribution) and for mutation rates inferred from the HCCG model itself (blue distribution), both based on 100,000 simulations of N = 11,020 mutations and binned in 1Mb windows. By sampling the same number of mutations, we can ensure that comparisons between rate estimates are meaningful. The effect of the sampling is illustrated by the mean correlation of the HCCG with itself at only 0.33, which would asymptotically reach unity with infinite sampling. The correlation is also given for the observed *de novo* mutation rate (red dot, N = 11,020 mutations) and the observed *de novo* mutation rate adjusted for local recombination rates [31] (green dot, N = 11,020 mutations). Correlation with the observed *de novo* mutation rate (r = 0.18) is stronger than correlations with rates based on the uniform model (mean r = 0.032, $p < 1 \times 10^{-5}$), indicating that the HCCG model captures to some degree that mutation rates are regionally variable. However, the correlation with the observed *de novo* mutation rates is weaker than correlations with rates based on the HCCG model itself (mean r = 0.33, $p < 1 \times 10^{-5}$), suggesting that other factors also contribute. When adjusting the observed *de novo* mutation rate for local recombination rates, the correlation is 0.37, which illustrates that substitution rates computed from the HCCG model capture both mutation rates as well as orthogonal evolutionary forces such as local recombination rates.

genome-wide rates, we also calculated gene-level mutation rates, separately estimating synonymous, missense and nonsense mutation rates. This mutation rate map will be instrumental for evolutionary inferences based on human mutation rates and patterns and for the identification of disease genes with recurrent *de novo* mutations.

We describe here the most extensive catalog to date of *de novo* germline mutations in healthy individuals, revealing several mechanisms influencing the distribution of mutations along the genome. In particular, clustered mutations suggest the existence of a novel mutagenic mechanism, and the effect of replication timing on germline mutations depends on paternal age. Mutation rate heterogeneity substantially influences genomic variation in the rate of sequence evolution, adding to the effects of evolutionary forces acting at the population level.

Methods

The Genome of the Netherlands data

This study uses *de novo* mutation data from the Genome of the Netherlands (GoNL) project, for which all data generation and processing steps were detailed in a previous publication [13] . A brief version is included here.

The Genome of the Netherlands project includes 250 Dutch parent-offspring families (231 trios, 8 quartets with dizygotic twins, 11 quartets with monozygotic twins) sampled throughout the Netherlands without phenotypic ascertainment. For this study, we used all 250 parents as well as 258 genetically unique offspring, removing one of the two twins (chosen randomly) in each of the monozygotic twin pairs.

Samples were sequenced using 91bp paired-end with 500 insert size libraries on Illumina HiSeq2000. The alignment and variant calling were devised based on the Genome Analysis Toolkit (GATK) best practices v2 [35,36] : The sequence data were mapped to the human reference genome build 37 using bwa 0.5.9-r16 [37] , duplicate reads were removed using Picard tools (http://picard.sourceforge.net), local indel realignment was performed around indels using GATK IndelRealigner and base qualities were recalibrated using GATK BaseQualityScoreRecalibration. Variants were called using GATK UnifiedGenotyper v1.4 on all samples simultaneously and filtered using GATK VariantQualityScoreRecalibration.

De novo mutation detection was performed using the trio-aware genotype caller GATK PhaseByTransmission which leverages familial, population and mutation rate,

followed by filtering using a random forests machine-learning classifier (trained on 592 true positives and 1,630 false positive putative *de novo* mutations validated experimentally). We obtained a set of 11,020 high confidence mutations in the 269 children of the GoNL project with an estimated 92.2% accuracy [13] . All putative *de novo* mutations found in the 11 monozygotic twin quartets were subjected to validation in both twins. Out of the 680 mutations detected and validated in either twin, 660 were shared by both twins and 20 were unique to a single child. We therefore estimate that 97% of the mutations in our data are germline and 3% are somatic. Using GATK ReadBackedPhasing we assigned parental origin to 1,991 paternal and 630 maternal *de novo* mutations based on phase-informative reads.

Simulation of *de novo* mutations

We simulated *de novo* mutations at the read level to create a null distribution (uniform) while accounting for the effect of coverage fluctuation inherent to high-throughput sequencing. We generated 264k random positions throughout the GoNL accessible genome [13] (i.e. ~1/1000bp), excluding any position that was polymorphic in GoNL or outside the accessible genome. For each of these positions, we generated a random non-reference allele to be used as a decoy mutation. For each trio separately, we extracted children reads overlapping each of the positions to insert the decoy mutation. Since *de novo* mutations are always heterozygous, each read had a 50% probability to be selected to carry the mutation. For all reads selected to carry the mutation, we replaced the reference base with the decoy mutation base. Base and mapping qualities were kept intact under the assumption that altering a single base in 90 would not affect these significantly. We then applied our entire *de novo* mutation calling pipeline to each decoy mutation.

Using these simulations, our *de* novo mutation calling pipeline had an average sensitivity of 67.9. This was heavily influenced by the coverage across the entire trio (R = 0.87). One outlier sample showed abnormally low sensitivity (-5.8sd) but was kept in the study since there were no quality concerns based on earlier QC [13].

Based on these simulations, we estimated the power to call a *de novo* mutation as a function of coverage in each individual in the trio. We found that simulated mutations covered by at least 9 reads in each parents, 4 reads in the child and 30 reads across the entire trio were detected with 92.5% sensitivity. On average 68.8% of the genome was covered in each trio using these thresholds. We considered all bases covered by these thresholds as high confidence bases in our analyses.

To derive a null distribution for *de novo* mutations based on the simulations above, we randomly sampled a single child at each of the 264k sites at which we inserted decoy mutations. The sampling was done regardless of whether the simulated mutation was called in the child or not and lead to a total of 179,845 called mutations that we used to compare our *de novo* mutations against.

Paternal age influence on the genomic location of mutations

We annotated each *de novo* mutation with replication timing measured in lymphoblastoid cell lines (LCL) [38] , expression levels in LCL [39] , recombination rates [31] and DNAse1 hypersensitivity sites and histone marks (H3K27ac, H3K4me1, H3K4me3) measured in lymphocytes (GM12878) from the ENCODE project [40] . We then used a linear regression model to investigate possible relationships between paternal age and the localization of *de novo* mutations with respect to the epigenetic variables above (DNA replication timing, recombination rate, DNase I hypersensitivity, expression levels, and the histone marks H3K27ac, H3K4Me1 and H3K4Me3), while correcting for GC content, CpG sites and sequencing coverage. We used a stepwise AIC approach, starting with a saturated model including all variables and their interactions, to derive a parsimonious model. The resulting parsimonious model only contained DNA replication timing ($p = 0.0022$) and histone H3K4me3 levels ($p = 0.35$) due a weakly significant interaction between the two ($p = 0.035$). This interaction is possibly caused by the correlation between replication timing and histone H3K4me3 levels. We estimated the significance of the other epigenetic variables by adding each one by one into the model and comparing the resulting model against the parsimonious model using an ANOVA test (**Supplementary Fig. 1**).

We dichotomized our data based on the age of the father to contrast replication-timing profiles of younger and older fathers. We ran an exhaustive search for a threshold that maximizes the difference between the two groups. For each of the 23 possible age thresholds, we used a Kolmogorov-Smirnoff test to compare the replication timing profile of the younger and older fathers (**Supplementary Fig. 2**). We found a peak around 28 years of age ($p = 5.7 \times 10^{-4}$) and therefore used this as an age threshold. Hereafter, we will refer to fathers who were <28 years old at conception as "younger fathers", and fathers who were ≥28 years old at conception as "older fathers".

We compared the distribution of replication timing of mutations from younger and older fathers using a Mann-Whitney (MW) test and found that those of younger fathers were significantly shifted towards later replicating regions ($p = 1.3 \times 10^{-4}$). We also compared the distribution of the replication timing of simulated

mutations against offspring of younger and older fathers and found these to be shifted towards later replication regions ($p = 4.9 \times 10^{-4}$) and similar ($p = 0.68$), respectively.

We repeated the same analyses using independent replication timing data [41] in four cell types (lymphoblastoid cells, neural precursor cells, embryonic stem cells (of four separate lines) and induced pluripotent stem cells (of two separate lines)) and observed consistent results across all cell types (**Supplementary Fig. 3**).

To delineate whether the effect we observed was paternal, maternal or both, we used mutations for which we could unambiguously determine parental origin and ran the linear regression model using the father's age on the 1,991 paternally inherited mutations ($\beta = 0.0092$, $p = 0.038$) and using the mother's age on the 630 maternally inherited mutations ($\beta = -0.0096$, $p = 0.26$) separately. Because of the difference in sample size between the mutation sets, we resampled 10,000 sets of 630 mutations from the paternally inherited mutations and ran the linear regression on these sets. We found that the expected paternal effect was significantly greater than the maternal one with the same number of mutations ($p = 0.0023$, **Supplementary Fig. 4**).

Next, we ran a robust linear regression model between the percentage of genic mutations in each of the 258 offspring and paternal age correcting for coverage and found a significant association ($\beta = 0.0026$, $p = 0.0085$). We used a robust linear regression model is to account for a single sample that showed an abnormally high percentage of genic mutations (>8sd away from the mean). This sample was no different from others in terms quality metrics such as coverage, SNP heterozygosity, proportion of known and novel SNPs and possible contamination.

We used linear regression models to compute the increase of mutations with paternal age for genic ($\beta = 0.52$, $p < 2 \times 10^{-16}$) and intergenic ($\beta = 0.32$, $p = 3.7 \times 10^{-14}$) mutations separately while correcting for coverage. Based on these, we estimated that an offspring born to a father aged 20 would receive on average 9.63 genic and 22.68 intergenic mutations whereas an offspring with a father aged 40 would receive on average 19.06 genic and 35.24 intergenic.

Mutations clusters

We tested whether the intra- and inter-individual distribution of *de novo* mutations deviated from a simulated uniform distribution across the genome correcting for detection power by assigning a probability equal to the average number of high confidence bases (see Simulations) across all trios at the kilobase

scale. We used a Kolmogorov-Smirnoff test and found that both intra-individual ($p = 3.3 \times 10^{-4}$) and inter-individual $p = 5.8 \times 10^{-5}$) were enriched in more closely spaced mutations than expected (**Fig. 3**). The strongest enrichment was for intra-individual mutations up to ~20kbp and we therefore defined mutation clusters as regions of 20kbp or less containing two or more *de* novo mutations in the same sample. We observed 73 clusters of two and 5 clusters of three mutations. We ran 1mln permutations to test whether these clusters were due to generally hyper-mutated regions. In each permutation round, we permutated the samples to which each mutation belongs to and counted the number of clusters obtained. The maximum we found under this permutation scheme was 18 such clusters, far from the 78 we observe in total, indicating that clustered mutations are likely co-occurring rather than independent. We then looked at the substitution types for clustered *de novo* mutations and compared them against non-clustered *de novo* mutations using chi-square tests. We also looked for differences in larger context (multiple flanking nucleotides) but did not see any further signature.

Mutation rates in exonic regions

We annotated all observed and simulated *de novo* mutation with their coding status (exonic, intronic, intergenic) using UCSC CCDS track [42] . We used a chi-square test to investigate differences in the number of observed and simulated mutations between exonic and non-exonic and found a 28% enrichment of mutations in observed exonic regions ($p = 0.008$). When considering non-CpG sites only, there was no significant difference ($p = 1.0$). Using a bootstrapping approach (randomly removing mutations regardless of their coding status), we computed that we would have 83.5% power to detect the enrichment above when removing the exonic mutations if it was present.

Mutation rates in DNase1 hypersensitivity sites

Using the ENCODE [40] measurements of DNase1 hypersensitivity sites (DHS), we defined a set of conserved peaks present in at least 2 cell types and annotated all observed and simulated mutations as within or outside a DHS peak (DHSstatus is 0 if outside a DHS peak, 1 if within). We used a logistic regression using a dummy variable DNMstatus (0 for simulated mutations and 1 for observed mutations) as the response variable and distance to DHS as the explanatory variables. Under this model, DHSstatus was significantly associated with DNMstatus ($\beta = 0.11$, $p = 0.0041$). When adding a CpG covariate (0 for mutations outside CpGs, 1 for those at CpG sites) into the model, CpG was strongly associated with DNMstatus ($\beta = 2.74$, $p < 2 \times 10^{-16}$) and the distance to DHS was no longer significant ($\beta = 0.038$, $p = 0.34$).

Influence of mutation and recombination rates on nucleotide diversity

We annotated all observed and simulated mutations with recombination rates from the DECODE reombination map [31] . We compared the distribution of recombination rates at mutation sites in observed and simulated mutation using a logistic regression model with a dummy variable DNMstatus (0 for observed, 1 for simulated mutation) as the response variable and recombination rate, correcting for CpGs and GC content (1kb up/downstream). We found a significant positive association between recombination rates and DNMstatus ($\beta = 0.01$, $p = 0.0015$).

We then computed nucleotide diversity (π) in regions of 10kb around each observed and simulated mutation using VCFTools [43] . We ran a linear regression model with π as the response variable and DNMstatus and recombination rate as explanatory variables. We found that under this model both DNMstatus ($\beta = 3.1 \times 10^{-4}$, $p < 2 \times 10^{-16}$) and recombination rates ($\beta = 7.64 \times 10^{-6}$, $p < 2 \times 10^{-16}$) were independently associated with π. We repeated the analysis with π computed in 100kb regions around each mutation and found similar results. Correcting for local GC content (computed over the same region as π) and CpGs did not influence these associations either.

Influence of mutation and recombination rates on human-chimpanzee divergence

To estimate the influence of mutation rates on human-chimpanzee divergence, we studied the correlations between a human-chimpanzee comparative genomics (HCCG) model (described below) and mutation rates obtained using: (a) observed mutation rates based on 11,020 *de novo* mutations, (b) rates based on sampling 11,020 mutations based on the HCCG model, and (c) rates based on sampling 11,020 mutations based on a "null" context-dependent mutation rate model. All rate computations were done on 1Mb non-overlapping regions. Because our power to call *de novo* mutation varies along the genome, each of the regions was corrected for the average fraction of high confidence bases per trio in that region (based on simulations).

The HCCG substitution model was computed using genome-wide triple alignments using the Pecan 10 amniotes multiple alignments available at the Ensembl database version 56 [44,45] (restricted to human, chimpanzee, and macaque) and excluding exons and CpG islands. We inferred substitution rates $r_{t,i}$ for each substitution type *t* in {A→G , A→C, A→T, C→A, C→G, C→T ,CpG→TpG (to account for hyper-mutability of CpG sites)} in each region *i* using a maximum likelihood-based method as described elsewhere [30] . We assumed that the rates

of complementary substitution processes to be equal, but did not assume that the substitution process is time-reversible.

We next computed the total substitution rate per window using the rate inferred above for all bases *b* in {A, C} using the following formula:

$$n_{b,i} \sum_{t_b} r_{t_b,i} + 2n_{CpG,i} r_{CpG \to \mathrm{TpG},i}$$

where $n_{b,i}$ is the number of high confidence bases *b* in window *i*, and t_b is the set of substitutions in *t* where the ancestral base is *b* (e.g. for *b* = A, t_b= { A→C, A→G, A→T}).

The genome-wide averaged model was computed assuming a uniform substitution rate matrix, defined as the mean of the substitution rate matrices over all 1Mb regions. We then applied the same procedure as above to obtain a uniform rate but context-dependent substitution model.

Using the above HCCG substitution-rate model and the genome-wide averaged model, we drew 100,000 genomic profiles of 11,020 mutations (same as the number of observed *de novo* mutations) using the Poisson random number generator in R [46] . For each of these simulated substitution profiles, as well as the observed *de novo* mutation rates, we computed the correlation with the HCCG (**Fig. 4**).

To investigate the effect of local decode sex-averaged recombination rates on the HCCG model, we computed substitution rates s_i for each region *i* using the following Poisson regressing with both local male recombination rates ρ_i and observed mutation counts n_i:

$$\log (S_1) = \beta_0 + \beta_\rho \rho_i + \beta_n n_i$$

We then computed the Pearson's correlation between the HCCG and the above Poisson regression.

Strand asymmetry in transcribed regions

Using the UCSC CCDS track, we annotated all genic *de novo* mutations with the direction of transcription. We annotated each *de novo* mutation with its corresponding strand-dependent substitution type (A→G, G→A, A→C, A→T, C→A, C→G). We used chi-square tests to evaluate strand differences for each substitution type in transcribed regions (**Supplementary Fig. 6**).

We used a chi-square test to compare observed and simulated mutation counts in intronic and intergenic regions and found a modest 2.8% depletion in observed intronic regions ($p = 0.047$).

Tri-nucleotide context dependency

We utilized the context-dependent substitution matrix from fixed differences in human, chimp and baboon (i.e. from multiple alignments of the three species) available on the UCSC genome browser [42,47,48] to empirically calculate the "directed" 64x3 mutation matrix using SCONE [33] . We accounted for multiple mutational events and restricted our analysis to non-exonic regions and removed CpG islands. We then computed the same mutation matrix based on *de novo* mutations and computed Pearson's correlation coefficient between the two matrices ($r^2 = 0.993$).

Mutation rate map

Although our set of *de novo* mutations is the largest available to date, it is still a relatively sparse sampling across the genome. For this reason, we set on using a human-chimpanzee-macaque primate substitution model [30] and refine it using our observed mutations.

Local mutation rates derived from a human-chimpanzee-macaque primate substitution model [30] were corrected for biases due to local recombination rate [31] , types of mutations, and the direction of transcription along the strand. This correction was applied for 2,339 1Mb non-overlapping windows across the autosomes after excluding windows where (1) the decode sex-averaged recombination rate is unavailable for more than 10% of the window, (2) the sex-averaged recombination rate across the window is 0 or greater than 3 cM/Mb, (3) the primate local substitution rate was estimated to be 0 or extremely high, or (4) more than 800 kb on average were below our high confidence calling threshold per trio (based on simulations).

For each 1Mb window *i*, a substitution rate matrix was inferred using the context-dependent primate substitution model described in Duret *et al.* [30] for seven types of substitutions, parameterized by $r_{t,i}$ with *t* in {T→G, G→C, T→C, T→A, C→A, C→T, CpG→TpG (to account for hyper-mutability of CpG sites)}.

First, we tested if the observed *de novo* mutation rates co-vary with primate substitution rates across the genome using the following Poisson regression model with log link function:

$$\log(n_{t,i})$$

$$= \beta_{r,t} r_{t,i} + \beta_{r,t,0} + \log(N_{t,i}) \text{ for } t \neq \text{CpG}\rightarrow\text{TpG}$$

$$= \beta_{r,t}(r_{t,i} + r_{C\rightarrow T,i}) + \beta_{r,t,0} + \log(N_{t,i}) \text{ for } t = \text{CpG}\rightarrow\text{TpG}$$

where $n_{t,i}$ is the observed count of *de novo* mutations of type *t* in window *i*, $r_{t,i}$ is the substitution rate of type *t* in window *i*, and $N_{t,i}$ is the number of sites at which *de novo* mutations of type *t* can be detected with high confidence in window *i*. The offset term $\log(N_{t,i})$ was added since the number of called *de novo* mutations is dependent on detection power. The C→T mutation of CpG sites requires special treatment since it can be attributed to context-independent C→T substitution as well as hyper-mutability of CpG.

The primate substitution rates in the above Poisson regression model only had significant predictive power for local *de novo* mutation rates for S→W and W→W substitutions (Supplementary Table 1). For this reason, we only estimated local mutation rates based on the primate substitution rate for substitutions in t_{SW} = {T→A, C→A, C→T, CpG→TpG}. For the rest of substitutions (*t* in {T→G, T→C, G→C}), we used the genome-wide averaged mutation rates $r_t^{'}$ estimated from our observed mutations:

$$r_t^{'} = \frac{\sum_i n_{t,i}}{\sum_i N_{t,i}} \cdot \frac{1}{c}$$

$$c = \frac{\sum_i \sum_{t \epsilon t_{SW}} n_{t,i}}{\sum_i \sum_{t \epsilon t_{SW}} r_{t,i} N_{t,i} + \sum_i r_{C\rightarrow T,i} N_{CpG\rightarrow TpG,i}}$$

where *c* is a scaling factor to convert between *de novo* mutation rates and instantaneous substitution rates.

Second, we corrected for the biases due to local recombination rates. The observed local *de novo* mutation rates were not significantly correlated with recombination rates when considering each type of substitutions separately (Bonferroni-corrected p-value > 0.05). However, local substitution rates $r_{t,i}$ depend significantly on local sex-averaged recombination rates ρ_i for *t* = C→A and CpG→TpG (Supplementary Table 2). To eliminate the dependency on recombination rate, we fit the following linear regression model:

$$r_{t,i} = \beta_{\rho,t}\, \rho_i + \beta_{0,t}$$

and residualized $r_{t,i}$ by subtracting the ρ_i-dependent term.

The final formula we used to compute the mutation rates for each 1Mb window *i* is then:

$\mu_{t,i} = \left(r_{t,i} - \beta_{\rho,t}\,\rho_i\right) \cdot f_t$ for *t* in {C→A, CpG→TpG}

$\mu_{t,i} = r_{t,i} \cdot f_t$ for *t* in {T→A, C→T}

$\mu_{t,i} = r_t^{'}$ for *t* in {T→G, T→C, G→C}

where f_t is a global scaling factor for substitution of type *t* to match the observed frequencies of different types of *de novo* mutations (Supplementary Table 2). In particular, A→T mutation is over-represented in primate substitutions by 12% compared to our *de novo* data. For each *t* in $t_{SW,}$ f_t is defined to satisfy the following conditions:

$\sum_i \mu_{t,i} N_{t,i} = \frac{1}{c} \sum_i n_{t,i}$ for *t* in {T→A, C→T, C→A}

$\sum_i \left(\mu_{t,i} + \mu_{C\to T,i}\right) N_{t,i} = \frac{1}{c} \sum_i n_{t,i}$ for *t* = CpG→TpG

Finally, the mutation rate μ was scaled so that the overall mutation rate across the autosome is 1.2×10^{-8} per nucleotide per generation.

To evaluate the fit of the estimated mutation rates to observed *de novo* mutations, we examined the likelihood L_t of the observed data given mutation rates, assuming homogenous Poisson process for each type of mutation *t* within each window *i*:

$$L_t\left(data \middle| \mu_{t,i}\right)$$

$$= \prod_i Poisson\left(n_{t,i} \middle| \lambda = \frac{1}{c} \mu_{t,i} N_{t,i}\right) \text{ for } t \neq \text{CpG→TpG}$$

$$= \prod_i Poisson\left(n_{t,i} \middle| \lambda = \frac{1}{c'} \left(\mu_{t,i} + \mu_{C\to T,i}\right) N_{t,i}\right) \text{ for } t = \text{CpG→TpG}$$

$$c^{'} = \frac{\sum_i \sum_t n_{t,i}}{\sum_i \sum_t \mu_{t,i} N_{t,i} + \sum_i \mu_{C\to T,i} N_{CpG\to TpG,i}}$$

The likelihood of the observed data under different models is summarized in Supplementary Table 3.

We estimated functional mutation rates in protein-coding region for autosomal protein-coding transcripts (downloaded from Ensembl [45] v74). Excluding 24,508

transcripts (3,808 genes) outside our analysis windows for bias correction, we computed bias-corrected mutations rates for a total of 54,310 transcripts (15,462 genes). For maximum coverage of genes, however, we provide two additional functional mutation rates based on uncorrected local primate substitution rates $r_{t,i}$ and the uniform genome-wide averaged mutation rates $r_t^{'}$ derived from our observed data.

For each transcript, the local mutation rate was determined by the 1Mb genomic window that overlapped the coordinate of midpoint between transcription start and end sites. Based on this rate, all possible nonsense, missense, synonymous and 4-fold degenerate synonymous mutations were examined with respect to the reference genome, and their mutation rates were aggregated over the entire transcript.

While we assumed the equal rate of $\mu_{A\to G,i}$ and complementary $\mu_{T\to C,i}$ in non-coding region, we adjusted for their strand bias in protein-coding region as follows:

$$\mu_{A\to G,i}^{tx} = \frac{N_A^{tx} + N_T^{tx}}{N_A^{tx}} \frac{\gamma_{sb}}{1+\gamma_{sb}} \mu_{T\to C,i}^{nc}$$

$$\mu_{T\to C,i}^{tx} = \frac{N_A^{tx} + N_T^{tx}}{N_A^{tx}} \frac{1}{1+\gamma_{sb}} \mu_{T\to C,i}^{nc}$$

$$\gamma_{sb} = \frac{\mathrm{n}_{A\to G}^{tx}}{\mathrm{n}_{T\to C}^{tx}}$$

where $\mu_{T\to C,i}^{nc}$ (= $\mu_{A\to G,i}^{nc}$) is the local mutation rate of A:T→G:C in non-coding region, $\mu_{T\to C,i}^{tx}$ and $\mu_{A\to G,i}^{tx}$ are the local mutation rates of T→C and A→G in protein-coding with respect to the transcribed strand, N_A^{tx} and N_T^{tx} are the total numbers of protein-coding A and T bases in transcribed strand across the autosomes, and $\mathrm{n}_{T\to C}^{tx}$ and $\mathrm{n}_{A\to G}^{tx}$ are the genome-wide counts of observed T→C and A→G *de novo* mutations with respect to the transcribed strand in our dataset. γ_{sb} was estimated to be 1.389 and $N_A^{tx}/(N_A^{tx} + N_T^{tx})$ to be 0.543 in our data.

Acknowledgements

Sequence data have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI), under accession number EGAS00001000644. The mutation rate map can be found on the Genome of the Netherlands website at http://www.nlgenome.nl. The GoNL Project is funded by the Biobanking and Biomolecular Research Infrastructure (BBMRI-NL), which is financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007). S.S, P.P. and S.C. are funded by NIH grants 1 R01 MH101244 and 1 R01 GM078598. We thank Dmirty Gordenin for very helpful comments.

References

1. Sawyer, S. A. & Hartl, D. L. Population genetics of polymorphism and divergence. *Genetics* **132,** 1161–76 (1992).
2. Felsenstein, J. & Churchill, G. A. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13,** 93–104 (1996).
3. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–50 (2005).
4. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–13 (2005).
5. Veltman, J. & Brunner, H. De novo mutations in human genetic disease. *Nature reviews. Genetics* **13,** 565–75 (2012).
6. Friedberg, E. C., Walker, G. C. & Siede, W. DNA repair and mutagenesis. (1995).
7. Kondrashov, A. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases . *Human Mutation* **21,** 12–27 (2003).
8. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 961–8 (2010).
9. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12,** 756–66 (2011).
10. Schaibley, V. M. *et al.* The influence of genomic context on mutation patterns in the human genome inferred from rare variants. *Genome Res.* **23,** 1974–84 (2013).
11. Michaelson, J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151,** 1431–42 (2012).
12. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–5 (2012).
13. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–25 (2014).
14. Jenkins, T. G., Aston, K. I., Pflueger, C., Cairns, B. R. & Carrell, D. T. Age-associated sperm DNA methylation alterations: possible implications in offspring disease susceptibility. *PLoS Genet.* **10,** e1004458 (2014).
15. Koren, A. DNA replication timing: Coordinating genome stability with genome regulation on the X chromosome and beyond. *Bioessays* **36,** 997–1004 (2014).
16. Arndt, P. F., Petrov, D. A. & Hwa, T. Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. *Mol. Biol. Evol.* **20,** 1887–96 (2003).
17. Schmidt, S. *et al.* Hypermutable non-synonymous sites are under stronger negative selection. *PLoS Genet.* **4,** e1000281 (2008).
18. Subramanian, S. & Kumar, S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res.* **13,** 838–44 (2003).

19. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* **32,** 71–5 (2014).

20. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463,** 184–90 (2010).

21. Campbell, C. D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44,** 1277–81 (2012).

22. Roberts, S. A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46,** 424–35 (2012).

23. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149,** 979–93 (2012).

24. Chan, K., Resnick, M. A. & Gordenin, D. A. The choice of nucleotide inserted opposite abasic sites formed within chromosomal DNA reveals the polymerase activities participating in translesion DNA synthesis. *DNA Repair (Amst.)* **12,** 878–89 (2013).

25. Arndt, P. F., Hwa, T. & Petrov, D. A. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *J. Mol. Evol.* **60,** 748–63 (2005).

26. Hellmann, I., Ebersberger, I., Ptak, S. E., Pääbo, S. & Przeworski, M. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72,** 1527–35 (2003).

27. Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* **356,** 519–20 (1992).

28. Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18,** 337–40 (2002).

29. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genetics* (2002). doi:10.1038/ng917

30. Duret, L. & Arndt, P. F. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* **4,** e1000071 (2008).

31. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467,** 1099–103 (2010).

32. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5,** e1000471 (2009).

33. Asthana, S., Roytberg, M., Stamatoyannopoulos, J. & Sunyaev, S. Analysis of sequence conservation at nucleotide resolution. *PLoS Comput. Biol.* **3,** e254 (2007).

34. Gratten, J., Visscher, P. M., Mowry, B. J. & Wray, N. R. Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. *Nat. Genet.* **45,** 234–8 (2013).

35. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

36. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297–303 (2010).

37. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–95 (2010).

38. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *American journal of human genetics* **91,** 1033–40 (2012).

39. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–11 (2013).

40. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

41. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research* **20,** 761–70 (2010).

42. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

43. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics (Oxford, England)* **27,** 2156–8 (2011).

44. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18,** 1814–28 (2008).

45. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41,** D48–55 (2013).

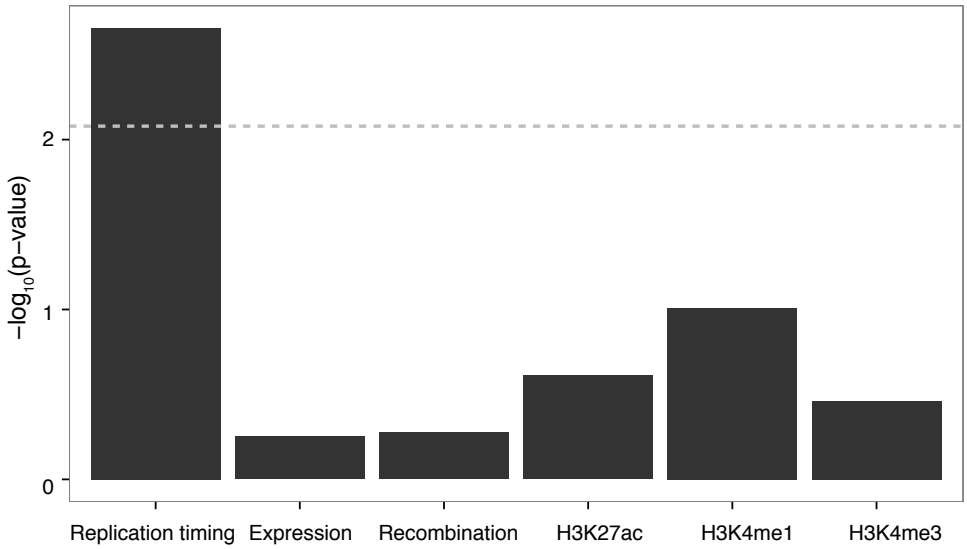
46. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2014).

47. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–15 (2004).

48. Murphy, W. J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294,** 2348–51 (2001).

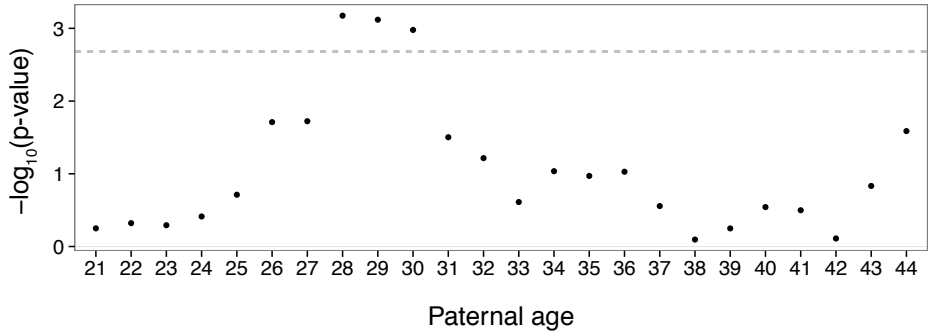
Chapter 4

Supplementary information



Supplementary Figure 1 | Association of paternal age *de novo* mutation location with respect to epigenetic variables

Using a linear regression model, we tested the association of 7 epigenetic variables (DNA replication timing, expression levels, recombination rates and histone modifications H3K27ac, H3K4Me1 and H3K4me3), while correcting for GC-content, CpG status and sequence coverage. Here, we plot the significance of the associations we found with the different epigenetic variables along with the significance threshold level after Bonferroni correction for the 6 tests we performed (grey dashed line).



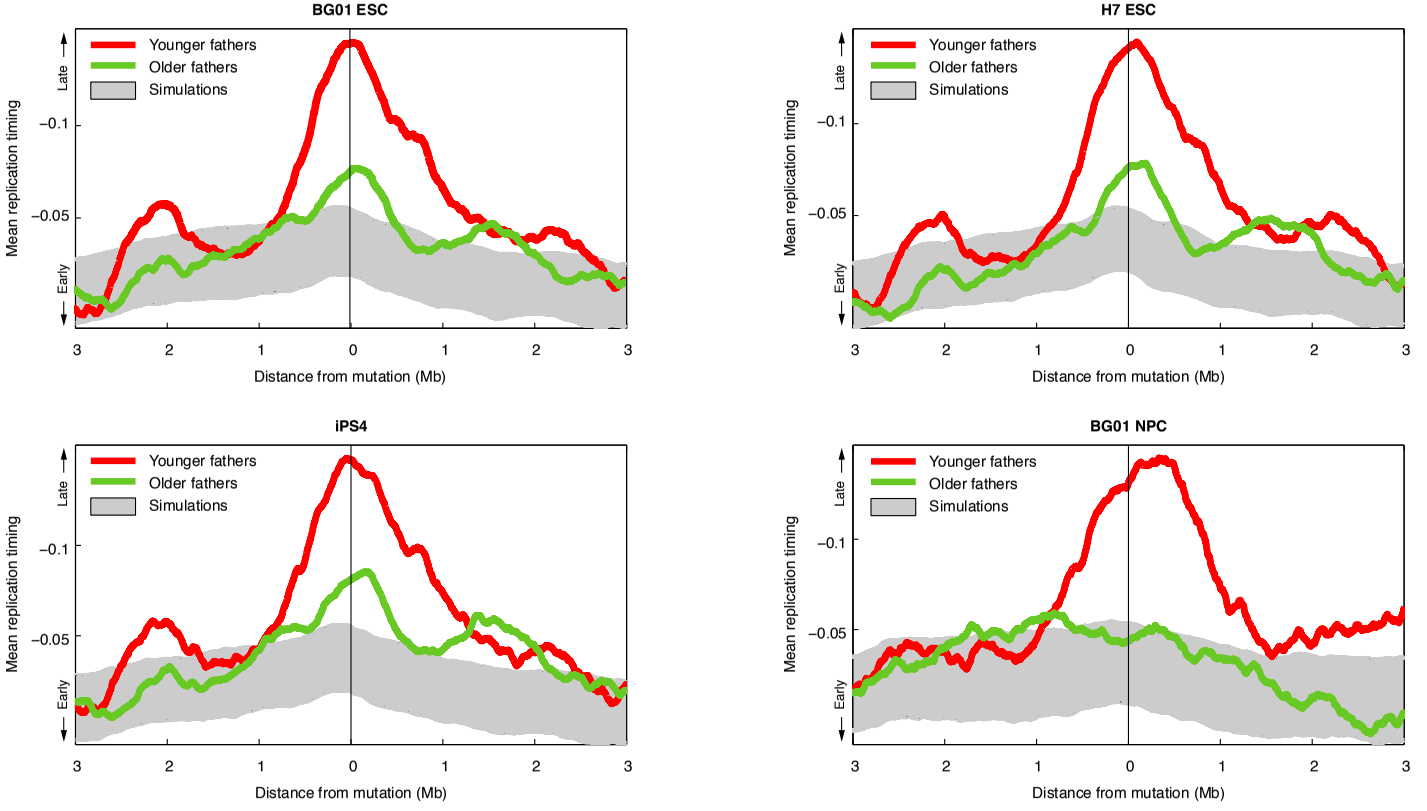
Supplementary Figure 2 | Separation of younger and older fathers

We separated our *de novo* mutation data in two groups based on paternal age. To select a threshold that maximizes the difference in the groups, we considered every integer age in our study as a possible threshold and applied a Kolmogorov-Smirnoff test to compare the distribution of replication timing of *de novo* mutations between the groups. This plot shows the p-values obtained for each of the 27 tests as well as the significance threshold afer Bonferroni correction (grey dashed line).

Cell type	Cell line	Replicate	Linear regression		M-W younger vs older fathers		M-W younger fathers vs simulations		M-W older fathers vs simulations	
			p	β	p	β	p	β	p	β
Lymphoblastoid cells	6 lines	-	**0.0022**	0.158	**1.3E-04**	-0.070	**4.9E-04**	-0.053	0.6839	0.004
	C0202	1	**0.0305**	0.110	**1.7E-03**	-0.042	**2.7E-03**	-0.033	0.83	0.002
	C0202	2	**0.0285**	0.111	**9.9E-04**	-0.050	**2.8E-04**	-0.046	0.70	-0.004
Neural precursor cell	BG01	1	**0.0265**	0.108	**1.5E-03**	-0.057	**9.5E-04**	-0.049	0.84	-0.002
	BG01	2	**0.0257**	0.111	**7.5E-04**	-0.051	**7.6E-04**	-0.042	0.93	-0.001
Embryonic stem cell	BG01	-	**7.7E-04**	0.171	**3.6E-04**	-0.057	**5.0E-05**	-0.055	0.37	-0.009
	BG02	1	**0.0036**	0.148	**5.2E-04**	-0.048	**4.4E-05**	-0.047	0.27	-0.009
	BG02	2	**0.0047**	0.146	**4.7E-03**	-0.044	**1.1E-03**	-0.042	0.35	-0.009
	H7	-	**0.0036**	0.146	**3.2E-03**	-0.051	**1.5E-04**	-0.055	0.16	-0.015
	H9	-	**0.0021**	0.155	**7.1E-04**	-0.050	**1.3E-05**	-0.054	0.15	-0.013
Induced pluripotent stem cell	iPS4	1	**6.1E-04**	0.172	**1.1E-03**	-0.054	**5.6E-05**	-0.056	0.31	-0.011
	iPS4	2	**0.0046**	0.143	**3.0E-04**	-0.065	**8.4E-05**	-0.059	0.68	-0.004
	iPS5	1	**8.9E-04**	0.167	**3.6E-04**	-0.060	**3.2E-04**	-0.051	0.93	-0.001
	iPS5	2	**8.7E-04**	0.171	**5.5E-03**	-0.045	**2.7E-03**	-0.040	0.65	-0.004

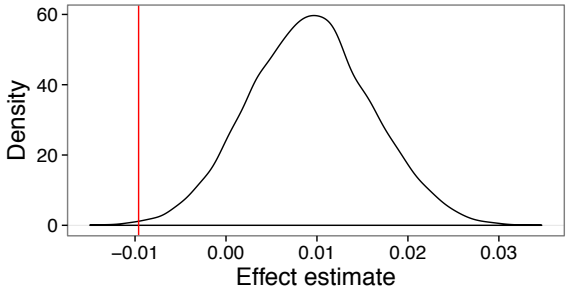
Supplementary Figure 3 | Paternal age effect on *de novo* mutation replication timing measured in 6 cell types

a. The source of the replication timing data was Koren *et* al.[45] for the 6 lymphoblastoid cell lines (1st row) and Ryba *et al.*[47] for all other cell lines. The linear regression column contains p-values and estimates (β) for the parsimonious model described in Methods. The M-W test columns contain the p-values and estimated difference using a Mann-Whitney test between the distribution of mutation replication timing values of: offspring of younger (<28 years old) vs older (≥28 years old); offspring of younger fathers vs simulations; offspring of older fathers vs simulations. Significant p-values are highlighted in bold font.



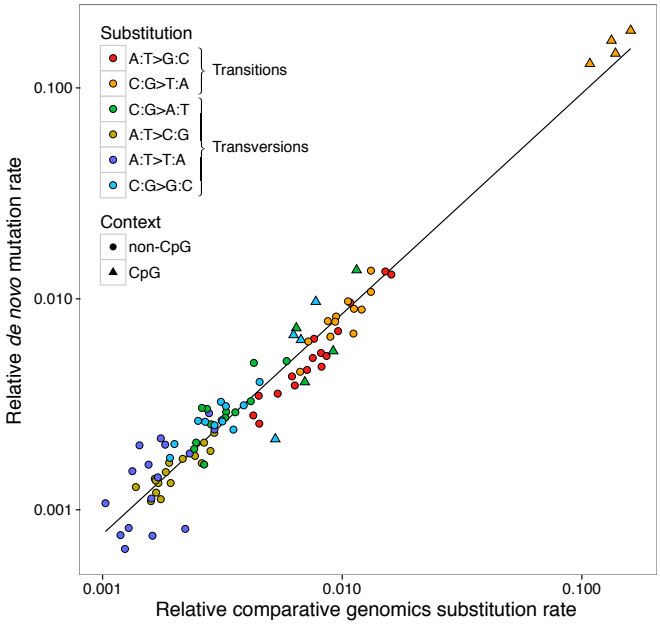
Supplementary Figure 3 | Paternal age effect on *de novo* mutation replication timing measured in 6 cell types

b. The distribution of replication timing around *de novo* mutations in offspring of younger fathers (red curves, < 28 years old), older fathers (green curves, ≥ 28 years old) and simulations (grey areas, 200 simulation sets of 11,020 mutations) in three cell types from four cell lines[47]: embryonic stem cells (BG01 ESC, H7 ESC), induced pluripotent stem cells (iPS4) and neural precursor cells (BG01 NPC).



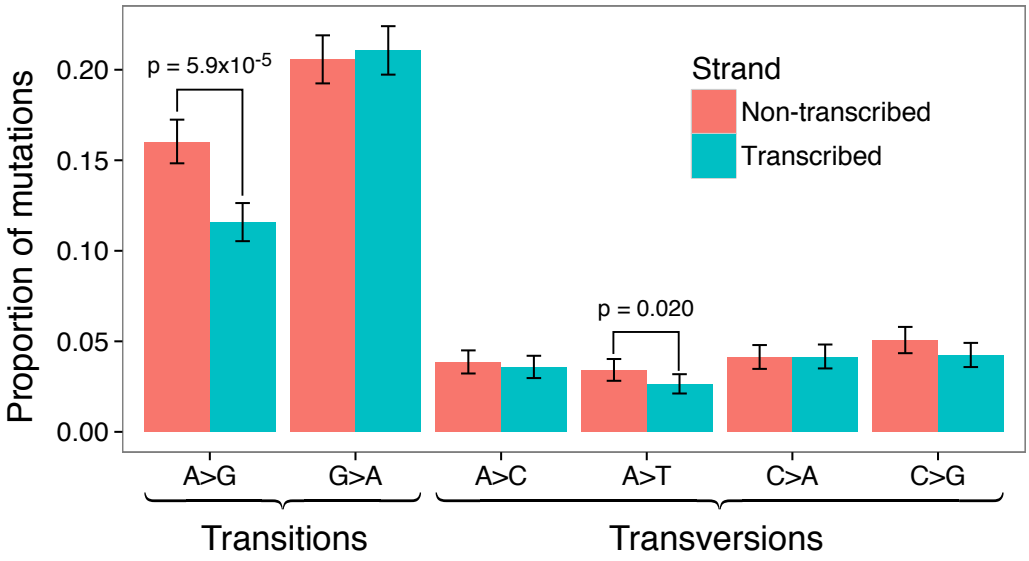
Supplementary Figure 4 | Paternal and maternal age effect estimate on replication timing based on *de novo* mutations with known parental origin.

All effect estimates are computed using a linear regression model. The red line represents the effect estimate from the 630 maternal mutations. The black curve shows the effect estimate for 10,000 sampling of 630 paternal mutations out of the 1,991 available. The effect estimate is significantly larger for paternal mutations (p = 0.0019) when considering the same number of mutations.



Supplementary Figure 5 | Comparison of nucleotide context-specific mutation rates based on comparative genomics and observed *de novo* mutations

Each point represents one of the 96 substitutions in a specific tri-nucleotide context. The black line shows the best fit (r^2 = 0.993). The rate of transition mutations is 2.15x greater than the transversion rate (Ti/Tv ratio). The highest mutation rates are observed for cytosine bases in a CpG context.



Supplementary Figure 6 | Mutational spectrum in transcribed regions.

The proportion of *de novo* mutations of each substitution type in transcribed regions classified based on their corresponding strand (transcribed or non-transcribed). There is a strong asymmetry of mutations between the two strands, with significantly elevated A>G substitutions on the transcribed strand, consistent with the action of transcription-coupled nucleotide excision repair.

	$\hat{\beta}_{\sigma,t}$	$P(\hat{\beta}_{\sigma,t} \neq 0)$	$\hat{f}_t$
G→A	-2.485×10^{-5}	0.23	1.034
T→A	2.597×10^{-6}	0.649	0.891
C→A	-9.857×10^{-5}	$< 2 \times 10^{-16}$	1.014
CpG→CpA/TpG	-0.00449	$< 2 \times 10^{-16}$	0.988

Supplementary Table 1 | Estimated values of correction factors for the mutation rate map

This table summarizes the estimated values of correction factors between comparative genomics and *de novo* rates for the different substitution types.

Log Likelihood	GoNL uniform	Primate $r_{t,i}$	Corrected $\mu_{t,i}$(%/%)*		
Recomb. rates	No	No	Male	Sex-Averaged	Female
log $L_{T>G}$	-1,591.0	-1,605.3	-1,584.6 (0.4% / 1.3%)	-1,584.6 (0.4% / 1.3%)	-1,584.6 (0.4% / 1.3%)
log $L_{G>A}$	-2,974.5	-2,961.7	-2,959.2 (0.5% / 0.1%)	-2,959.2 (0.5% / 0.1%)	-2,957.7 (0.6% / 0.1%)
log $L_{G>C}$	-1,735.6	-1,736.5	-1,728.3 (0.4% / 0.5%)	-1,728.4 (0.4% / 0.5%)	-1,728.4 (0.4% / 0.5%)
log $L_{T>C}$	-3,176.2	-3,183.5	-3,152.0 (0.8% / 1.0%)	-3,152.0 (0.8% / 1.0%)	-3,152.0 (0.8% / 1.0%)
log $L_{T>A}$	-1,439.8	-1,437.7	-1,434.5 (0.4% / 0.2%)	-1,434.5 (0.4% / 0.2%)	-1,434.5 (0.4% / 0.2%)
log $L_{C>A}$	-1,852.2	-1,836.7	-1,832.5 (1.1% / 0.2%)	-1,831.9 (1.1% / 0.3%)	-1,831.7 (1.1% / 0.3%)
log $L_{CpG>CpA/TpG}$	-2,696.2	-2,439.6	-2,435.8 (9.7% / 0.2%)	-2,435.0 (9.7% / 0.2%)	-2,434.9 (9.7% / 0.2%)
Total	**-15,465.6**	**-15,201.0**	**-15,126.9 (2.2% / 0.5%)**	**-15,125.6 (2.2% / 0.5%)**	**-15,123.8 (2.2% / 0.5%)**

* In parenthesis is the percent change of log likelihood of $\mu_{t,i}$ compared to GoNL uniform model and uncorrected primate rate model $r_{t,i}$.

Supplementary Table 2 | Likelihood of the observed data under different mutation rate models

Likelihood of the observed *de* novo mutation data by substitution type based on (a) a uniform mutation rate model derived from the observed mutations, (b) the uncorrected primate rate matrix $r_{t,i}$ and (c) the computed mutation rate matrix $\mu_{t,i}$ (Methods).

Chapter 5

Characteristics of *de novo* structural changes in the human genome

Wigard P. Kloosterman*, Laurent C. Francioli*, Fereydoun Hormozdiari, Tobias Marschall, Jayne Y. Hehir-Kwa, Abdel Abdellaoui, Eric-Wubbo Lameijer, Matthijs H. Moed, Vyacheslav Koval, Ivo Renkens, Markus J. van Roosmalen, Pascal Arp, Lennart C. Karssen, Bradley P. Coe, Robert E. Handsaker, Eka D. Suchiman, Edwin Cuppen, Djie T. Thung, Mitch McVey, Michael C. Wendl, Genome of the Netherlands Consortium, Andre Uitterlinden, Cornelia M. van Duijn, Morris Swertz, Cisca Wijmenga, Gertjan van Ommen, P. Eline Slagboom, Dorret I. Boomsma, Alexander Schönhuth, Evan E. Eichler, Paul I. W. de Bakker, Kai Ye* and Victor Guryev*, *Revised manuscript under review at Genome Research*

Abstract

Small insertions and deletions (indels) and large structural variations (SVs) are major contributors to human genetic diversity and disease. However, mutation rates and characteristics of *de novo* indels and SVs in the general population have remained largely unexplored. We report 332 validated *de novo* structural changes identified in whole genomes of 250 families, including complex indels, retrotransposon insertions and interchromosomal events. These data indicate a mutation rate of 2.93 indels (1-20bp) and 0.16 SVs (>20bp) per generation. Structural changes affect on average 4.1kbp of genomic sequence and 29 coding bases per generation, which is 91 and 52 times more nucleotides than *de novo* substitutions, respectively. This contrasts with the equal genomic footprint of inherited SVs and substitutions. An excess of structural changes originated on paternal haplotypes. Additionally, we observed a non-uniform distribution of *de novo* SVs across offspring. These results reveal the importance of different mutational mechanisms to changes in human genome structure across generations.

Introduction

Genomic mutations drive human evolution and phenotypic diversity. They are generally divided into three classes: single nucleotide variations (SNVs), small insertions and deletions (indels) and larger structural variants (SVs). Comparative genomics studies highlighted important small base-level and large-scale differences between human and chimpanzee genomes and noted a larger impact of segmental duplications compared to SNVs [1]. Whereas interspecies comparisons provide us with insight into long-range processes such as genetic drift and selection, the information derived from direct measurements of the *de novo* mutation spectrum and rates across generations is crucial for understanding mechanisms of mutation formation and inter-individual differences [2]. While several projects have started to investigate the rates and characteristics of *de novo* SNVs [3–6], those of *de novo* indels and large SVs have been much less studied [7].

Copy number variations (CNVs) and SVs contribute substantially to human genetic variation [8–11] and the phenotypic impact of CNVs may be larger than of SNVs [12–14]. The impact of novel changes in genome structure is further illustrated by their role in human genetic disease [15,16]. Copy number variations (CNVs) are widely studied and have been implicated in a variety of neurological disorders, such as autism [17], schizophrenia [18] and intellectual disability [15]. *De novo* CNVs are highly enriched in these patients as compared to healthy controls and contribute to disease in more than 10% of cases [19,20]. Recent large-scale exome sequencing studies have

uncovered *de novo* SNVs and short indels causing various disease phenotypes, ranging from complex neurological disease to rare Mendelian disorders [21].

Given the significant contribution of *de novo* mutations to human disease and evolution, studying genome-wide mutation rates and patterns is important for understanding mutation origins, locating hotspots, estimating disease risk and interpreting novel disease-associated mutations. Here, we surveyed the entire spectrum of *de novo* indels (1-20bp) and SVs (>20bp) in the human population at nucleotide-resolution using whole genome sequencing data of 250 families from the Genome of the Netherlands (GoNL) project [4,22].

Results

Study design and variant detection

The Genome of the Netherlands project includes 231 parent-offspring trios, 11 quartets with monozygotic twins and 8 quartets with dizygotic twins for a total of 258 genetically distinct children. DNA material was obtained from peripheral blood mononuclear cells to avoid problems with accumulated somatic mutations routinely observed in DNA isolated from cell lines [23]. The medium coverage (14.5x median sequence depth; 38.4x median physical depth) of paired-end sequencing data combined with a family-based design enabled the construction of a high-quality dataset of genomic variation [4].

Indels (1-20bp) were called using GATK UnifiedGenotyper [24], GATK HaplotypeCaller (http://www.broadinstitute.org/gatk/), and Pindel [25], and then filtered according to recommendations for each tool (**Fig. 1**). We focused exclusively on variants that were detected only in a single child by at least one algorithm with high confidence (**Supplemental Table 1**). We performed experimental validation assays for all 1,176 candidate *de novo* indels in 110 children from 92 families (11 quartets with monozygotic twin pairs, 7 quartets with dizygotic twin pairs, 74 trios). We successfully re-sequenced 917 candidates in these families, of which 291 indels (203 deletions, 74 insertions and 14 complex indels) were confirmed as *de novo* events. All 31 *de novo* mutations validated in monozygotic twin pairs were concordant between the two twins, showing that most of the mutations we report are germline mutations. After validation, we randomly excluded one of the twins from each monozygotic twin pairs, leaving 99 children for *de novo* indel analysis. We only focused on regions where we had sufficient indel calling power by requiring at least 4 reads in the child and 10 reads in each parents. Using these thresholds a median of 77% of the genome was covered with sensitivity of 93.2% based on comparison of singletons in

11 twin pairs and 83.3% based on comparison of singletons in whole-exomes of 24 parents sequenced independently at deep coverage. Based on these experiments, we found a lower sensitivity for insertions (92.6% based on twin comparison, 75.1% based on whole-exome comparison) than for deletions (93.5% based on twin comparison, 87.4% based whole-exome comparison).

Structural variants (>20bp) were predicted by a selection of 11 tools (1-2-3-SV, Breakdancer [26], CNVnator [27], DWAC-Seq, FACADE [28], GATK Unified Genotyper [24], GATK HaplotypeCaller, GenomeSTRiP [29], MATE-CLEVER [30], Mobster [31], Pindel [25]) that together use information from gapped reads, split-reads, discordant read-pairs and read depth to capture the full spectrum of SV sizes and types (**Fig. 1**). In order to maximize sensitivity, each tool was run using permissive settings and all *de novo* calls were visually inspected using the Integrated Genome Viewer (IGV) [32]. We identified a total of 601 *de novo* SV candidates in the 258 GoNL offspring (**Supplemental Table 1**). All candidates were subjected to experimental validation, resulting in a final set of 41 confirmed *de novo* SVs ranging in size from 20bp to 327kbp (**Supplemental Figs. 1, 2**). This set includes 27 deletions, 8 tandem duplications, 5 retrotransposon insertions and 1 complex interchromosomal event (that also involves a retrotransposon segment). We estimate the sensitivity of our calling for SVs sized 20-99bp and SVs larger than 100bp to be 69.4% and 85.8% that of deep coverage data, respectively. Further, nearly the complete genome, (an average of 98.8% of the haploid genome excluding assembly gaps) was covered by four or more read-pairs, a minimum threshold for calling SVs in our data (**Methods**). The sensitivity for detection of retrotransposon insertions was tested based on a previously published set of validated variants and found to be 77.6% for heterozygous retrotransposon insertions [33]. To empirically estimate the sensitivity for calling large SVs (>100 kb), we analyzed Illumina high-density SNP array data that were generated for 57 families (**Supplemental Table 2**). We detected a single *de novo* deletion (113 kb) in these data, which was already identified by whole genome sequencing.

In total, we confirmed 332 *de novo* structural changes (291 indels of size 1-20bp and 41 SVs larger than 20bp), which were used for downstream analyses (**Fig. 2A, Supplemental Table 2**). All 332 *de novo* variations are uniquely present in a single individual in the GoNL cohort. We also examined the overlap with public databases and found that 3 large SVs (>80% reciprocal overlap; Database of Genomic Variants; 1000 Genomes Phase 1) and 8 rare indels (exact match; dbSNP build 142; allele frequency < 1.5%) are overlapping, suggesting that these events are recurring in the population.

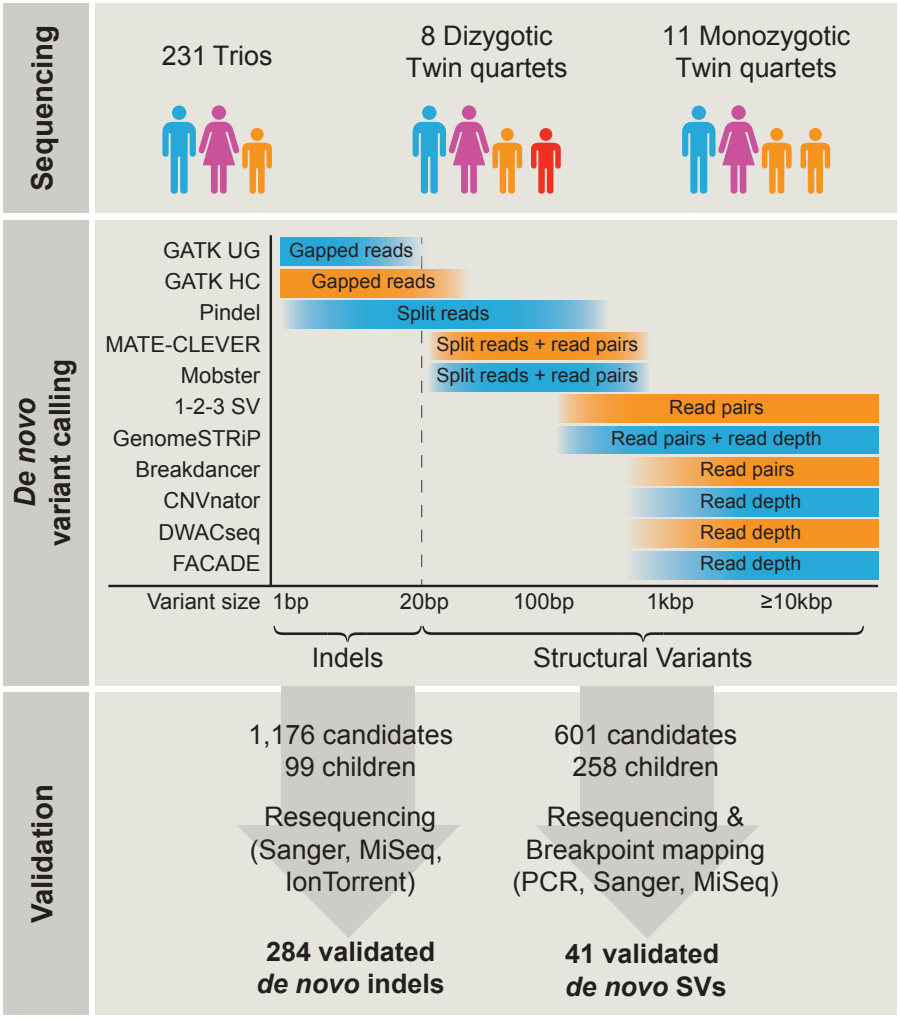


Figure 1 | Overview of study design

A total of 250 parent-offspring families were sequenced at 13x coverage. *De novo* indel and structural variant (SV) calling was performed using 11 algorithms combining gapped reads, split reads, discordant read-pairs and read depth approaches to cover the entire mutation size spectrum. All candidate indels (1,176 in 99 children) and SVs (601 in 258 children) were subjected to experimental validation leading to 284 validated de novo indels and 41 validated *de novo* SVs.

Indel and SV mutation rates

Previous estimates of the human indel mutation rate range from 0.53 to 1.5 x 10^{-9} per base per generation [6,7,34–36]. The mutation rate for copy number variants was estimated to be 0.03 CNVs larger than 500bp [12] and 0.012 CNVs larger than 100kbp [37] per haploid genome. Our data indicate a mutation rate of 0.68 x 10^{-9} indel (1-20bp) per base per generation and 0.08 SVs (>20bp) per haploid genome (or 0.16 SVs per generation). The higher SV rates reported here in comparison to previous array CGH studies result from greater power to interrogate the full size range of structural changes and the ability to capture retrotransposon insertions (**Fig. 2A**). For example, when considering only CNVs larger than 500bp or larger than 100kbp our data provide a rate of 0.041 and 0.0077 per haploid genome, respectively. In addition, a substantial proportion (15%) of the observed *de novo* SVs were retrotransposition events, allowing us to empirically estimate the rate of retrotransposition in the population to 0.023 (1/43) per generation. This is in line with estimates based on diseased subjects and on comparative genomics studies, which range from ~1/20 for Alu elements to ~1/100-200 for L1 long interspersed elements (LINEs) per generation [38,39].

Although the above *de novo* SV rate implies that only one in seven children bears such a mutation, we found six offspring with two and one with three *de novo* SVs (**Supplemental Table 2**). Such co-occurrence of multiple SVs is unexpected under a uniform distribution of the 41 *de novo* SVs across the 258 children ($p = 0.0074$). One individual carries two *de novo* deletions (327kbp and 1.5kbp) on maternal chromosome 18 within a distance of 202kb of each other. This close placement of two *de novo* SVs is unlikely to be random ($p = 1.35 \times 10^{-4}$). Together, these data suggest possible differences in the effects of environmental factors or the vulnerability for acquiring *de novo* SVs per family [40]. We did not find evidence for a non-uniform distribution of the *de novo* indels across offspring ($p = 0.061$).

Elevated paternal mutation rates

Large-scale genome sequencing of families with disorders has shown that most *de novo* SNVs have a paternal origin, with a significant increase of *de novo* mutation burden with paternal age [3–5,40,41]. In addition, the majority of sporadic *de novo* CNVs and cytogenetically balanced genomic rearrangements in patients with congenital disorders are also paternal in origin [20,42]. However, it is unclear whether this bias is also present for *de novo* SVs and indels occurring in the general population. Using reads spanning neighboring phase-informative polymorphisms, we assigned a parental haplotype to 20% of the indels (39 paternal, 20 maternal) and 71% of the SVs (20 paternal, 9 maternal). We observed a significantly larger

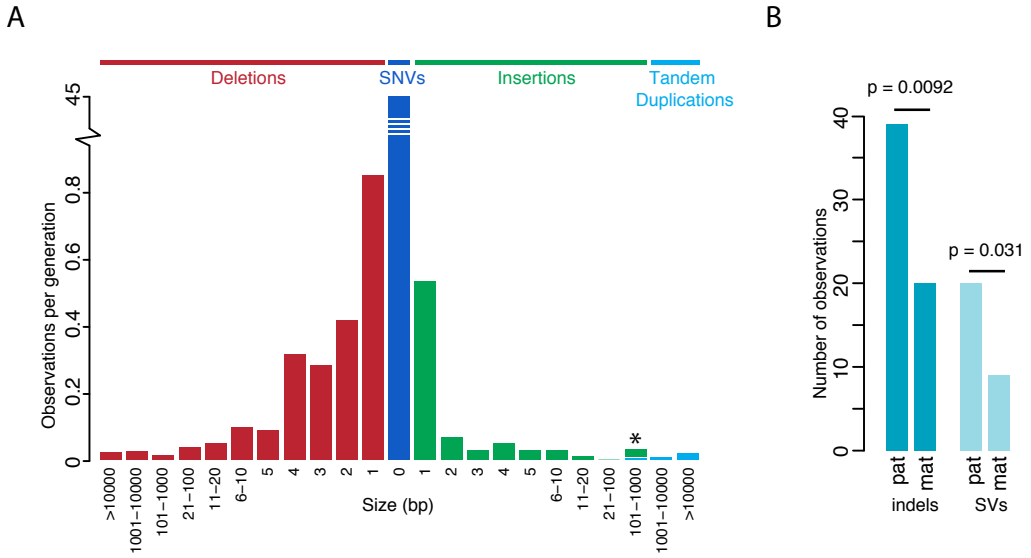


Figure 2 | Frequency of *de novo* indels and SVs

(A) Size-frequency distribution of 325 validated *de novo* indels and SVs identified in this study. In addition, the frequency of *de novo* SNVs is shown [4]. The asterisk denotes a size bin containing one *de novo* tandem duplication and six *de novo* retrotransposon insertions.

(B) Barplot indicating the numbers of *de novo* indels and SVs on paternal and maternal haplotypes.

fraction (66.1%) of indels and SVs arising on paternal chromosomes than on maternal chromosomes (p_{indel} = 0.0092, p_{SV} = 0.031, **Fig. 2B**), further emphasizing the contribution of the paternal germline to human mutations. There was no significant correlation between *de novo* structural change occurrence and paternal age, possibly due to the limited number of observations.

Indel formation

We found a total of 277 simple indels with a deletion to insertion ratio of 2.74:1. This ratio is consistent with previous reports [43,44], although it is possible that this number is influenced by differences in detection power between insertions and deletions. To investigate the mechanisms of formation of these indels, we categorized their sequence content and flanking context (**Table 1**). Most of the *de novo* indels in our data (59.9%) were found in repeat regions or resulted in local copy count changes, meaning that the long allele can be obtained by copying part or all of the short allele. More specifically, we found 28 indels in homopolymer runs (HR), 27 in tandem repeats (TR) and 111 indels resulting in a copy count change outside repeat regions (CCC). Copy-count-changing indels show a relatively balanced deletion to insertion ratio of 1.5:1. They likely arose through polymerase slippage, a process by which the leading and lagging strand become mispaired during DNA replication causing a few bases to be duplicated or deleted. Although we confirm a strong enrichment for indels in homopolymer runs (HR, $p < 2.2 \times 10^{-16}$) and tandem repeats (TR, $p < 2.2 \times 10^{-16}$) [43], they only represent 19.9% of our observations. This is significantly less than what we observe in polymorphic indels in our data (44.2%) and in previous reports (46.0%, Montgomery et al. 2013), possibly indicating low selective pressures on these repetitive regions (**Fig. 3A**).

The remaining 40.1% of the observed *de novo* indels occurred in non-repeat regions and did not lead to a copy count change (non-CCC). These indels are likely the result of imperfect double-stranded DNA break repairs by non-homologous end-joining (NHEJ) which can create indels at the repair junction. The very high deletion to insertion ratio of 12.9:1 we observe for these indels supports their occurrence through NHEJ [45]. This provides a mechanistic explanation for the relative depletion of short insertions in the overall size spectrum of *de novo* variation (**Fig. 2A**). We found palindromic sequences (≤20bp away, ≥6bp long) flanking eight of these deletions, suggesting that a secondary structure such as a hairpin loop played a role in their formation [43,45]. Another five non-CCC indels presented microhomologies of at least 4bp, possibly indicating emergence through microhomology-mediated end joining (MMEJ) [46].

Class	Example[1]	Observations	Possible Mechanisms	Sequence Features
Homopolymer Run	Ref. CTGAGGAAGAGTTTTTTTTACA *De Novo* CTGAGGAAGAG–TTTTTTTACA	21 insertions 7 deletions	Polymerase slippage	Repeat context
Tandem Repeats	Ref. CTACCCCAGGCAGAGAGAGAAA *De Novo* CTACCCCAGGC----AGAGAAA	8 insertions 19 deletions	Polymerase slippage	Repeat context
Copy Count Changing	Ref. CAGAAGG----TAGCTAGTCAG *De Novo* CAGAAGGTAGCTAGCTAGTCAG	37 insertions 74 deletions	Polymerase slippage	Local copy count change
	Ref. CTAAAGGGCAGTCTTGCAAAAG *De Novo* CTAAAGGGCAG--TTGCAAAAG	8 insertions 90 deletions	NHEJ[2]	Blunt or microhomolgy at breakpoints
Non Copy Count Changing	Ref. AGTCAAAAACCAAAGTTTTGAA *De Novo* AGTCAAAAACCA---TTTTGAA	8 deletions	NHEJ[2] / hairpin loop	Palindrome (≥6bp) in surrounding context (≤20bp)
	Ref. GGGGAGAATTGAGACTTGATCA *De Novo* GGGGAGAA-------TTGATCA	5 deletions	NHEJ[2] / MMEJ[3] / replication slippage	Microhomology ≥4bp at breakpoints
	Ref. ACTCACAAAAAAATTTTTTTCC *De Novo* ACTCACAAAAA-TTTTTTTTCC	2 variants	Polymerase slippage	Repeat context
Complex	Ref. CACATGGGCTTCC-----TGTC *De Novo* CACATGGGCTGGAGCCCATGTC	8 variants	SD-MMEJ[4] TMEJ[5]	Palindromic or templated insertion
	Ref. CCAAAGTGCTGGGATTACAGGC *De Novo* CCAAAGTGCTC-GATTACAGGC	4 variants	Unknown	None

Table 1 | Indel classes and mechanisms

[1]All examples are chosen from observed validated *de novo* indels and their positions are given with respect to the start of the variant on the human reference genome build 37. In the alleles column, "A" denotes the ancestral allele and "D" the derived allele. Differences between the ancestral and derived alleles are highlighted in bold. Repeats and palindromes are underlined with straight and wavy lines respectively.

[2]NHEJ: Non-homologous end joining

[3]MMEJ: Microhomology-mediated end joining

[4]SD-MMEJ: Synthesis-dependent microhomology-mediated end joining

[5]TMEJ: Theta-mediated end joining

In addition to the 270 simple indels, we also identified 14 complex indels (**Table 1 and Fig. 3B, Supplemental Table 3**) replacing multiple bases (2-10bp) by a different sequence (1-11bp). Although similar types of complex indels have been described previously [47], this class of variants has largely been neglected in sequencing studies and is therefore absent from variant repositories. As they represent 4.8% of the *de novo* indels in our data, we speculate that this type of polymorphism may be relatively common. Indeed, we found that 5.1% of inherited indels in the GoNL samples seem complex. One of the difficulties posed by such variation when studying polymorphisms is that they can be due to a combination of multiple separate indels or SNVs or as a single complex variant. We provide here the first *de novo* observation of such variations in humans, showing that they indeed arose as part of a single mutational event.

In contrast to simple indels, only two complex indels are located in repetitive regions, indicating that polymerase slippage is unlikely to be a major contributor to their formation. Strikingly, five of them form palindromic repeats (≥6bp), a proportion significantly elevated when compared to simple insertions ($p = 0.0015$). The inserted bases for another three variants appeared to have been templated from the neighboring sequence. Such palindromic and templated complex indels have been reported in model organisms around double-stranded break repairs through synthesis-dependent microhomology-mediated end joining (SD-MMEJ) [48] and theta-mediated end joining (TMEJ) [49]. The formation of these indels likely follows a multi-step process involving resection of break ends, hairpin formation, microhomology-mediated annealing and DNA synthesis. **Fig. 3B** shows an example of how a *de novo* complex event we observed could have arisen through SD-MMEJ.

SV formation

To obtain insights into the origin of *de novo* SVs in the general population, we experimentally fine-mapped their breakpoints at base-pair resolution and assigned a formation mechanism (**Fig. 4A, Supplemental Table 2**) [50]. The majority (N = 24, 58.5%) of the SVs larger than 20bp likely arose via non-homologous repair (NHR) as their breakpoints presented little or no homology (0-6bp, N = 19) or short inserted sequences (1-18bp, N = 5). The breakpoints junctions of eight SVs (19.5%) contained long homologous sequences (28bp to 12kb) indicating formation by non-allelic homologous recombination (NAHR) (**Supplemental Table 2**). Three variants (7.3%) were found within a region with a variable number of tandem repeats (VNTR).

We also identified 6 *de novo* mobile element insertions (14.6% of SVs), all short interspersed elements (SINE) retrotransposon insertions of the AluY family (**Supplemental Fig. 2, Supplemental Table 2**). Class I transposable elements,

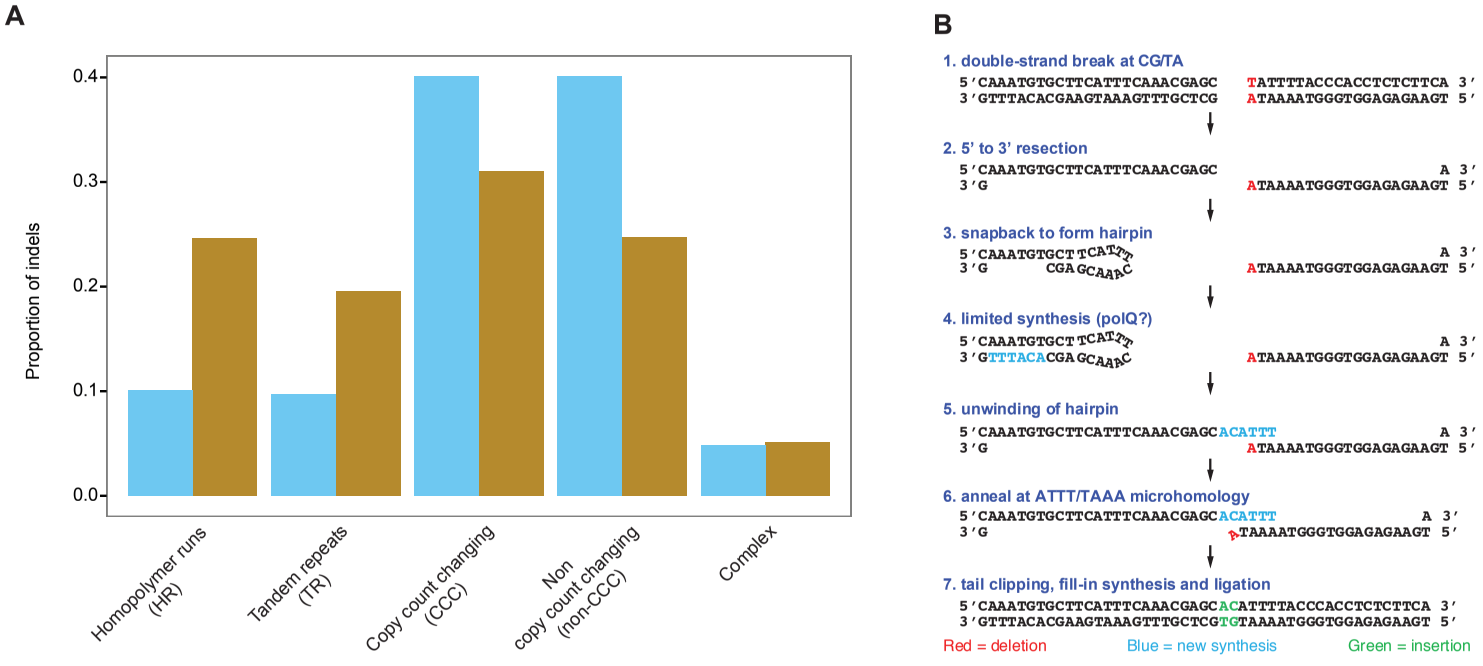


Figure 3 | Overview of *de novo* indel classes and their formation mechanisms

(A) Proportion of *de novo* and inherited indels by class. Inherited indels exhibit a 2.3 fold enrichment in indels located in homopolymer runs (HR) and tandem repeats (TR) when compared to *de novo* indels, suggesting lower selective pressures in these regions.

(B) Outline of a plausible 7-step process that could account for the formation of a complex *de novo* indel by SD-MMEJ.

including SINE and LINE elements, constitute nearly half of the human genome, although most of these elements were acquired before separation of the human lineage [39]. The sequences of the breakpoint junctions of the *de novo* AluY retrotransposon insertions all indicate the presence of target site duplication (TSD) of 3-16bp, and poly-A tails (**Supplemental Fig. 2**); both well-known signatures of retrotransposon integration [39].

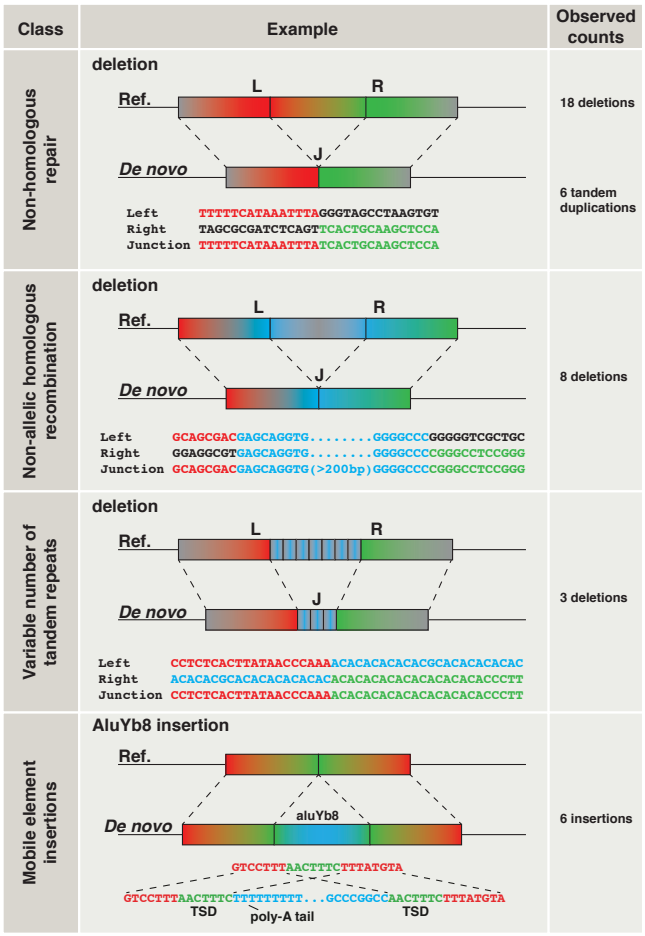
Remarkably, in one instance of interchromosomal integration we found three breakpoint junctions leading to the joining of two small DNA fragments – one from chromosome 3 (163bp) and another from chromosome 19 (179bp) – into chromosome 4 (**Fig. 4B**). We propose that this complex rearrangement has also occurred through retrotransposition, because the fragment from chromosome 19 contains part of an AluY element and no DNA is lost at the original genomic positions of the inserted sequences. Furthermore, the breakpoint on chromosome 4 likely involved a staggered cut with three overhanging nucleotides, which appear as TSDs in the final product. The fragment on chromosome 3 is close (1.7 kbp) to the 3'UTR of the *PPARG* gene. We hypothesize that the fragment could represent a retrocopy of an RNA product from this region, e.g. an elongated version of the *PPARG* mRNA or another transcript.

We compared the proportion of *de novo* SVs derived from each of four mechanisms with inherited SVs from the GoNL project. This revealed a larger proportion of mobile element insertions (MEI, 40.8%, $p = 0.029$) for inherited SVs and a lower proportion of NHR (30.3%, $p = 0.0072$), while similar proportions of VNTR (10.5%) and NAHR (18.4%) mediated variants were found. In addition, we compared the proportion of each SV mechanism with those reported previously [51–54] (**Supplemental Table 4**). We found substantial differences between studies, which probably reflect different methods for variant detection and for assigning SV mechanism (Pang et al. 2013). Furthermore, we should note the caveat that SVs involving long stretches of homologous sequence can be missed by short-read sequencing.

Functional impact of *de novo* structural changes

Although none of the *de novo* indels overlapped with protein-coding exons, in total 6 large *de novo* SVs (3.7kbp – 327kbp) affect coding regions, resulting in exonic duplications of *BANK1* (1 exon), *PROC*, *GCNT3*, *GTF2A2* and *BNIP2* (complete genes), and deletions in *LYN* (1 exon), *PTPRM* (6 exons) and *UBR5* (8 exons) (**Fig. 5, Supplemental Table 2**). Four SVs potentially disrupt gene function by changing reading frames, introducing premature stop codons or truncating the protein. The duplication of exon 1 of *BANK1* possibly leads to a premature stop-codon

A



B

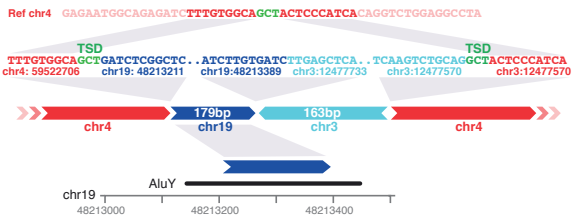


Figure 4 | Mechanisms contributing to the formation of *de novo* SVs

(A) Overview of four SV formation mechanisms, including examples and observed counts for each of these. L=left flank; R=right flank; J=junction.

(B) Schematic structure of a complex *de novo* interchromosomal SV involving an insertion of DNA from chromosomes 3 and 19 into chromosome 4. TSD: target site duplication.

depending on splicing mode. Exonic deletions within *LYN* and *UBR5* cause frame-shifts, while deletion of 6 exons of *PTPRM* leads to an in-frame, but shortened transcript. Examination of these genes in the exome sequencing database from the Exome Aggregation Consortium (ExAC, exac.broadinstitute.org) revealed that all of them contain heterozygous loss of function mutations in the population. Since the ExAC data are derived from persons without severe pediatric phenotypes, the heterozygous changes in these genes possibly have no severe early developmental consequences. Mutations in two of the affected genes - *BANK1* and *PROC* - are associated with systemic lupus erythematosus [55] and thrombophilia [56], respectively. However, *PROC* and *BANK1* duplications – as observed in our study – have not been reported to be associated with a clinical phenotype and the offspring carrying these *de novo* SVs appeared healthy at the time of sampling (aged 39 and 32).

Next, we compared the genomic footprints of *de novo* SVs and indels with SNVs. Consistent with recent studies involving families with disorders [3,5,41,57], an average of 45 *de novo* SNVs per child were detected in the GoNL Project [4]. While the cumulative burden of *de novo* indels was only 7.1bp per child, we found that despite their lower frequency *de novo* SVs affected on average 4,084 genomic bases (**Fig. 6A**). This relatively large impact of SVs was also found in coding regions where an average of 28.62 coding bases per generation were affected by *de novo* SVs, while only 0.55 coding bases per generation were mutated by *de novo* SNVs (**Fig. 6B**). The larger number of affected bases for SVs relative to SNVs is largely due to their difference in size. We observed that per offspring 17.9 times more genes are hit by *de novo* SNVs (0.55) versus SVs (0.03) (**Fig. 6C**). However, only 5% of *de novo* SNVs is potentially disruptive (stop gained, stop lost, splice-site change), whereas 50% (4/8) of the *de novo* SVs possibly have a major impact on gene structure and function (**Fig. 5**).

Finally, we investigated differences in the genomic footprint of *de novo* and inherited SVs and SNVs identified in the GoNL data. We found that on average large *de novo* SVs (>20bp) affect 90.6 times more genomic bases, 52.0 times more coding bases and 60.1 to 114.7 times more bases marked by histone modifications than *de novo* SNVs (**Fig. 6D**). In contrast, inherited SVs affected on average only 1.6 times more bases when compared to inherited SNVs. Altogether, these data demonstrate the overall impact of *de novo* SVs on the genome when compared to *de novo* SNVs and indels.

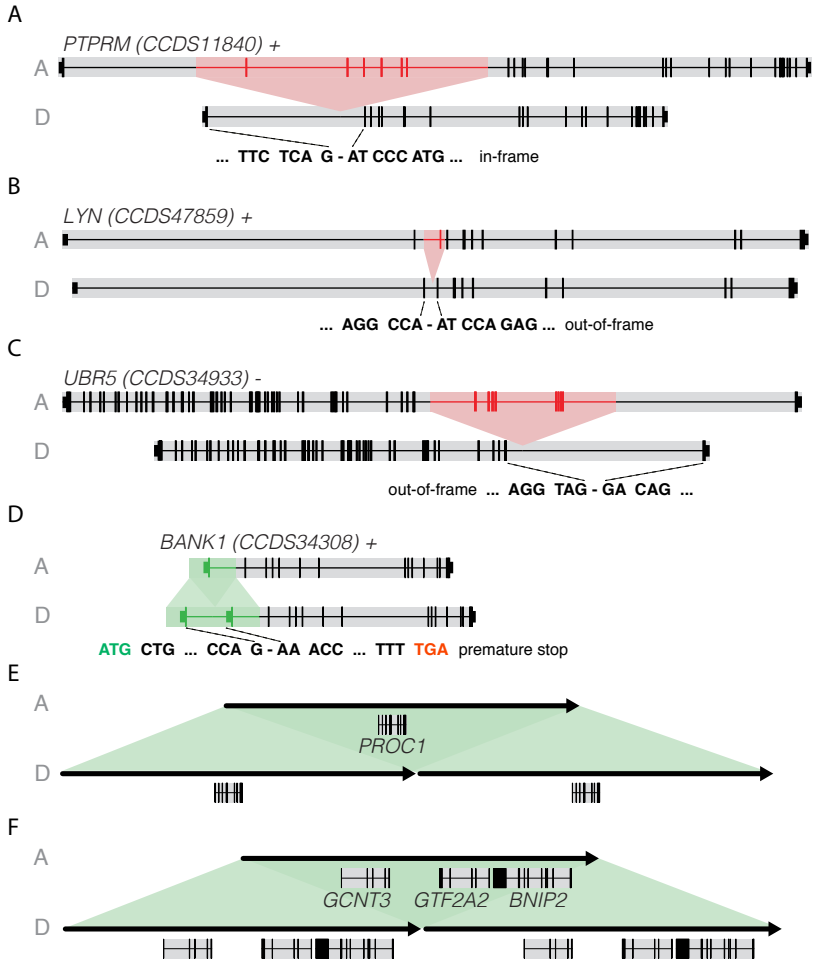


Figure 5 | Effect of *de novo* SVs on protein coding genes

(A) Deletion of 6 exons of *PTPRM* resulting in an in-frame shortened gene.

(B) Deletion of 1 exon of *LYN* causing an out-of-frame effect at the transcript level.

(C) Deletion of 5 exons of *UBR5* causing an out-of-frame effect at the transcript level.

(D) Duplication of 1 exon of *BANK1*, possibly resulting in a premature stop. Although *BANK1* variations are associated with systemic lupus erythematosus [55], the offspring carrying this duplication appeared healthy at age 39.

(E) Duplication of the entire *PROC1* gene, variations in which have been associated with thrombophilia [56] but the offspring was healthy at age 32.

(F) Duplication of 3 entire genes (*GCNT3, GTF2A2, BNIP2*). The *de novo* SVs outlined in panels A-D result in novel gene structures. Duplications are shown in green and deletions in red. A, ancestral allele; D, derived allele.

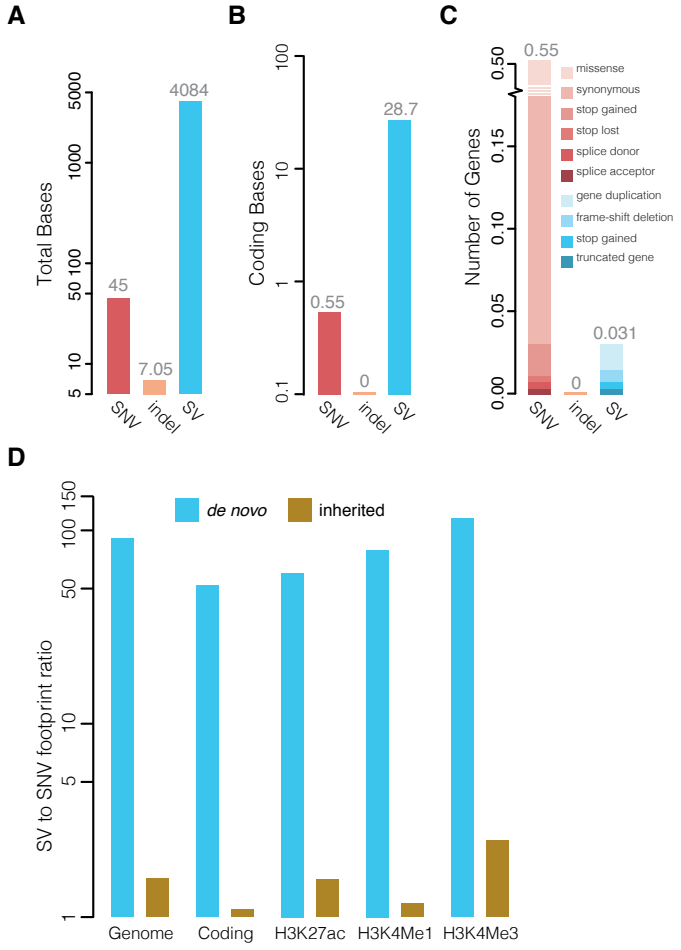


Figure 6 | Functional impact of *de novo* indels and SVs

(A) Average number of genomic bases affected by *de novo* SNVs, indels and SVs per child.

(B) Average number of coding bases affected by *de novo* SNVs, indels and SVs per child.

(C) Average number of genes affected by *de novo* SNVs, indels and SVs per child. The relative frequencies of the effects of the variations on the gene are indicated.

(D) Comparison of the footprint of *de novo* (blue bars) and inherited (brown bars) large SVs (>20bp) relative to the footprint of SNVs. The footprint was computed genome-wide, in protein-coding regions and genomic regions marked by H3K4Me1, H3K4Me3 and H3K27Ac based on data from 16 cell lines from the ENCODE project [62]. The y-axis shows the ratio of the average number of affected bases per offspring relative to SNVs.

Discussion

The human genome continuously evolves as a result of mutation and selection. The relatively low rate of SV and indel formation in the human genome requires large sample sizes involving parent-offspring families to capture the full spectrum of *de novo* changes that alter genome structure every generation [7]. Moreover, the detection and genotyping of these variants remains challenging given their diversity in both size and type [58]. By leveraging multiple calling approaches, we have provided a detailed picture of the landscape of *de novo* SVs and indels in the human genome based on whole-genome sequencing of 250 families. The comprehensive detection and validation approach we undertook should be broadly applicable to future studies of *de novo* indels and SVs.

Our work demonstrates that both *de novo* indels and SVs originate primarily in the paternal germline. These data complement recent findings, which indicated a strong paternal bias for *de novo* SNVs [3–5]. Furthermore, we provide empirical estimates for the rate of *de novo* SVs and indels across the complete size spectrum, including relative frequencies of different variant sizes and types. These rates define a baseline for the general population and will help guide the interpretation of *de novo* indels and SVs in the diagnosis of individual patients [16]. Roughly 15% of patients with intellectual disability or congenital abnormalities harbor an apparently causative CNV, most of which occur *de novo* [59]. Estimating pathogenicity of these CNVs is based on their overlap with known disease CNVs, protein coding genes and control databases, but should also consider the background rate of large CNVs as described here. Specifically, we find that changes in gene structure - i.e. deletion or tandem duplication of entire exons - occur at a rate of 1 in 43 offspring in the general population.

In spite of their low frequency large *de novo* SVs have a substantial impact on the genome. Due to their larger size, the average genomic footprint of *de novo* SVs is much larger than that of *de novo* SNVs and they are much more likely to hit a coding region. Indeed, 14.6% of the *de novo* SVs we observed affected exons, whereas only about 1.3% of the *de novo* SNVs did. The considerable influence of *de novo* SVs is however primarily driven by a limited number of *de novo* SVs altering multiple kilobases of genomic sequence in a single generation. These rare but large variants may be quickly removed from the population by purifying selection, particularly when they hit genes or other important genomic elements (**Fig. 5**) [12,51]. This may explain why inherited SVs and SNVs affect a similar number of bases.

Previous studies have convincingly shown that large and dramatic genome changes introduced by large structural mutations can be associated with a multitude of pathological conditions [16]. In this study we demonstrated that a broad range of *de novo* indels and structural mutations is also characteristic for individuals obtained from a general human population.

Methods

Whole genome sequencing and alignment

A total of 250 Dutch families (231 trios, 8 quartets with dizygotic twins and 11 quartets with monozygotic twins) were selected for the study, without phenotypic ascertainment. Genomic DNA from nucleated blood cells was obtained from each individual. Library construction and whole genome sequencing was performed at BGI using the Illumina HiSeq 2000 platform (500 bp insert size, 90bp paired-end reads).

Reads were aligned to the GRCh37/hg19 human genome reference using BWA 0.5.9-r164 [60] and processed following the Genome Analysis Toolkit (GATK) best practices v2 [24]: duplicate reads were marked using Picard tools (http://picard.sourceforge.net), reads were realigned around indels using GATK IndelRealinger and base quality scores were recalibrated using GATK BaseRecalibrator. Additional details regarding the study design, sequencing and alignment can be found in previous publications [4].

Detection of *de novo* variants

Indels were called using GATK UnifiedGenotyper [24] and PINDEL [25] and all calls were further genotyped with GATK HaplotypeCaller (**Supplemental Methods**). We used GATK PhaseByTransmission (PBT) to call *de* novo variants from the GATK UnifiedGenotyper and HaplotypeCaller calls using a sensitive mutation prior of 10^{-4} per base per generation. We then kept only calls with (a) no evidence of the non-reference allele in the parents, (b) no non-reference allele called in any other GoNL sample, (c) at least 2 reads supporting the non-reference allele, (d) a PBT posterior of at least Q20. PINDEL calls with non-reference reads in the child only, at least 2 reads supporting the non-reference allele in the child and no significant strand bias were kept as putative *de novo* indels. All putative *de novo* indels from either method were experimentally validated in 92 of the families (including 7 quartets with dizygotic twin pairs).

De novo structural variant (SV) were called and filtered independently by 11 algorithms based on (a combination of) the following approaches: gapped/split

read mapping (PINDEL [25], GATK UnifiedGenotyper [24], GATK HaplotypeCaller), analysis of discordant pairs (Breakdancer [26], 1-2-3-SV, GenomeSTRiP [29], MATE-CLEVER [30]), read depth analysis (CNVnator [27], DWAC-Seq, FACADE [28]). In addition, Mobster was used to call *de novo* mobile element insertions (MEIs) [31]. For each algorithm, variant calls confined to kid(s) of a single family, but not detected in any other GoNL samples were selected and visually evaluated with IGV to discard evident inherited events due to false negatives in parental samples. We then created a union of all remaining calls by merging variants detected by multiple methods in the same child based on SV type and overlapping coordinates. We retained the most precise breakpoints for each variant based on the calling algorithm (in order: split-read, discordant read-pairs, read-depth). Local *de novo* assembly (SOAPdenovo [61]) was used for breakpoint fine-mapping. A detailed description of the tools, settings, filtering and variant calls is provided in the **Supplemental Methods**.

Experimental validation

To exclude inherited variants and alignment artifacts, all candidate mutations were manually checked within samples of the corresponding family using the IGV genome browser [32]. For large SVs (greater than 100bp), breakpoints of each candidate mutation were reconstructed by local *de novo* assembly of corresponding regions using SOAPdenovo [61]. Oligonucleotide primers for amplification of a genomic segment containing the variant (for mutations smaller than 100bp) or variant breakpoints (for larger structural variants) were designed using Primer3 software. PCR amplifications were performed using DNAs of each member of the affected family. PCR products were visualized on 1% agarose gels and resequenced with Sanger, IonTorrent or MiSeq (2x250 bp) technologies. A detailed description of the protocol used for genotyping the resequenced variants is provided in the **Supplemental Methods**.

Accessible genome and detection sensitivity analysis

Coverage from NGS data fluctuates both in systematic and random ways affecting *de novo* variation discovery power. Depending on variant size and type different signals from the data are used for discovery, which leads to variations in detection power among the different classes of variants. Indel detection mainly relies on base coverage. Thus, we restricted our indel analyses to regions of the genome covered by at least 5 reads in the child and 10 reads in each parents and no more than 100x across the entire trio. To estimate the indel sensitivity in these regions compared to deep sequencing, we combined the data of each of the 11 monozygotic twin pairs, leading to 11 samples with an average

base coverage of 28x. We then called singleton polymorphisms (in the parents) discovered in GoNL [4] by GATK UnifiedGenotyper and PINDEL in these merged samples (filtered calls). For each heterozygous genotype found in these merged samples, we assessed whether it was also called in the original separate twins. We only considered genotypes with quality of at least Q20 (both in the merged and separate samples) to ensure that these would have been confidently called as *de novo* indels. We found that 93.2% of the indels (92.6% of the insertions and 93.5% of the deletions) found at 28x were also found at 14x. In addition, 24 parents from the GoNL Project were independently deeply sequenced whole-exome on the Illumina HiSeq platform. We called indels in these exomes using PINDEL, GATK UnifiedGenotyper and GATK HaplotypeCaller and considered all indels called by at least two algorithms reliable without further filtering. We then considered all singletons across the 24 whole exomes that were either not discovered (false negatives) or also singletons (true positives) in our whole-genome data to evaluate our sensitivity for *de novo* calls. We found that 83.3% of the singleton indels detected in the deep exome sequencing data were found back in the medium coverage whole-genome sequencing data (75.1% of the insertions and 87.4% of the deletions). We addressed false positives by experimental validation of all *de novo* indel candidates in 92 families.

For large SVs, the algorithms we used mostly rely on discordantly mapped read-pairs, supported by read depth analysis. Discovery power therefore mainly depends on the physical coverage. Given the relatively small number of *de novo* large SV calls in our dataset, all events showing a non-reference allele in a child but absent from the parents were subjected to experimental validation regardless of the coverage in the parents. We used 4 discordant read-pairs as a minimum threshold to call an SV in our discovery phase. On average, 98.8% of the known genome (non-N bases) of each haplotype were physically covered by 4 or more read pairs in the children.

Similarly to the indel sensitivity analysis, we used 11 MZ trio samples to compare the sensitivity of SV calling between 14x and 28x coverage data. First, we selected deletions, which were detected in the combined MZ set (28x coverage), retaining only those having evidence for the alternative allele in only one of parents. The resulting set encompasses deletions in a heterozygous state, which is also expected for *de novo* variants. For each heterozygous variant in a combined MZ twin pair (28x coverage) we checked whether it was also found in each individual MZ sample (14x coverage) by one or more algorithms (PINDEL, MATE-CLEVER, 123SV and Breakdancer). We observed 69.4% sensitivity (204,624 calls detected at 14x vs 294,826 detected with 28x) for short deletions sized between 20 and

99 bases. Larger variants, exceeding 100bp are detected with 85.8% sensitivity (74,039 calls at 14x vs 86,276 at 28x).

The sensitivity for detection of MEIs by Mobster was separately tested using a set of 134 validated MEIs (127 Alu; 6 L1; 1 SVA) from 1000 Genomes sample NA12878 [33]. We subsampled the original data for NA12878 resulting in an average coverage of ~14X. By running Mobster on this subsampled dataset we could detect 104/134 (77.6%) validated MEIs based on a combination of both single and double cluster predictions.

Parental origin

We used genotypes from phased haplotypes [4] to interrogate the parental origin of *de novo* indels and SVs. For indels, we identified read-pairs containing both the *de novo* allele and a phase-informative SNP allele (heterozygous in the child). The parental origin was derived from the phased SNP allele.

Parental haplotypes for SVs were determined from allele ratios at polymorphic genome positions that overlap with a *de novo* variant. Assignment to the paternal or maternal haplotype was made if: i) one or more homozygous alleles in the offspring are located inside a *de novo* deletion and could only be inherited from one parent; ii) one or more polymorphic SNPs in offspring are located inside a *de novo* duplication and have a 2:1 (or 1:2) ratio with the reference allele and can be assigned unambiguously to either the paternal or maternal haplotype; iii) a SNP in the offspring was found to locate specifically within discordant read pairs supporting the *de novo* SV and could be assigned to either the paternal or the maternal genome.

Paternal and familial biases

We tested for enrichment of *de novo* mutations on the paternal haplotypes using a one-tailed binomial test and found that both indels ($p = 0.092$) and SVs ($p = 0.031$) were indeed enriched. Additionally, we fit a linear model to the number of *de novo* indels in the 99 independent offspring and the father's age at conception correcting for coverage but did not find a significant association ($p = 0.24$). It is possible that we are underpowered to find such an effect given the relatively low number of mutations and the narrow distribution of the father ages around 29.4 years old. For *de novo* SVs, we found that the median age of the father at conception was the same (29 years old) for children carrying a *de novo* SVs and for those who did not.

In order to test whether the distribution of *de novo* variants across children was random, we used a multinomial model with equal probability for each child to receive a *de novo* variant. We performed a goodness-of-fit test of the observed data to the model. Given the low number of observations, we computed the p-value using a Monte Carlo test with 100,000 replicates. We also explored the effect of unequal genome coverage among samples on detection on *de novo* SVs and found that it has a very limited effect.

In one sample, we observe two SVs occurring on maternal chromosome 18 at a distance of 201kbp. We computed the probability of observing 2 independent deletions so closely located by direct enumeration. Let E_1 and E_2 be the smaller and larger deletion events, respectively, having respective lengths of L_1 and L_2 bases. Neglecting edge effects at the ends of chromosomes, the number of ways E_1 could be placed in the genome is $(G - L_2 + 1) - (L_1 + L_2 - 1)$, where G is the nominal genome size. The first term represents the possible placements of E_1, while the second represents the number of inadmissible placements that would result in the collapse of both events into a single indistinguishable one. If D is the observed distance in bp between the two events, then the number of the total placements that are significant, i.e. at least as extreme as the observation, is 2 D, since E_1 could be on either side of E_2 implying a "two-sided" test. The ratio of these two counts represents the tailed P-value. Given $G \approx 3 \times 10^9$ and the observed values L_1 = 1,552, L_2 = 326,954, and D = 201,790, we find a P-value of 1.35×10^{-4}.

Computation of mutation rates

To compute the indel rate, we used validated *de novo* indels in 99 children from 92 families, including 11 quartets with monozygotic (MZ) twins, 7 quartets with dizygotic (DZ) twins and 74 trios. We only used one child from each of the MZ twin pairs and considered the 14 children from the 7 DZ twin pairs as independent for this analysis. To rule out possible inter-sibling correlation with respect to *de novo* indels, we tested whether the absolute difference in number of *de novo* mutations (referred to as distance below) was smaller in DZ twin pairs when compared to pairs of unrelated children. We found that the distribution of distances between DZ twin pairs was not different that of unrelated children pairs using a one-tail Kolmogorov-Smirnoff test ($p = 0.68$). We further showed that the mean distance in a DZ twin pair was no different from the mean distance in an unrelated children pair ($p = 0.59$) through 10,000 permutations (for each permutation a random 7 pairs of unrelated children was drawn).

To compute the per-base indel rate in our study, we only considered accessible regions bases of the genome as described above. The estimated rate was then computed as the sum of *de novo* indels divided by the sum of accessible bases in the 99 children.

The SV rate was computed over 258 children from 250 families including mono- and dizygotic twin pairs. Only one child was considered for each of the monozygotic twin pairs and children from dizygotic twin pairs were considered as genetically independent with respect to *de novo* SVs. The rate was calculated by dividing the total number of *de novo* SVs observed (N = 41) by the 258 children times 2 transmitted haplotypes. We also report the rate for *de novo* MEIs (N = 6, including one interchromosomal event which involved an *AluY* element) computed in a similar fashion.

Indel and SV formation mechanisms

Indels were annotated using the classification proposed by Montgomery *et al.* [43], except for (i) Predicted Hotspots (PR) that we did not use since they were not readily available and (ii) Complex indels that are new in our data. A full description of indel classification and calculation of frequencies for each category is given in **Supplemental Methods**.

Analysis of mutation formation mechanisms of SVs was performed using BreakSeq software v. 1.3 [50]. Genome references (GRCh37/hg19, panTro4, rheMac3, ponAbe2), repeat-masked reference, Blat program and annotation databases were downloaded from the UCSC website (http://genome.ucsc.edu). We used default settings for annotation of variants longer than 50 base pairs (annotation of shorter variants is not supported by BreakSeq). Because BreakSeq does not support mechanism prediction for tandem duplications, we used BLAST to determine the similarity between 200bp genomic segments at the beginning and end of the tandemly duplicated segment (duplication start +/- 100bp and duplication end +/- 100bp). Tandem duplications showing homologous sequences at the breakpoints with a similarity > 90%, at least 50bp in length and in the correct orientation (+/+ alignment) were assigned to the NAHR mechanism, while the remaining tandem duplications were classified as NHR.

Genome context and functional effects

To determine the number of affected genomic bases by indels, we calculated one base per insertion, the variant length for each deletion and the number of replaced genomic bases for complex indels. For SVs, we took the SV length for duplications and deletions, while for MEIs the lengths of the target site duplications were used. To determine the number of affected coding bases, we intersected *de novo* substitutions, indels and SVs with protein coding exons retrieved from the Ensembl database (version 74). Both 3'-UTR, 5'-UTR sequences and non-coding exons were excluded from the intersection. *De novo* variants were overlapped with three types of ENCODE histone modification datasets (H3K4Me1, H3K4Me3 and H3K27Ac) derived from 16 different cell lines (http://genome.ucsc.edu/cgi-bin/hgFileSearch?db=hg19, **Supplemental Table 5**) [62]. For each of these histone modification data we determined the number of bases that were affected by *de novo* substitutions (n=11,618), indels (n=284) and SVs (n=41), respectively. Subsequently, the average number of affected bases per individual was calculated (n=258 for substitutions and SVs; n=99 for indels). We repeated the procedure for inherited variation (indels, inversions, deletions, duplication, MEIs) found in the GoNL project [4] in order to compare the footprint of inherited and *de novo* variation.

Data access

Sequence data have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI), under accession number EGAS00001000644.

Acknowledgements

We thank Matt Wyczalkowski for help with illustrations and Craig Grove for textual editing. The GoNL Project is funded by the Biobanking and Biomolecular Research Infrastructure (BBMRI-NL), which is financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007).

References

1. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437,** 88–93 (2005).
2. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13,** 824–824 (2012).
3. Michaelson, J. J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151,** 1431–1442 (2012).
4. Francioli, L. C. *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–25 (2014).
5. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).
6. Besenbacher, S. *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat. Commun.* **6,** 5969 (2015).
7. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. *Trends Genet.* **29,** 575–584 (2013).
8. Korbel, J. O. *et al.* Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science (80-.).* **318,** 420–6 (2007).
9. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305,** 525–528 (2004).
10. Lafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36,** 949–951 (2004).
11. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37,** 727–732 (2005).
12. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).
13. Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315,** 848–853 (2007).
14. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–454 (2006).
15. Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43,** 838–846 (2011).
16. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61,** 437–455 (2010).
17. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316,** 445–449 (2007).
18. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320,** 539–543 (2008).

19. Malhotra, D. & Sebat, J. CNVs: Harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148,** 1223–1241 (2012).

20. Hehir-Kwa, J. Y. *et al.* De novo copy number variants associated with intellectual disability have a paternal origin and age bias. *J. Med. Genet.* **48,** 776–778 (2011).

21. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13,** 565–575 (2012).

22. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* **22,** 221–7 (2014).

23. Londin, E. R. *et al.* Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. *BMC Genomics* **12,** 464 (2011).

24. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).

25. 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,** 2865–71 (2009).

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

27. 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).

28. Coe, B. P., Chari, R., MacAulay, C. & Lam, W. L. FACADE: A fast and sensitive algorithm for the segmentation and calling of high resolution array CGH data. *Nucleic Acids Res.* **38,** (2010).

29. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43,** 269–276 (2011).

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

31. Thung, D. T. *et al.* Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biol.* **15,** 488 (2014).

32. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).

33. Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.* **7,** (2011).

34. Kondrashov, A. S. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Hum. Mutat.* **21,** 12–27 (2003).

35. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 961–968 (2010).

36. Ramu, A. *et al.* DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat. Methods* **10,** 985–7 (2013).

37. Itsara, A. *et al.* De novo rates and selection of large copy number variation. *Genome Res.* **20,** 1469–81 (2010).

38. Belancio, V. P., Hedges, D. J. & Deininger, P. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Res.* **18,** 343–358 (2008).

39. Burns, K. H. & Boeke, J. D. Human transposon tectonics. *Cell* **149,** 740–752 (2012).

40. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43,** 712–714 (2011).

41. Jiang, Y. *et al.* Detection of Clinically Relevant Genetic Variants in Autism Spectrum Disorder by Whole-Genome Sequencing. *Am. J. Hum. Genet.* **93,** 249–263 (2013).

42. Batista, D. A. S., Pai, G. S. & Stetten, G. Molecular analysis of a complex chromosomal rearrangement and a review of familial cases. *Am. J. Med. Genet.* **53,** 255–263 (1994).

43. Montgomery, S. B. *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* **23,** 749–61 (2013).

44. Bhangale, T. R., Rieder, M. J., Livingston, R. J. & Nickerson, D. a. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.* **14,** 59–69 (2005).

45. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10,** 551–64 (2009).

46. McVey, M. & Lee, S. E. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet.* **24,** 529–538 (2008).

47. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** 2113–2144 (2007).

48. Yu, A. M. & McVey, M. Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. *Nucleic Acids Res.* **38,** 5706–5717 (2010).

49. Roerink, S. F., van Schendel, R. & Tijsterman, M. Polymerase theta-mediated end joining of replication-associated DNA breaks in C. elegans. *Genome Res.* **24,** 954–62 (2014).

50. Lam, H. Y. K. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.* **28,** 47–55 (2010).

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

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

53. Pang, A. W. C., Migita, O., Macdonald, J. R., Feuk, L. & Scherer, S. W. Mechanisms of Formation of Structural Variation in a Fully Sequenced Human Genome. *Hum. Mutat.* **34,** 345–354 (2013).

54. Kidd, J. M. *et al.* A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143,** 837–847 (2010).

55. Kozyrev, S. V *et al.* Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. *Nat. Genet.* **40,** 211–216 (2008).

56. Romeo, G. *et al.* Hereditary thrombophilia: identification of nonsense and missense mutations in the protein C gene. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 2829–2832 (1987).

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

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

59. Hochstenbach, R., Buizer-Voskamp, J. E., Vorstman, J. A. S. & Ophoff, R. A. Genome arrays for the detection of copy number variations in idiopathic mental retardation, idiopathic generalized epilepsy and neuropsychiatric disorders: Lessons for diagnostic workflow and research. *Cytogenet. Genome Res.* **135,** 174–202 (2011).

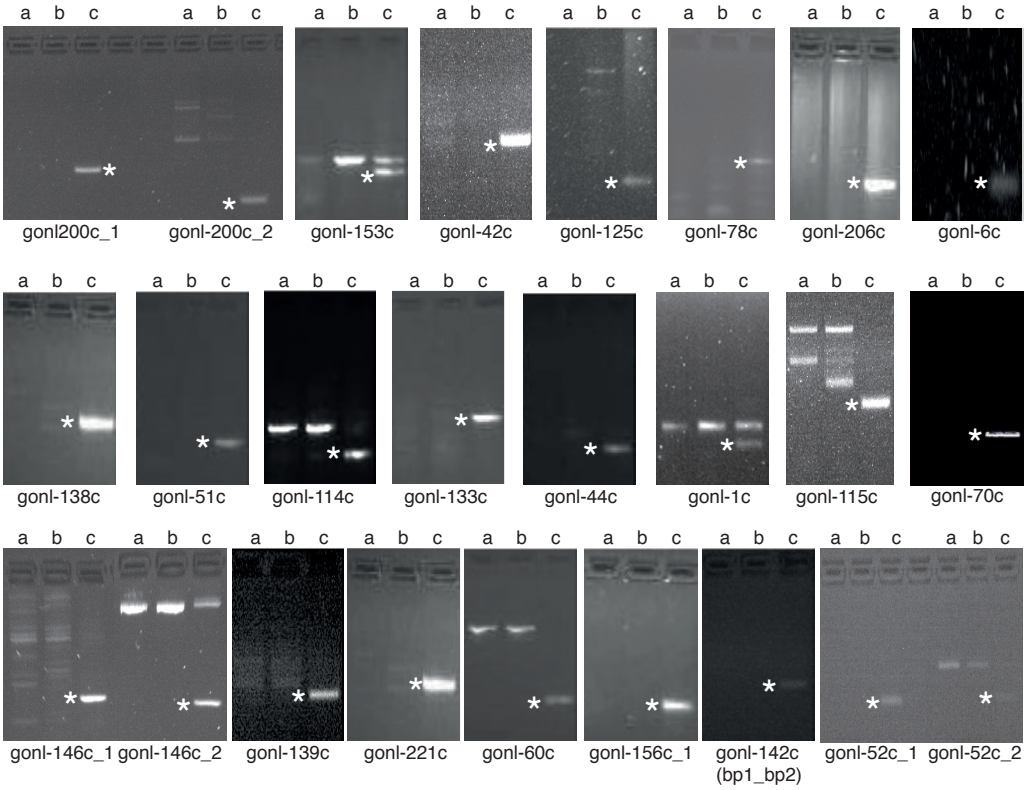
60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

61. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1,** 18 (2012).

62. The ENCODE Project Consortium. Identification and analysis of functional elements in 1\% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).

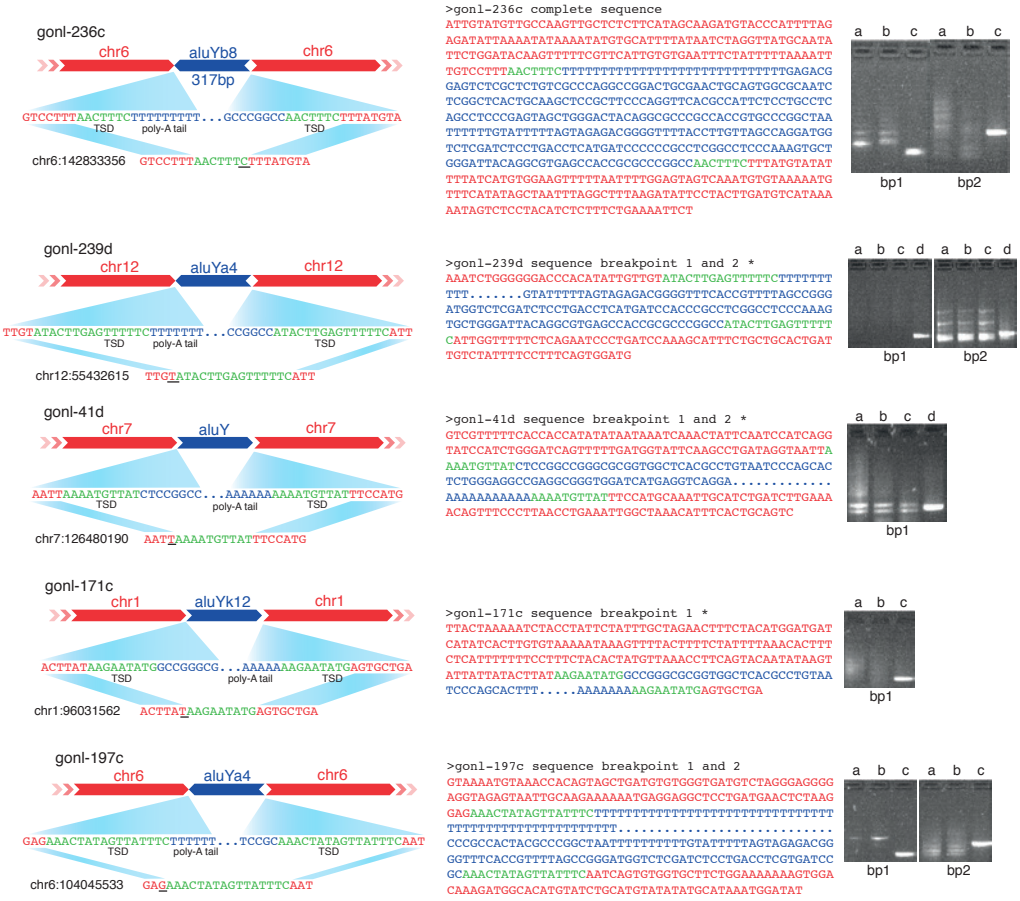
Chapter 5

Supplementary information



Supplemental Figure 1 | Overview of confirmatory PCR results for *de novo* SV breakpoints (> 100bp) identified in the GoNL dataset

Primer sets for *de novo* SV breakpoints were tested on the father (a), mother (b) and child (c). For each breakpoint a unique PCR band is visible in the child. The PCR product was sequenced by Sanger sequencing to confirm the predicted SV breakpoint (*).



Supplemental Figure 2 | Overview of five *de novo* mobile element insertions identified in the GoNL dataset

Breakpoint junction sequences were derived from Sanger sequencing or MiSeq sequencing of PCR products crossing the breakpoint junctions. For those instances where we could only obtain one of the breakpoint junctions by PCR, we used the HiSeq data for assembly of the other junction sequence based on discordant read pairs (*). TSD=target site duplication. (a) father, (b) mother, (c) child.

Variant type	Algorithm#	Approach	Reported in Francioli et al, 2014, Nat Gen 46(8):818-25	Candidate *de novo* calls	Filtered, assembled breakpoints, entered validation*	Validated as *de novo*
Indels[$]	GATK UG	Gapped reads	1.2 million	9948	1122	223
Indels[$]	GATK HC	Gapped reads		10934	1212	228
Indels[$]	Pindel	Split-reads		1557	559	137
Deletions	GATK UG	Gapped reads	27.5 thousand	138	6	1
Deletions	GATK HC	Gapped reads		317	31	4
Deletions	Pindel	Split-reads		64	14	11
Deletions	MATE-CLEVER	Split-reads, read pairs		616	85	12
Deletions	1-2-3 SV	Read-pairs		330	236	15
Deletions	Breakdancer	Read-pairs		912	29	5
Deletions	GenomeSTRIP	Read-pairs, read depth		13	13	7
Deletions	CNVnator	Read depth		202	40	0
Deletions	DWACseq	Read depth		296	53	1
Deletions	FAÇADE	Read depth		116	10	5
Tandem duplications	Pindel	Split-reads	-	3	0	0
Tandem duplications	1-2-3 SV	Read-pairs	-	62	34	8
Tandem duplications	Breakdancer	Read-pairs	-	32	7	4
Duplications	CNVnator	Read depth	-	139	18	1
Duplications	DWACseq	Read depth	-	218	24	1
Duplications	FAÇADE	Read depth	-	166	20	4
Inversions	Pindel	Split-reads	-	0	0	0
Inversions	1-2-3 SV	Read-pairs	-	27	16	0

Inversions	Breakdancer	Read-pairs	-	8	3	0
Mobile element insertions	Mobster	Split-reads, read-pairs	-	160	95	6
Interchromosomal events	1-2-3 SV	Read-pairs	-	30	17	1
Interchromosomal events	Breakdancer	Read-pairs	-	5	1	0

$ We defined indels as insertion/deletions with a size difference smaller than 21 bp between the reference and alternative alleles.

Individual variants may be discovered by multiple methods and thus can contibute to multiple values listed in this table.

* SVs were selected and validated across all 250 families; indels were only selected and validated in 92 families (11 quartets with monozygotic twin pairs, 7 quartets with dizygotic twin pairs, 74 trios)

Supplemental Table 1 | Summary on *de novo* variant calling, filtering and confirmation

Supplemental Table 2 | List of validated de novo SVs and indels

Available upon request as an Excel spreadsheet.

Child	Position	Alleles	
gonl-41d	1:199325670	Reference	AGCCAACTTTAAAAAGAAG**TA/GTATGGTA**TATAATCACAAGAAAT
		Mutation	AGCCAACTTTAAAAAGAAG**AA/G**TATAATCACAAGAAAT
gonl-125c	2:228554844	Reference	AACAAAATAACAGGAGT/**AAGTCCTTAC**TTATCAATCTAAACTCTTC
		Mutation	AACAAAATAACAGGAGT/**TTAGT**TTATCAATCTAAACTCTTC
gonl-239d	3:31723929	Reference	CTGCTCCAGCCACATGGGCT/**TCC**TGTCTCCATCCCCGTGCC
		Mutation	CTGCTCCAGCCACATGGGCT/**GGAGCCCA**TGTCTCCATCCCCGTGCC

Supplemental Table 3 | Sequences of complex *de novo* indels and outline of possible formation via SD-MMEJ

Comment	Mechanism
Templated insertion	Deletion: TAGTATGGTA Insertion: AAG Mechanism: DSB to the right of AAAGAAGTA, resection of TA on top strand (2 nt) Loop out AAG on the left and synthesis of AAGTA from bottom strand on left side of break AGCCAACTTTAAAAAGAAGAAGTA Anneal with CAT on bottom strand of other side of break ⇒ SD-MMEJ consistent, 3 nt loop out on left side, 3 nt annealing on right side
Palindrome	Deletion: AAGTCCTTAC Insertion: TTAGT Mechanism: DSB after GGAGT, resection Snapback to form hairpin with TAA/ATT stem, TCAATC in loop Synthesize TGA TCAATCTAAACTCTT ATTTGA Unwind hairpin AGTTTATCAATCTAAACTCTT Snapback to form hairpin with CT/GA stem, TTTATCAAT in loop Synthesize TTTGAG TTTATCAATCTAAACTCTT GATTTGAG Unwind hairpin GAGTTTAGTTTATCAATCTAAACTCTT Anneal with CTCA on the bottom strand of the other side of the DSB ⇒ SD-MMEJ consistent, 2 snapbacks on right, (3nt +3nt synthesis, 2nt + 6nt synthesis), 4nt annealing on left side
Palindrome	Deletion: TCC Insertion: GGAGCCCA Mechanism: DSB to the right of GGGCT, Snapback to form hairpin with AGCC/TCGG stem, ACATG in loop Synthesize GGAGC CTGCTCCAGCCACATG CGAGGTCGG Unwind hairpin CTGCTCCAGCCACATGGGCTGGAGC Snapback to form hairpin with GCT/CGA stem, GG in loop Synthesize CCATGTG CTGCTCCAGCCACATGGGCTGG TGTACCCGA Unwind hairpin CTGCTCCAGCCACATGGGCTGGAGCCCATGT Anneal to ACA on right side ⇒ SD-MMEJ consistent, 2 snapbacks on left (4 nt + 5 nt synthesis, 3 nt + 6 nt synthesis), 3 nt annealing on right side

Child	Position	Alleles	
gonl-91c	3:43390907	Reference	GATTTCCATGTAGATAGAAGA**ACT**ATAGGGCCTAGTACAGAAGG
		Mutation	GATTTCCATGTAGATAGAAGA**C**ATAGGGCCTAGTACAGAAGG
gonl-236c	3:108389073	Reference	ATATTCACAGAGGTTTACAGT/**CATCACC**ACTATCTATTTCCAGAAC
		Mutation	ATATTCACAGAGGTTTACAGT/**GAATAA**ACTATCTATTTCCAGAAC
gonl-51d	4:42806684	Reference	GATTTTTTGCTTTCAGTACTG/**AGG**TTGAGAAGTCAGATGCCAGTC
		Mutation	GATTTTTTGCTTTCAGTACTG/**C**TTGAGAAGTCAGATGCCAGTC
gonl-88c	5:73329721	Reference	AAGGAAGTTGGCCAACTCACAAAAA**AA**TTTTTTTCCAAATGGCTC
		Mutation	AAGGAAGTTGGCCAACTCACAAAAA**T**TTTTTTTCCAAATGGCTC
gonl-7c	5:138236182	Reference	ATAGGAAGCCAGGAGCAGTTA**AGGA**AGTGGGAAACAACTTGTGTT
		Mutation	ATAGGAAGCCAGGAGCAGTTA**GGG**AGTGGGAAACAACTTGTGTT
gonl-44c	7:42199409	Reference	ACCTTTTACAGGAGGTGCATG/**CAGT**ATTTTGAAAGCAAAAATTCT
		Mutation	ACCTTTTACAGGAGGTGCATG/**TAA**ATTTTGAAAGCAAAAATTCT
gonl-158c	9:105647370	Reference	CTACATGAAACCATTGTAGT/**AAGG**TGCTGTCTTGCTTATAAAGT
		Mutation	CTACATGAAACCATTGTAGT/**T**TGCTGTCTTGCTTATAAAGT
gonl-13c	11:65939317	Reference	CCTCGGCCTCCCAAAGTGCT**GG**GATTACAGGCTTGAGCCACC
		Mutation	CCTCGGCCTCCCAAAGTGCT**C**GATTACAGGCTTGAGCCACC
gonl-180c/d	14:58160893	Reference	CTTCATACCCTCTTGAAATCCTA**CCTCT**ACCTCCAAG
		Mutation	CTTCATACCCTCTTGAAATCCTA**GCTCTACCTCC**ACCTCCAAG

Supplemental Table 3 (cont.)

Comment	Mechanism
	Deletion: ACT Insertion: C ⇒ nothing SD-MMEJ consistent
Palindrome	Deletion: CATCACC Insertion: GAATAA Mechanism: DSB to the right of ACAGT Snapback to the left to form hairpin, 4 nt stem with A/A mismatch at position 2 ATATTCACAGAGGTTA ATAAGTGAC Unwinding to form: ATATTCACAGAGGTTACAGTGAATA (synthesis of A still required, mechanism unclear) ⇒ Partially SD-MMEJ consistent
	Deletion: AGG Insertion: C Mechanism: DSB to the right of TACTG Loop out at TG, synthesize CTT Unwind, anneal TT to AA on right side of break ⇒ SD-MMEJ consistent, 2 nt loop out on left, 2 nt annealing on right
Repeat context	⇒ Not SD-MMEJ consistent, looks like polymerase slippage – lose 2 A's, gain one T
Templated insertion	⇒ Not SD-MMEJ consistent, but suspicious repeats
Palindrome	Deletion: CAGT Insertion: TAA Mechanism: DSB to right of ATG Snap back on left to form a hairpin TTTTACAGGAGGTGCA AAATGT Unwinding to form TTTTACAGGAGGTGCATGTAAA Annealing with A on right side of break ⇒ SD-MMEJ consistent, 2nt snapback on left, 1 nt annealing on right (but short)
Quasi-palindrome	Deletion: AAGG Insertion: T Mechanism: DSB to right of TGTAGT, resection of AAGG on right side Loop out on right side and pairing at TGCT Synthesis of T But nothing to anneal to on left side ⇒ Not SD-MMEJ consistent, but suspicious repeats
	Deletion: GG Insertion: C ⇒ Not SD-MMEJ consistent
Templated insertion	Deletion: CCTCT Insertion: GCTCTACCTCC ⇒ Not SD-MMEJ consistent, although the ACCTCC duplication is suspicious

Child	Position	Alleles	
gonl-60c	21:15632281	Reference	TTGTTCCCACAGCCAACCATC**CT**TTTTTGCCTTGGAAAGGTGA
		Mutation	TTGTTCCCACAGCCAACCATC**A**TTTTTGCCTTGGAAAGGTGA
gonl-179c	X:5924662	Reference	CAAATGTGCTTCATTTCAAACGAGC/**T**ATTTTACCCACCTCTCTTCA
		Mutation	CAAATGTGCTTCATTTCAAACGAGC/**AC**ATTTTACCCACCTCTCTTCA

Supplemental Table 3 (cont.)

Comment	Mechanism
Repeat context	Deletion: CT Insertion: A ⇒ Not SD-MMEJ consistent, likely polymerase slippage
Palindrome	Deletion: T Insertion: AC Mechanism: DSB to the right of AAACGAGC Snap-back to form hairpin with GCT/CGA stem, TCATTTCAACG in loop Synthesis of ACATTT CAAATGTGCTTCATTTCAAACG TTTACACGA Unwinding of hairpin CAAATGTGCTTCATTTCAAACGAAGCACATTT Annealing with TAAA on right side of break ⇒ SD-MMEJ consistent, 3nt snapback on left side, 4nt annealing on right side

Mechanism	GoNL *de novo*	GoNL inherited	Mills et al. (2011)b	Pang et al. (2013)	Kidd et al. (2010)
NHR	24 (58.54%)	11797 (30.3%)	4579 (32%)	572834 (72.51%)	549 (52.09%)
NAHR	8 (19.51%)	7153 (18.4%)	1722 (12%)	2050 (0.26%)	275 (26.09%)
VNTR	3 (7.32%)	4074 (10.5%)	367 (2.6%)	213566 (27.03%)	30 (2.85%)
MEI	6 (14.63%)	15847 (40.8%)	7637 (53.4%)	1542 (0.20%)	200 (18.98%)

[a] Proportions represent the total of SVs reported in each study and have not been stratified by SV class

[b] Numbers include all MEIs reported in this study (2266 MEIs detected based on deletions ancestral allele and 5371 MEIs called separately)

Supplemental Table 4 | Formation mechanisms for large SVs identified in GoNL and other studies

Cell Type	H3K4Me3	H3KMe1	H3K27ac
Dnd41	GSM1003468	GSM1003460	GSM1003462
GM12878	GSM733708	GSM733772	GSM733771
H1-hESC	GSM733657	GSM733782	GSM733718
HMEC	GSM733712	GSM733705	GSM733660
HSMM	GSM733637	GSM733761	GSM733755
HSMMtube	GSM733738	GSM733661	GSM733666
HUVEC	GSM733673	GSM733690	GSM733691
HeLa-S3	GSM733682	GSM798322	GSM733684
HepG2	GSM733737	GSM798321	GSM733743
K562	GSM733680	GSM733692	GSM733656
Monocytes CD14+ RO01746	GSM1003536	GSM1003535	GSM1003559
NH-A	GSM733747	GSM733710	GSM733763
NHDF-Ad	GSM733650	GSM1003526	GSM733662
NHEK	GSM733720	GSM733698	GSM733674
NHLF	GSM733723	GSM733649	GSM733646
Osteoblasts	GSM1003506	GSM733704	GSM733739

Supplemental Table 5 | Histone modification datasets used for functional assessment of *de novo* variants

Detection of *de novo* variants

In order to create reliable *de novo* indel and structural variant (SV) call sets, we used a combination of 11 algorithms based on 6 approaches: (i) gapped reads, (ii) split-read, (iii) read pair, (iv) read depth, (v) combined approaches, and (vi) *de novo* assembly. *De novo* indels were defined as insertions and deletions of size smaller than 21bp and SVs encompass all events larger than 20bp.

De novo indels were called using three methods: Genome Analysis Toolkit (GATK)[1] UnifiedGenotyper, GATK HaplotypeCaller and PINDEL[2]. All confident calls made by either of the three methods were merged (exact match) and all resulting putative *de novo* indels in 92 families were subjected to experimental validation. Details of the calling and filtering for each of the three algorithms can be found in the sections below.

In order to call *de novo* SVs, we used the calls from each individual GoNL SV set [3] where the variant was predicted to be present in a single child in the entire dataset. We specifically aimed at detecting *de novo* variants with high sensitivity and therefore even *de novo* calls with marginal read support were still included. All calls were then manually inspected using IGV[4] to eliminate obvious false positives due to a miscalled parent (evidence of the SV in a small portion of the reads of one of the parents). We then took the union of all remaining calls and we merged variants which (i) were detected by multiple methods in the same child, (ii) were of the same SV type and (iii) had overlapping coordinates. We retained the most precise breakpoints for each variant based on the calling algorithm (in order: split-read, discordant read-pairs, read-depth). Local *de novo* assembly (SOAPdenovo [5]) was used for breakpoint fine-mapping and all calls were then subjected to experimental validation.

Because many of these algorithms were already used for calling segregating polymorphisms within the Genome of the Netherlands (GoNL) project, the description of the *de novo* variant calling and filtering starts with the GoNL raw or filtered calls [3]. The sections below highlight the additional calling or filtering steps undertaken in order to obtain the *de novo* candidates with each of the methods except for Breakdancer [6], CNVnator [7] and FACADE [8] for which no additional filtering was performed. The numbers of *de novo* indel and SV candidates identified by each of the tools are provided in **Supplemental Table 1**.

GATK UnifiedGenotyper

All bi-allelic autosomal GoNL indel calls [3] were re-genotyped using GATK PhaseByTransmission (PBT, manuscript in preparation), a trio-aware genotyper that reports the most likely genotypes (along with a phred-scaled confidence score) in

a trio given a mutation prior and the observed allele frequency of the site in the population. We used a relaxed mutation prior of 1 x 10^{-4} in order to increase the sensitivity of our initial call set. We note that since GATK PBT does not support sex chromosomes, only autosomal chromosomes were called using this method.

All calls for which both parents were genotyped as homozygous reference and the offspring as heterozygous were extracted and the following additional filters were applied to obtain the high confidence set:

- No evidence of non-reference allele in the parents reads
- No other GoNL sample genotyped with this non-reference allele
- At least 2 reads containing the non-reference allele found in the offspring
- The PBT confidence score was >Q30

GATK HaplotypeCaller

We used the GATK HaplotypeCaller to discover and genotype the union of regions previously called as putative indels in GoNL [3] raw indel calls as well as PINDEL calls [2] (**section 1.3**), including 1kbp flanking each variant. We filtered these calls using GATK VariantQualityScoreRecalibration (VQSR) using the following parameters:

a) Training sets
   i. Mills-Devine 1KG gold standard indel set [9]

b) Features
   i. Quality / Depth
   ii. Fisher's test on strand bias
   iii. Haplotype score
   iv. Read position rank sum test
   v. Inbreeding coefficient

We genotyped all autosomal calls passing VQSR with GATK PhaseByTransmission using a mutation prior of 10^{-4} to obtain sensitive trio-aware genotypes. All calls where both parents were genotyped as homozygous reference and the offspring as heterozygous were extracted and the following additional filters were applied to obtain the high confidence set:

- No evidence of non-reference allele in the parents reads
- No other GoNL sample genotyped with this non-reference allele
- At least 2 reads containing the non-reference allele found in the offspring
- The PBT confidence score was >Q30

PINDEL

We applied PINDEL v0.2.4t [2] on the complete GoNL alignment BAM files [3] using the default parameters. For this analysis, all chromosomes were split into bins of 20 Mb, with an overlap of 0.1 Mb, resulting in 113 genomic regions, spread over 75,000 files. All reads except for perfectly mapped ones were collected for indel/SV detection. Each variant calling job processes one of 113 genomic regions on all samples.

All variants observed in only a single child with at least 2 supporting reads and where the difference of reads supporting each allele on the forward and the reverse strand < $2\sqrt{(\#reads)}$ were kept as confident *de novo* candidates.

1-2-3 SV

From the 1-2-3 SV[10] calls, only clusters that are limited to one family and exhibiting Mendelian error (i.e. contain 4 or more discordantly mapped read pairs belonging to offspring(s), but not to parents) were considered for further analysis.

DWAC-seq

We only considered DWAC-seq (http://tools.genomes.nl/dwac-seq.html) calls where the estimated copy number change between the offspring and his/her parents was above 30%.

MATE-CLEVER

MATE-CLEVER, LASER and several auxiliary tools from the CLEVER Toolkit (CTK) v2.0-rc1-14-gad61a0d [11] were used to run the following customized pipeline:

1. Lanes/libraries are encoded in read groups in the GoNL BAM files. Insert size distributions were estimated for each read group separately.
2. CLEVER was run on each individual separately using options -A and -R to use the read-group distributions computed in (1). Deletions supported by at least 5 read pairs were retained.
3. CLEVER deletion calls were merged per family and reads from regions of +/- 750bp around these calls were extracted and mapped using LASER with parameters "-M 50000 --extra-sensitive -w 0.1".
4. For each family, a list of insertions and deletions found in at least one individual of that family by CLEVER and in at least one individual of that family by LASER were retained. CLEVER and LASER calls were considered

to be the same if their center point distance was <100bp and their length difference <20bp. Breakpoint coordinates reported by LASER were retained

5. This set of putative deletions was used to recalibrate all BAM files created by LASER (using laser-recalibrate). All deletions present in primary alignments after recalibration were extracted. Their support was summed up over all individuals in the cohort, and the resulting list was sorted by this cumulative support.
6. The merge-to-vcf program was run (with parameters "-o 100 -z 20 -O 20 -Z 5 -l 10") using this ranked list of deletion calls, all CLEVER calls and LASER BAM files of the whole cohort to compute a list of high-confidence deletion candidates.
7. Using this set of candidates, BAM files of all individuals were recalibrated again.
8. Then, all these candidates where genotyped family-wise using the genotyping utility in the CTK with the following parameters: "--min_phys_cov 5 --min_gq 10 --denovo_threshold 1e-5 --variant_prior 0.1 --mapq 30". Again, read-group-wise insert size statistics were used. Using parameter "--denovo_threshold 1e-5" ensures that only *de novo* calls are reported that are unlikely to occur in the parents ($p < 0.00001$) and are likely to occur in the child ($p > 0.99999$). In all other cases, genotypes compatible with the Mendelian laws of inheritance are reported.
9. From the returned *de novo* calls, those that occur elsewhere in the population were discarded, leading to 32 candidates across all samples (out of which 15 made part of the list of validated calls we report in this paper).

GenomeSTRiP

GenomeSTRiP [12] sites with a Mendelian violation were extracted from all GoNL deletions considering only genotypes with an associated genotype quality (GQ) higher than Q13. Only variants called in a single child and with no contributing read-pair from the parents were considered.

Mobster

Mobile element insertions (MEIs) were called using Mobster 0.1.6 with default parameters[13]. Analysis was run separately for each family. Only candidate MEIs supported by 5 or more reads were considered. Offspring-specific candidates where estimated integration site ("border5-border3 range") did not overlap any other integration site scored in other families were selected for validation.

Breakpoint mapping

De novo assembly was performed to map breakpoints of *de novo* SV candidates, using SOAPdenovo [5]. Scaffolds were aligned to the GRCh37 reference using the NCBI BLAST software [14]. Scaffolds with two high scoring segment pairs (HSPs) that are consistent with predicted structural variants were used for design of validation assays. Forward and reverse primers were placed in scaffold regions that corresponded to an HSP upstream and downstream of a predicted variant, respectively. Primers were picked using Primer3 software [15] to amplify a 300bp or bigger fragment that includes a predicted variant.

Validation

Technologies

Sanger validation

PCR primers were designed for predicted *de novo* variants (**Supplemental Table 2**) using Primer3 software. PCR was performed on families using AmpliTaq Gold (Life Technologies) and PCR reads were subjected to Sanger sequencing. For *de novo* indels, the resulting sequence was genotyped visually using Phred software [16]. For *de novo* SVs, analysis of Sanger traces was done manually using BLAST and BLAT functionalities available from the UCSC [17] and Ensembl [18] browsers.

Miseq validation

PCR primers were designed for predicted *de novo* variants (**Supplemental Table 2**) using Primer3 software. PCR was performed on families using AmpliTaq Gold (Life Technologies). PCR amplicons were pooled per family member (separate pools for PCR products from father, mother and child) and MiSeq libraries were generated based on the PCR amplicon pools using Nextera DNA sample preparation reagents (Illumina). Libraries were sequenced on MiSeq in 2x250bp mode.

For *de novo* indels, the MiSeq sequence data were aligned to the UCSC human reference genome build37 using BWA [19] and variants were genotyped using both GATK UnifiedGenotyper and GATK HaplotypeCaller. Only genotypes concordant between both methods with a genotype quality (GQ) >Q30 and at least 20x coverage in the parents and 15x in the child were considered informative.

For *de novo* SVs, reads were mapped using BWA. Breakpoints were detected by performing local *de novo* assembly of MiSeq reads followed by BLAST analysis of the assembled contig against the GRCh37 genome reference.

IonTorrent validation

A Fluidigm AccessArray was created from a list of candidate *de novo* indels and library preparation was done according to the Fluidigm protocol. These were sequenced on IonTorrent PGM and the data were de-multiplex using a Perl script using first 10 bases as one of MID set barcodes. The sequencing reads were mapped to the human genome reference build 37 using by TMAP aligner v. 3.4.1 (Life Technologies). The putative *de novo* indels were genotyped using the GATK UnifiedGenotyper and GATK HaplotypeCaller. Only calls concordant between both methods with a genotype quality (GQ) >Q30 and at least 20x coverage in the parents and 15x in the child were considered informative.

De novo indel results

We assayed 1,176 candidate *de novo* indel sites in 99 children from 92 families (11 quartets with monozygotic twin pairs, 7 quartets with dizygotic twin pairs and 74 trios). In total 917 were successfully re-sequenced and 284 were confirmed as *de novo*. All 148 sites (31 *de novo*) successfully re-sequenced in families with monozygotic twin pairs had concordant results between the two twins. We used 3 technologies (**section 3.1**) for these assays:

41 sites were assayed using both Sanger and MiSeq. 29 sites were successfully assayed using both technologies and 17 were confirmed *de novo*. All calls were concordant between the two technologies. Another 8 sites were successfully assayed on one technology and 3 were confirmed *de novo*.

1,031 sites were assayed using MiSeq. 746 sites produced enough sequence to be informative and 236 were confirmed *de novo*.

104 sites were assayed using IonTorrent. 85 sites produced enough sequence to be informative and 28 were confirmed *de novo*.

De novo SV results

The subsections below give the details of all validation experiments we undertook. **Supplemental Figures 1 and 2** show the supporting evidence for all 41 validated *de novo* SVs. In general, we used the union of all putative *de novo* calls by the different methods. Whenever an event was reported by multiple

methods, we designed the validation assay based on the coordinates of the most precise breakpoints by giving priority to methods using split-read, then discordant read-pairs and finally read-depth.

Pilot validation of split-reads and discordant reads predictions

For an initial validation experiment we selected 39 candidate SVs predicted by 1-2-3 SV, BreakDancer, Pindel and GenomeSTRiP. Although the majority of candidates were inherited (i.e. false negatives in the parents), three candidates were validated as true *de novo* events, confirming the presence of *de novo* variants in our callsets.

Large SVs / discordant pair and read depth approaches

We checked 70 calls, predicted by discordant pair (1-2-3 SV, BreakDancer), read density (DWAC-Seq, CNVnator) or combined (GenomeSTRiP) approaches. Large (>1kbp) discordant pair-based candidates were cross-checked for showing expected read density profiles. We validated another 12 events as *de novo*, including 11 sized above 1kbp.

Large SVs / read depth-based methods

In an attempt to verify other large events we selected calls (n=69) exceeding 600 bases, which were called using at least one of the read-depth approaches (FACADE, DWAC-Seq, CNVnator), but not found by other methods. Inspection of the target regions using IGV browser [4] (IGV_check) or DWAC-Seq read depth profiles (RD_check) showed that these mostly represent inherited events where RD methods failed to make a call in one of the parents. None of the variants with only read-depth support were validated as *de novo*.

Short SVs produced by MATE-CLEVER

We used the *de novo* calls from MATE-CLEVER described in an earlier publication [3]. A total of 63 calls were tested by PCR/MiSeq sequencing, and only one additional 29 bp deletion validated as *de novo*. We analyzed the reasons for this low validation rate and found that slightly different insert size distributions of different libraries lead to these false positive calls. We adapted the MATE-CLEVER pipeline to be lane/library-aware and re-ran it as described in **Section 1.6**. The results of that run of MATE-CLEVER were included in the next round of validation (**Section 3.3.5**).

Large scale validation

We validated the remaining three hundred candidates (excluding candidates supported only by BreakDancer, which contributed over 9k *de novo* candidates). Another 20 *de novo* events were confirmed in these validation rounds.

Mobile element insertions validation

We identified 95 mobile element insertion (MEI) candidates, including 61 LINEs, 30 SINEs, 3 SVA, 1 HERV and attempted to validate them all. Oligonucleotide primers were designed for 5'- and 3'- breakpoints based on homologous sequences shared by discordant pairs belonging to regions upstream and downstream of a candidate integration site, respectively. All primers were PCR-amplified and sequenced using Sanger/Miseq. We could confirm 6 MEIs as *de novo*, all SINEs (**Supplemental Figure 2**).

Indel classification

We used the classification proposed by Montgomery and al. [20] in order to classify our indels, except for (i) Predicted Hotspots (PR) that we did not use since they were not readily available and (ii) Complex indels that are new in our data. The indel classes we used were thus defined as follows:

- HR: Indels in homopolymer runs of at least 6 identical nucleotides.
- TR: Indel in tandem repeats of at least
    - 9 nucleotides for repeat unit of 2 nucleotides
    - 12 nucleotides for repeat unit of 3 nucleotides
    - 14 nucleotides for repeat unit of 4 nucleotides
    - 15 nucleotides for repeat unit of 5 nucleotides
    - 17 nucleotides for repeat unit of 6 nucleotides
    - 19 nucleotides for repeat unit ≥7 nucleotides
- NR, CCC: Indels in non-repeat regions (not HR nor TR) but inducing a local change in copy count.
- NR, non-CCC: Indels in non-repeat regions (not HR nor TR) which do not introduce a local change in copy count
  Complex: Indels substituting multiple bases for others

Homopolymer runs were annotated using the GATK VariantAnnotator. Tandem repeats were found using Sciroko [21] (for TR of unit size <7bp) and Tandem Repeat Finder [22] (for TR of larger unit size). We required perfect sequence identity for

tandem repeats of unit size <7bp and 90% sequence identity for tandem repeats of larger unit size.

Formation mechanism

We assumed that all indels that are located in repeat regions or that lead to a copy count chance (HR, TR, NR CCC), have arisen by polymerase slippage.

For simple non copy count changing indels (NR non-CCC), we expect most of these to emerge through non-homologous end-joining (NHEJ) given the short size of these events and the breakpoints presenting little or no microhomology and therefore assigned NHEJ as a possible mechanism for all of these [23]. In addition to NHEJ, we also looked whether some:

- contained larger microhomologies (>3bp), possible signatures of replication slippage or microhomology-mediated end joining (MMEJ) [23–25]
- were located between palindromes, possibly indicating that the formation of a secondary structure played a role in their formation. We only looked at perfect palindromes of at least 6bp as these were previously shown to be overrepresented around indels when compared to the genomic background [20].

For complex indels, we noticed that many of the derived alleles were forming palindromes with neighboring sequence. We used a minimum size of 6bp for palindromes, consistent with palindromic annotations of the context around simple NR non-CCC indels. Such complex indels were previously described in cancer studies and template switching was proposed as a possible mechanism for their emergence [26,27]. We also annotated complex indels where the inserted sequence was partially or entirely templated from neighboring sequence. Such replacements with a templated sequence were observed in *C. elegans* and theta-mediated end joining (TMEJ) was proposed as a mechanism [28].

Summary counts of mechanisms can be found in **Table 1**.

Frequency of indel categories

Using the definitions of homopolymer runs (HR) and tandem repeats (TR) as described above, we computed the total number of reference bases spanned by such runs in the GoNL accessible genome [3] in order to estimate the genomic background. We used a proportion test to assess whether the proportion of *de novo* indels in TR and HR were compatible with their relative proportion in the reference genome.

While our results show very clear indel enrichment in these regions, it is likely that these numbers are conservative since the discovery and genotyping of indels in such repetitive regions of the genome in short-read sequencing data is more challenging than in non-repetitive regions. In our data, this is partly reflected by elevated proportions of inconclusive validation assays (27.9% vs 18.8%, chisq p = 0.0016) and of false positives (82.4% vs 63.1%, chisq p = 8.0×10^{-7}) in repetitive vs non-repetitive regions, respectively.

When looking at complex indels, we noticed that a 5/14 of the derived alleles formed palindromic repeats with neighboring sequence whereas this was observed for only 1 of the 70 insertions in our set. We used a chi square test to assess whether this difference was significant.

Comparison against inherited indels

To study the difference between *de novo* and inherited indels, we compared the *de novo* indels observed against the polymorphic indels discovered in the GoNL dataset [3]. Because the algorithms used to find GoNL indels could not identify complex indels, we ran the GATK HaplotypeCaller using the --mergeVariantsViaLD option on the GoNL data. The resulting set of indels (union of previously identified indels and complex indels) were used as comparison. We note that using this strategy on the complex *de novo* indels, the HaplotypeCaller correctly identified 12/14 (86%) of these as complex events. The inherited indels in the GoNL data were annotated with respect to their class and mechanisms using the same pipeline as for the *de novo* ones.

References

1. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).
2. 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).
3. Francioli, L. C. *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46,** 818–825 (2014).
4. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–92 (2013).
5. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1,** 18 (2012).
6. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6,** 677–81 (2009).
7. 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).
8. Coe, B. P., Chari, R., MacAulay, C. & Lam, W. L. FACADE: a fast and sensitive algorithm for the segmentation and calling of high resolution array CGH data. *Nucleic Acids Res.* **38,** e157 (2010).
9. Mills, R., Luttig, C. & Larkins, C. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome …* **16,** 1182–1190 (2006).
10. Kloosterman, W. P. *et al.* Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum. Mol. Genet.* **20,** 1916–24 (2011).
11. Marschall, T., Hajirasouliha, I. & Schönhuth, A. MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics* **29,** 3143–3150 (2013).
12. Handsaker, R. E., Korn, J. M., Nemesh, J. & Mccarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43,** 269–276 (2011).
13. Thung, D. T. *et al.* Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biol.* **15,** 488 (2014).
14. Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* **12,** 656–664 (2002).
15. Rozen, S. & Skaletsky, H. Primers3 on the WWW for General Users and for Biologist Programmers. *Methods Mol. Biol.* **132,** 265–386 (2000).
16. Ewing, B., Hillier, L., Wendl, M. & Green, P. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome Res.* 175–185 (1998). doi:10.1101/gr.8.3.175

17. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

18. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41,** D48–55 (2013).

19. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

20. Montgomery, S. B. *et al.* deletion variants identified in 179 human genomes The origin , evolution , and functional impact of short insertion – deletion variants identified in 179 human genomes. 749–761 (2013). doi:10.1101/gr.148718.112

21. Kofler, R., Schlötterer, C. & Lelley, T. SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics* **23,** 1683–5 (2007).

22. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).

23. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. **10,** (2009).

24. Albertini, M., Hofer, M., Calos, M. P. & Miller, J. H. On the formation of Spontaneous Deletions: The Importance of Short Sequence Homologies in the Generation of Large Deletions. *Cell* **29,** 319–328 (1982).

25. Mcvey, M. & Lee, S. E. MMEJ repair of double-strand breaks (director ' s cut): deleted sequences and alternative endings. (2008). doi:10.1016/j.tig.2008.08.007

26. Ripley, L. S. Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. **79,** 4128–4132 (1982).

27. Ripley, L. S. Frameshif Mutation: Determinants of Specificity. *Annu. Rev. Genet.* **24,** 189–213 (1990).

28. Roerink, S. F., Schendel, R. Van & Tijsterman, M. Polymerase theta-mediated end joining of replication-associated DNA breaks in C . elegans. (2014). doi:10.1101/gr.170431.113

Chapter 6

Summary and discussion

Summary

In this thesis, I have analyzed Next Generation Sequencing (NGS) data collected from 250 Dutch families to explore inherited and *de novo* variation in their genomes. Thanks to fantastic collaborations, the Genome of the Netherlands (GoNL) Consortium has created one of the most extensive catalogues of human genetic variation. Indeed, the GoNL project has added 8 million single nucleotide variants (SNVs), 650,000 short insertions and deletions (indels), and 27,000 structural variants (SVs) to databases and produced the largest set of observed *de novo* variants to date comprising 11,000 SNVs, 284 indels and 41 SVs.

Chapter 2 focused mainly on inherited genetic variation in the Dutch population. Thanks to the intermediate sequencing coverage, the familial design, and the use of state-of-the-art variant detection tools, it was possible to detect variation across the entire size and allele frequency spectrum. Of particular interest, mid-size deletions (50 – 500 bp) could be described for the first time at the level of a population.

By leveraging the parent-offspring design of the GoNL project, SNVs and indels could be phased accurately, significantly improving the downstream imputation accuracy in samples genotyped using SNP arrays, when compared to other existing panels such as the 1000 Genomes Phase 1 haplotypes [1]. It is, however, the combination of multiple haplotype panels that will give the best downstream imputation accuracy, warranting collaborative efforts such as the Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org/) that aims to integrate existing haplotype panels into a single imputation resource.

Chapter 2 also explored population genetic analyses in the Dutch population, which benefitted from genome-wide coverage of both common and rare alleles and from accurate haplotype phasing. A fine-scale structure across the country consistent with multiple ancient migrations was unveiled, possibly reflecting population movements due to the frequent changes in sea levels and flooding that affected the Netherlands in the last millennia.

The parent-offspring design of the GoNL project provided a unique opportunity to investigate *de novo* variation on an unprecedented scale. Identifying *de* novo mutations in NGS data is very challenging as such events are rare [2] and NGS data suffers from relatively high error rates and fluctuating coverage. In **Chapter 3**, I described the tool called PhaseByTransmission, which I developed for the detection of *de novo* SNVs and indels, and showed that it outperforms current methods using simulated data. **Chapter 2** described its use for the detection of *de novo* SNVs on the autosomes and introduced a machine-learning filtering

algorithm to further increase the detection accuracy. In **Chapter 5**, I used PhaseByTransmission to identify *de novo* indels.

In **Chapters 2 and 4,** I studied the effect of parental age on *de novo* SNVs. In addition to replicating the paternal age effect on the number of *de novo* SNVs previously described [3,4] (**Chapter 2**), I showed in **Chapter 4** that paternal age also influences their chromosomal distribution, which is a novel finding. Notably, mutations in offspring of older fathers are enriched in earlier replicating, genic regions compared to those of younger fathers.

Chapter 4 further described the properties and patterns of *de novo* SNVs in the general population and evaluated the factors influencing their distribution, including sequence context, functional and epigenomic properties. Results showed that elevated mutation rates in functional regions of the genome previously described [4] can be explained by local sequence context and possibly indicate purifying selection in these regions [5,6]. Furthermore, the spectrum of mutations in transcribed regions exhibits a strong asymmetry, compatible with the action of transcription-coupled repair [7]. Beyond these effects, I observed that mutations have a tendency to cluster within the individual. These mutation clusters represent 1.5% of the *de novo* SNVs and have a unique mutational spectrum with reduced transitions and highly elevated C>G transversions. Altogether, these observations provide novel insights into the mechanisms of *de novo* SNVs.

Chapter 4 also revisited long-standing hypotheses on mutation properties inferred from comparative genomics and population genetics. By using observed *de novo* mutations, it was possible to study the relative contributions of mutations and population genetics factors in these models. Notably, I showed that the possible mutagenic effect of recombination does not explain the enrichment of mutations around recombination hotspots previously described [8,9].

In **Chapter 5**, I characterized *de novo* indels and SVs and showed the remarkable complexity of these events ranging from small changes to large interchromosomal rearrangements. In particular, despite being mostly absent from genetic variation repositories, complex indels represent almost 5% of all observed *de novo* indels in GoNL suggesting that this type of variation is relatively common and has been understudied.

The relative impact of *de novo* SVs on the genome exceeds that of *de novo* SNVs; they affect about 55 times more genomic and protein coding bases on average. This number is in contrast with the relative difference between inherited SVs and SNVs, which affect similar numbers of bases, suggesting strong purifying selection for novel structural variants.

The parental distribution of *de novo* indels and SVs are such that both of these variant types are significantly elevated in the paternal germline. Interestingly, differences in mutation rate across families were observed, suggesting underlying environmental effects or varying familial susceptibility to *de novo* structural changes.

Overall, this thesis presents key results of large-scale NGS data collected within a single population. The parent-offspring design has proved beneficial, allowing in-depth characterization of common, rare and *de novo* SNVs, indels and SVs. Analyses of these data have revealed novel insights into inherited and *de novo* variation in human genomes.

Challenges in NGS data analysis: lessons learned

While NGS offers unprecedented possibilities for exploring individual genomes, it is not without technical issues and limitations. In this thesis, I have been confronted with these shortcomings when analyzing the Genome of the Netherlands (GoNL) project data. In the sections below, I will review some of the challenges inherent to NGS data analysis and illustrate them with specific examples.

First, I will discuss the limitations of NGS technologies for capturing repetitive and complex regions of the genome. As a result, a non-trivial portion of the genome is currently not "accessible", leading to possible problems and biases with the interpretation of the variation in these regions.

Second, I will present some of the challenges in obtaining a comprehensive catalogue of genetic variation, and describe how different technologies, different tools and different versions of tools can influence genetic variation detection. I will also examine the methodology used in the GoNL project to detect and integrate genetic variation ranging from single nucleotide variants (SNVs) to short insertions and deletions (indels) and large structural variants (SVs).

Third, I will examine the meaning of the genotypes inferred from NGS technologies and the factors responsible for biases in genotype quality. Because inferred genotype qualities can vary throughout the genome, downstream analyses require careful handling of this information to avoid spurious results.

Accessible genome

When sequencing a human genome using NGS technologies, not all bases of the genome are equally represented. On the one hand, this is due to the sequencing of random fragments; on the other hand, due to systematic biases that depend on

the regional sequence content [10–12]. In addition, when aligning the sequence data to the reference genome, not all regions are as easy to align to. Some regions of the genome are very similar to each other (due to homology) and therefore difficult to differentiate unambiguously.

As a result, the power to detect and genotype variants varies across the genome and is limited in certain regions. At the extreme, certain regions of the genome are simply "inaccessible" to current sequencing technologies, meaning that the sequence generated or mapped to these regions is too unreliable to accurately detect and genotype variants. In fact, 6.6% of the human reference genome is currently filled by so-called 'N' bases, indicating that the exact sequence content of these regions is unknown. These sequence gaps have yet to be filled and completed, probably with the aid of more sophisticated longer-read technologies [13].

In **Chapter 2**, I computed the regions of the genome that are "accessible" in our project, using an approach pioneered by the 1000 Genomes Project [14]. Thanks to the large number of samples in the GoNL project, random coverage fluctuations should average out when considering all samples together whereas systematic ones should add up. For each base, I therefore computed the total coverage across all samples and marked as inaccessible all bases where the total coverage deviated significantly from the mean. In addition, regions where more than 20% of the reads could not be mapped unambiguously were also marked as inaccessible. On top of the 6.6% of the reference genome lacking sequence content, I found that 2.9% of the genome had systematically low coverage, 0.4% systematically high coverage, and 2.3% could not be mapped unambiguously.

Although abnormal, some of these regions have overlapping reads and therefore could potentially be used for variant detection and genotyping. In fact these regions are often processed and analyzed in the same way as others despite their lower quality.

In GoNL, as many as 677,000 SNVs were detected in the inaccessible genome. When comparing the quality of the variants in inaccessible regions to other variants, I found that they had significantly lower power for their detection while exhibiting a higher proportion of novel variants (Fig. 1). This likely indicates higher level of spurious variants in these regions and warrants caution when using them for analyses.

With regard to the content of the inaccessible genome, it comprises the telomeres and centromeres of all chromosomes and is vastly enriched in repeat sequences such as microsatellites and mobile elements. The part of genome we are currently able to analyze is therefore biased towards unique regions of the genome, impacting the study of the more repetitive regions.

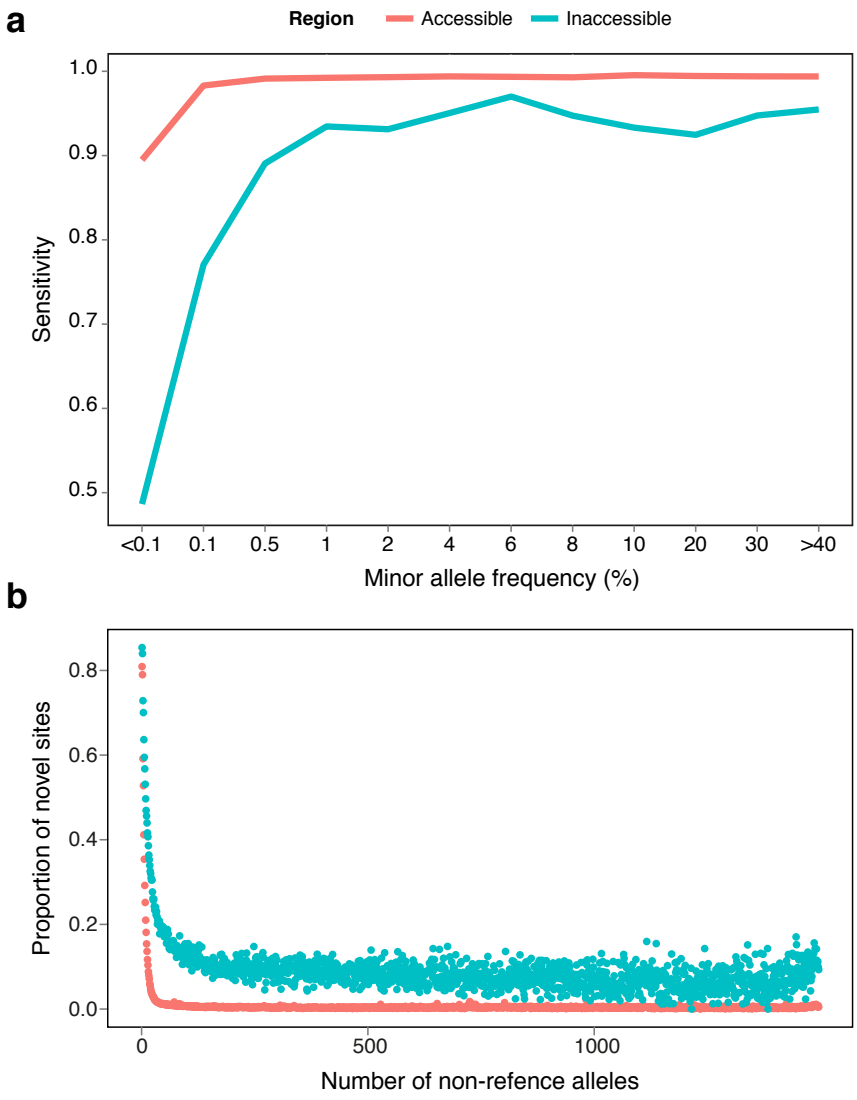


Figure 1 | Quality assessment of the accessible genome

a. The power to detect SNVs (y-axis) as a function of minor allele frequency (x-axis). Only SNVs present on the Immunochip array are considered (excluding the MHC region), because all GoNL samples were genotyped with this array. The results are stratified by accessible and inaccessible regions of the genome. The power to detect SNVs in the inaccessible genome is considerably lower than the accessible genome, especially for rare variation.

b. The proportion of novel sites (based on dbSNP 137) by allele count and stratified by accessible and inaccessible regions. Inaccessible regions have a higher proportion of novel sites at all allele frequencies, likely indicating a higher number of spurious variants.

For example, *de novo* mutations (as described in **Chapters 4 and 5)** cannot reliably be evaluated in these regions. However, *de novo* rates are likely elevated in inaccessible regions as most variant types show elevated mutations in repeat regions [15–18]. The mutation rates currently computed for these regions are therefore likely underestimates, but better data will be required to resolve this uncertainty unequivocally.

Characterizing genetic variation

Next-generation technologies (NGS) encompass many different technologies, such as Roche454, Illumina HiSeq, Illumina MiSeq, Life Technologies SOLiD4, Complete Genomics, or Pacific Biosciences SMRT. Although all these platforms sequence DNA fragments in a massively parallel way, they use distinctive processes that lead to very different error rates and biases in their output. In addition, the available downstream analysis tools differ per technology. As a consequence, the genetic variation detected using these different technologies typically varies substantially [10,12,19,20].

As part of the GoNL project, 20 unrelated samples were sequenced using both Illumina HiSeq2000 and Complete Genomics technologies. For the work described in **Chapter 2**, I compared the single nucleotide variants (SNVs) detected by these technologies in these samples, and found that about 92% of the variants identified by Complete Genomics were also detected using Illumina HiSeq2000, whereas Complete Genomics detected only about 70% of the variants identified using Illumina HiSeq2000. These numbers are in line with previous reports [19] and illustrate the difficulty of comparing results between platforms. The genotypes inferred on sites detected by both technologies were mostly concordant (99.9%), showing that in spite of the difference in detection, shared sites can robustly be genotyped.

Beyond intrinsic differences between sequencing technologies, analysis of the generated NGS data will depend on an automated pipeline composed of multiple bioinformatics tools. A growing arsenal of tools exists for each of the required steps in the pipeline and the genetic variation detected depends on the tools (and version of tools) used [20–22]. Moreover, because of the complexity and variability in the types of genetic variation, no single tool can identify all variation comprehensively.

Genetic investigators and diagnostic centers are thus faced with difficult choices in the implementation of their variant detection pipelines. First, while in principle running more methods and combining results should lead to more accurate

results, it can come at enormous costs in computing infrastructure and time. Second, tools and technologies change very rapidly; for example the widely used Genome Analysis Toolkit (GATK) [23] releases about 10 updates a year. Although each new version represents an improvement over the previous, comparing results produced by different versions might prove challenging as their sensitivity and specificity may vary.

The following sections illustrate these challenges with examples from the GoNL project, which used a large number of tools in order to catalogue the genetic variation in the Dutch population.

Single Nucleotide Variants (SNVs)

I used a single tool for the detection of SNVs in GoNL: the Genome Analysis toolkit (GATK) UnifiedGenotyper [23,24]. To assess the reproducibility of the GoNL SNV callset with respect to tools updates, I investigated the difference in detection between the results produced by the version of tools used for the GoNL SNV release 5 (BWA v0.5.9, GATK v1.4) and the same tools a year later (BWA v0.7.4, GATK v2.7) on a single parent-offspring trio.

Although 98% of the variants present in the GoNL SNV release 5 were also identified with the newer version of GATK, only 90% of the variants called by the newer version of GATK were part of the GoNL SNV release 5. The genotype concordance between the two versions of the tools on shared variants was however very high (99.8%).

While the newer version possibly improves results, the relatively large difference in the number of variants detected between the two versions of the pipeline can pose problems when comparing downstream data sets. These results illustrate the rapid evolution of tools for NGS data analysis and highlight the difficulty in comparing variants from NGS data even when pipelines only differ in the version of the tools they use.

Short insertions and deletions (indels)

Indels in GoNL were defined as events inserting or deleting 1 to 20 bp. Because of their size spectrum, three different methods were used for indel calling present in the release of the GoNL data described in **Chapter 2**: GATK UnifiedGenotyper, PINDEL [25] and MATE-CLEVER [26]. These methods leverage different signals from the data and thus are complementary in the size of variant they capture and in their error modes. PINDEL captures indels across their entire size spectrum,

whereas GATK UnifiedGenotyper is most sensitive to shorter indels (up to ~10 bp) and MATE-CLEVER to longer indels (>10 bp). In addition to these algorithms, the GATK HaplotypeCaller was subsequently run on all indel calls made by these three methods. All these algorithms were run and their results filtered according to their documentation.

Considering only indels detected by multiple methods has been previously shown to result in dramatically increased specificity [20]. I therefore filtered all indels detected only by a single tool. This approach however poses several challenges: first, in repetitive regions the same indel can be represented with different coordinates by the different methods since the exact location of the inserted or removed sequence within the repeat is ambiguous. To produce an unambiguous representation, I realigned all indels to their most upstream possible location. Second, PINDEL and MATE-CLEVER only report the presence or absence of an indel in a sample whereas GATK Unified Genotyper and GATK HaplotypeCaller infer a genotype for each of the samples. I therefore used a priority scheme for assigning genotypes using the methods reporting the most precise genotype to the least precise genotype: GATK HaplotypeCaller, GATK UnifiedGenotyper, PINDEL, MATE-CLEVER.

Beyond the technical considerations in merging indel data sets, it is interesting to compare the relative performance of the indel calling methods (Fig. 2). From the 2.1 million indels called across the different methods, 76.4% were called by at least two methods, 53.5% by three methods and 0.2% of the events were detected by all methods. The GATK UnifiedGenotyper, GATK HaplotypeCaller and PINDEL all display good sensitivity for this size range, whereas MATE-CLEVER appears to be better suited for larger variants.

Structural Variants (SVs)

The detection of SVs is an even more challenging process than indels since variants size ranges from tens to millions of bases. Their detection is based on different signals and patterns from the sequencing data to capture the full size spectrum. Therefore, we ended up using 10 tools for the GoNL project to combine methods based on mapped reads (GATK UnifiedGenotyper), split-read (PINDEL, MATE-CLEVER), read-depth (CNVnator [27], DWACSeq (http://tools.genomes.nl/dwac-seq.html), Facade [28], GenomeSTRiP [29]), read-pairs (123SV (http://tools.genomes.nl/123sv.html), Breakdancer [30], GenomeSTRiP, MATE-CLEVER) and *de novo* assembly (SOAPdenovo [31]). The resulting calls were filtered according the tools documented best practices.

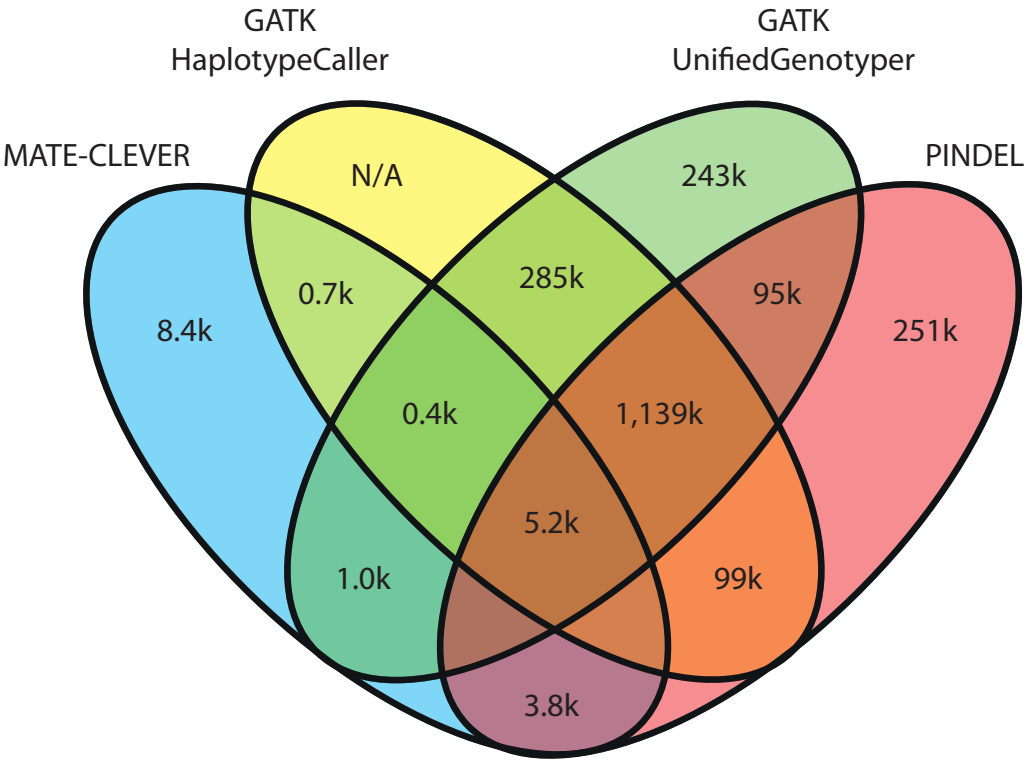


Figure 2 | Overlap of indel calling algorithms

This Venn diagram depicts the overlap in short indel calls (≤ 20bp) between the four methods used on the GoNL data. Note that the GATK HaplotypeCaller was only used on regions were indels were previously detected by one of the other methods and thus number of calls specific to the GATK HaplotypeCaller are not provided.

Each of these tools reports different types of SVs (e.g. deletions, duplications, inversions, etc.), with different precision (precise breakpoints vs. approximate genomic area) and different genotyping precision (no genotypes, presence/absence in each sample, genotypes for each sample). Because deletions were reported by all methods used, the initial GoNL SV release described in **Chapter 2** included only these events. Since tools are sensitive over a different size ranges, the deletions were split by size: 21-100 bp and >100 bp.

In total, 83,000 deletions of 20-100 bp in size were detected by four algorithms: GATK UnifiedGenotyper, PINDEL, MATE-CLEVER, SOAPdenovo. Similarly to indels, all deletions identified by these methods were realigned to produce unambiguous genomic coordinates and all deletions detected by a single tool (62,000) were filtered out. To further improve specificity, only deletions seen in at least 3 families and transmitted to at least 1 child were kept, leaving a final set of 18,500 deletions. PINDEL and MATE-CLEVER were the most sensitive methods for this size range as all deletions were detected by at least one of these tools and 62% were not identified by any other tools. As a consequence, genotype information is missing for most of these deletions since neither PINDEL nor MATE-CLEVER produces this information.

Larger deletions (>100 bp) were detected using 9 tools (123SV, Breakdancer, CNVnator, DWACSeq, Facade, GenomeSTRiP, MATE-CLEVER, PINDEL, SOAPdenovo). For deletions of this size, many algorithms do not produce precise breakpoints or coordinates. Assembling all events into a unified set thus required allowing for imprecise coordinates and deletions produced by different tools that had at least 80% reciprocal overlap were considered as representing the same event. Since coordinates still need to be assigned to each of the deletions, a priority scheme was used for selecting coordinates based on the precision of the signal used by the method for calling the event: split-read (PINDEL), Read pair / split read (MATE - CLEVER), *de novo* assembly (SOAPdenovo), read-pair (1-2-3-SV, Breakdancer), read-depth / read-pair (GenomeSTRiP), read-depth (DWACseq, CNVnator, FACADE). Of these tools, GenomeSTRiP is the only method to report inferred genotypes.

In total, 257,000 large deletions were called across the 9 tools after filtering, but this number is dominated by Breakdancer, which contributed a total of 211,000 deletions while other tools only contributed 8,000 to 25,000 calls. Ignoring calls unique to Breakdancer (192,000 deletions), 35% of the deletions were detected by more than one method. To further improve the specificity, deletions that were not present in at least three families and transmitted to at least one child (9,947 deletions filtered out) or not called using at least two different types of signal (e.g.

read depth and split read; 4,061) were also filtered out. The final set of deletions larger than 100 bp thus comprised 9,187 deletions.

Looking at the contribution of the different tools to the final set, methods using discordant read-pairs were the most sensitive (123SV, Breakdancer, GenomeSTRiP, MATE-CLEVER), each detecting about 80% of the events, followed by split-read (PINDEL, 60% of the deletions detected), *de novo* assembly (SOAPdenovo, 32% of the deletions detected) and read-depth methods (DWACSeq, CNVnator. Facade, 3% to 14% of the deletions detected).

Combining the deletions of all sizes, the GoNL project contributed 27,000 deletions, of which 93.3% were novel when compared with 1000 Genomes Phase 1 (80% reciprocal overlap). With an overall validation rate of 96.5% based on PCR and Sanger resequencing of 144 events, we can be confident to have produced a high-quality set of SVs. Although the sensitivity was not assessed, it is likely that this set is conservative given the stringent criteria used.

The detection of SVs currently requires the application, filtering and merging of many different tools. This certainly hampers the efforts for the global characterization of such events and highlights the need for novel tools combining the different signals in NGS data to produce robust calls. The high novelty rate of the variants detected in GoNL is likely a reflection of how much SV genome diversity is still to be discovered.

Interpreting genotypes from NGS data

Sequencing and genotyping has traditionally being achieved using Sanger sequencing [32], a method that is nowadays so well tuned that it achieves error rates in the order of 10^{-5} per-base [33,34]. During the past decade, the genetics community has heavily relied on SNP arrays to genotype samples at known polymorphic positions [35]. Although the genotypes reported by these arrays are not exempt from problems and errors, quality standards and statistical methods have been developed to overcome these shortcomings and produce robust results [36,37].

Due to the nature of sequencing, the quality of the genotypes inferred from NGS data is intrinsically tied to the evidence from the reads overlapping each genomic position. In particular, the number of reads, the quality of the mapping of each of these reads and the quality of the base overlapping the position should all be taken into account [24,38].

All of these quantities vary from one position to the next, influenced both by the random nature of NGS and by systematic factors such as the library preparation

protocol, sequencing technology used, the sequence content or the complexity of the region [11,12,24,39]. Moreover, when preforming targeted sequencing (e.g. whole-exome sequencing), the enrichment kit used introduces additional biases in the coverage of the different regions targeted [40,41]. Targeted sequencing also produces more variance in coverage than whole-genome sequencing, leading to larger differences in the quality of the genotype produced across the regions sequenced.

Most state-of-the-art NGS genotyping protocols aggregate the information across all reads, taking mapping and base qualities into account, to produce genotype likelihoods for all possible genotypes (given the alleles present in the data) at each position. Genotype likelihoods convey the statistical uncertainty of the different genotypes and as such provide the best information for downstream analyses. Figure 3 illustrates the dependency of genotype likelihoods with respect to the proportion of reference and non-reference alleles observed.

Genotype likelihoods are, however, not supported by all downstream methods and can sometimes be difficult to interpret [42]. For this reason, analyses often rely solely on the most likely genotype even though the quantification of the uncertainty for each genotype is lost. In order to reduce noise from genotyping errors, only genotypes over a certain confidence threshold are usually considered. Figure 3 illustrates the high and low confidence zones for a threshold of 95% confidence in a genotype.

The most common underlying reasons for a miscalled or low-confidence genotype depend on whether the true genotype is homozygous or heterozygous: most problems with homozygous genotypes stem from sequencing errors whereas most problems with heterozygous genotypes come from the sampling of the two alleles (especially at lower coverage). Because of these different error modes, homozygous genotypes have on average higher qualities given the same evidence in the sequencing data. This leads to a bias towards homozygous genotypes when restricting analyses to high confidence genotypes only, especially in regions with coverage below ~20x [42].

Depending on the type of analysis, the usage of genotypes from NGS data thus needs careful consideration in order to maximize usability and minimize error rates and genotyping biases. In addition, there is a need for analyses tools capable of considering and propagating the uncertainties inherent to NGS data genotypes to avoid biases related to the use of (sometimes arbitrary) thresholds and ad hoc filtering steps.

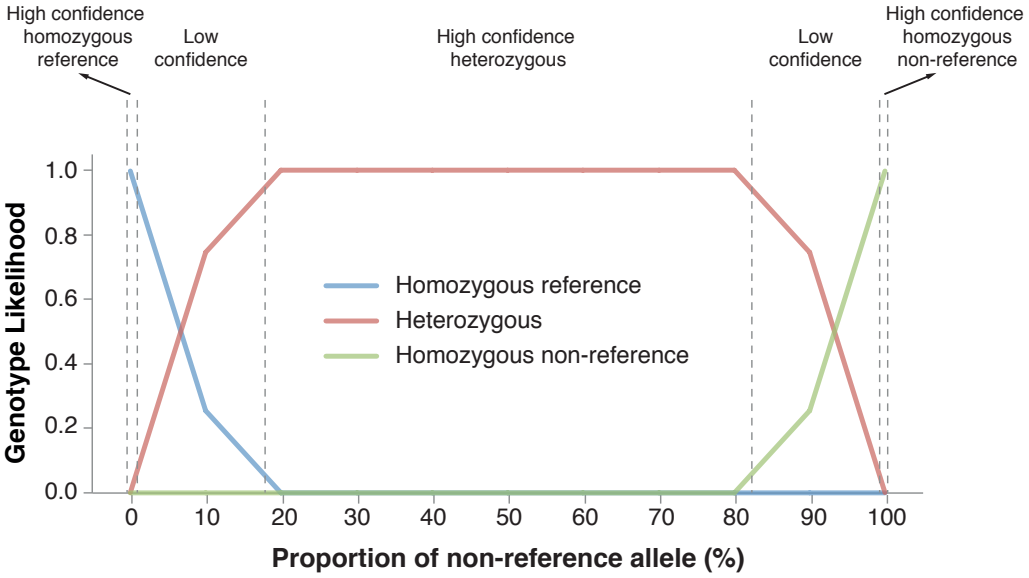


Figure 3 | Proportion of non-reference allele and genotype likelihoods

This plot shows the relationship between the proportion of non-reference allele at a bi-allelic locus covered by 10 reads and the genotype likelihood as modeled in the Genome Analysis Toolkit (GATK). The base and mapping quality are assumed to be constant across the reads. The blue line shows the confidence in the homozygous reference genotype, the red line in the heterozygous genotype and the green line in the homozygous non-reference genotype. High and low confidence zones (here defines as the confidence in the genotype >95%) are shown.

Conclusion

Fifteen years ago, the Human Genome Project (HGP) assembled the first human genome [43], providing the foundation for the research of human genetic variation. The HapMap Project characterized the common (>5% allele frequency) human single nucleotide variation [44–46], allowing for the first time to screen for disease associations without candidate genes. As a result genome-wide association studies (GWAS) identified thousands of risk alleles for common diseases, which have provided many clues about disease biology [47]. The total heritability explained by these alleles is however limited, suggesting that rare or more complex variants play a role in common disease too [48].

With next-generation sequencing (NGS) technologies it is now feasible to sequence the entire exome or genome in large number of samples, enabling the identification of common and rare genetic variation. The 1000 Genomes Project has now characterized most of the human genetic variation present in at least 1% allele frequency [1,14]. The Genome of the Netherlands (GoNL) Project presented in this thesis in one of many initiatives, e.g. SardiNIA Project (https://sardinia.irp.nia.nih.gov/) and UK10K (http://www.uk10k.org/) now identifying much rarer variants in specific populations.

The GoNL Project has characterized the genetic variants of all types in the general Dutch population and accurately phased these variants into a haplotypes. Combining these haplotypes with existing resources resulted in a marked improvement in downstream imputation accuracy of GWAS samples, especially for rare variants. These haplotypes should therefore contribute to a global effort in creating a cosmopolitan haplotype panel integrating all types of genetic variation, which will be instrumental in understanding the architecture of diseases.

The unique parent-offspring design of the GoNL Project also allowed the identification of *de novo* variation in the general population. Although rare events individually, *de novo* mutations are collectively common and could contribute substantially to some common diseases [49]. The observations of these mutations in the GoNL Project have allowed inferring background mutation rates in the population, which will be instrumental in calibrating the null expectations in *de novo* mutation studies.

With costs dropping, there is no doubt that larger projects will continue to assemble similar data sets as GoNL. Combining these data with functional annotation resources will be instrumental in the interpretation of the impact of individual alleles. Indeed, the ultimate goal in human genetics is to understand genotype-phenotype relationships at a functional level to give insight into the molecular underpinnings of disease.

References

1. McVean, G. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
2. Campbell, C. & Eichler, E. Properties and rates of germline mutations in humans. *Trends in genetics : TIG* **29,** 575–84 (2013).
3. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–5 (2012).
4. Michaelson, J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151,** 1431–42 (2012).
5. Schmidt, S. *et al.* Hypermutable non-synonymous sites are under stronger negative selection. *PLoS Genet.* **4,** e1000281 (2008).
6. Subramanian, S. & Kumar, S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res.* **13,** 838–44 (2003).
7. Green, P., Ewing, B., Miller, W., Thomas, P. J. & Green, E. D. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33,** 514–7 (2003).
8. Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18,** 337–40 (2002).
9. Hellmann, I., Ebersberger, I., Ptak, S. E., Pääbo, S. & Przeworski, M. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72,** 1527–35 (2003).
10. Koboldt, D. C., Ding, L., Mardis, E. R. & Wilson, R. K. Challenges of sequencing human genomes. *Brief. Bioinformatics* **11,** 484–98 (2010).
11. Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* **5,** 1005–10 (2008).
12. Glenn, T. C. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11,** 759–69 (2011).
13. Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* (2014).
14. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–73 (2010).
15. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **15,** 47–70 (2014).
16. Lindsay, S. J., Khajavi, M., Lupski, J. R. & Hurles, M. E. A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. *Am. J. Hum. Genet.* **79,** 890–902 (2006).
17. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10,** 551–64 (2009).

18. Montgomery, S. B. *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* **23,** 749–61 (2013).

19. Lam, H. Y. *et al.* Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **30,** 78–82 (2012).

20. O'Rawe, J. *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* **5,** 28 (2013).

21. Yu, X. & Sun, S. Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics* **14,** 274 (2013).

22. Pabinger, S. *et al.* A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinformatics* **15,** 256–78 (2014).

23. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297–303 (2010).

24. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

25. 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,** 2865–71 (2009).

26. Marschall, T., Hajirasouliha, I. & Schönhuth, A. MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics* **29,** 3143–50 (2013).

27. 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,** 974–84 (2011).

28. Coe, B. P., Chari, R., MacAulay, C. & Lam, W. L. FACADE: a fast and sensitive algorithm for the segmentation and calling of high resolution array CGH data. *Nucleic Acids Res.* **38,** e157 (2010).

29. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43,** 269–76 (2011).

30. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6,** 677–81 (2009).

31. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–72 (2010).

32. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74,** 5463–7 (1977).

33. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–45 (2008).

34. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–85 (1998).

35. LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.* **37,** 4181–93 (2009).

36. Pulit, S. L., Leusink, M., Menelaou, A. & Bakker, P. I. de. Association claims in the sequencing era. *Genes (Basel)* **5,** 196–213 (2014).

37. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9,** 356–69 (2008).

38. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–93 (2011).

39. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* **6,** 291–5 (2009).

40. Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29,** 908–14 (2011).

41. Sulonen, A.-M. M. *et al.* Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol.* **12,** R94 (2011).

42. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12,** 443–51 (2011).

43. Lander, ES, Linton, LM, Birren, B, Nusbaum, C & Zody, MC. Initial sequencing and analysis of the human genome. *Nature* (2001).

44. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).

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

46. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–8 (2010).

47. Visscher, P., Brown, M., McCarthy, M. & Yang, J. Five years of GWAS discovery. *American journal of human genetics* **90,** 7–24 (2012).

48. Eichler, EE, Flint, J, Gibson, G, Kong, A & Leal, SM. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* (2010).

49. Veltman, J. & Brunner, H. De novo mutations in human genetic disease. *Nature reviews. Genetics* **13,** 565–75 (2012).

Summary

The study of genetics started in the 19th century with Gregor Mendel's observation that traits were passed from one generation to the next in independent and unaltered units: the genes. In 1944, it was discovered that genes were encoded by the deoxyribonucleic acid (DNA) molecule. Throughout the 20th century, a few hundred of the ~20,000 human genes were identified and their sequence characterized, but the reading of the entire human genome DNA was only achieved in 2001. This milestone achievement started a new era in genetics by allowing the cataloguing of genetic variation between humans and the study of how these variations influenced traits. Today, almost all of the common genetic variation in humans is known and thousands of genetic variants and/or genes have been linked to traits and diseases. The identification of the rare genetic variation is progressing rapidly, but since rare variation varies regionally it requires large efforts in many populations. The work in this thesis presents the characterization of the genetic variation in the healthy Dutch population through the sequencing of 250 parent-offspring trios by the Genome of the Netherlands Project (GoNL).

In **Chapter 2**, I describe the spectrum of genetic variation found in the Dutch population and show that the knowledge about the genetic architecture of a single population can be instrumental in helping disease association studies. In addition, the comparison between genomes of Dutch people from different provinces provided insights about the history of the Dutch people. A fine-scale structure across the country consistent with multiple ancient migrations was unveiled, possibly reflecting population movements due to the frequent changes in sea levels and flooding that affected the Netherlands in the last millennia.

By sequencing the genomes of two generations, the GoNL project provides a unique opportunity to study *de novo* mutations, that is mutations present in the children but absent from the parents. These mutations were originally present in the sperm and egg of the parents and can be found in all the child's cells as they are derived from a single cell zygote formed by the fusion of the sperm and

egg. Such mutations occur between every generation and are the source of novel genetic variation that will then be subjected to natural selection and genetic drift. In **Chapter 3**, I present a novel computational approach for the discovery of *de novo* mutations in trio sequencing data. This method is applied in **Chapters 2, 4 and 5** to identify these mutations in the 250 families of the GoNL Project.

Chapters 2 and 4 focus on single nucleotide variations (SNVs): the smallest, but most abundant, type of variation where a single base of the genome is replaced with another. These mutations appear to be distributed non-randomly throughout the genome and that functional regions exhibit higher rates of mutations, possibly due to natural selection. In addition, there is also a strong enrichment in clusters of two or three mutations within individuals that display a markedly different mutational spectrum than other mutations. Finally, I show that the age of the father at conception influences the number and localization of *de novo* mutations in the offspring. Indeed, offspring of older fathers carry, on average, a larger number of mutations and these mutations also tend to locate more in genic (and therefore potentially functional) regions.

Chapter 5 presents the larger and more complex types of *de novo* variation found in the genomes of the children of the GoNL Project. The 332 *de novo* variants observed in this chapter highlight the remarkable complexity of these events ranging from small changes to large interchromosomal rearrangements. In particular, I found that a specific class of complex variants (complex indels) represents almost 5% of the observations while being almost absent from current genetic variation repositories. This indicates that this type of variation may be relatively common and has been understudied. The genomic footprint of larger *de novo* variants is much larger than that of SNVs in our data, both in functional and non-functional regions of the genome. This is in contrast with the variation segregating in the population and suggests that these large events are subject to strong selective pressures. Finally, similarly to SNVs, I found that most of these variants originated form the father's sperm rather than in the mother's egg.

Overall, this thesis presents the key results of a large-scale sequencing project in a single population. The parent-offspring design has proved beneficial, allowing in-depth characterization of common, rare and *de novo* SNVs, indels and SVs. Analyses of these data have revealed novel insights into inherited and *de novo* variation in human genomes.

Samenvatting

Met de observaties van de monnik Gregor Mendel (1822-1884) begon in de 19de eeuw het wetenschappelijk onderzoek naar de genetica. Mendel zag dat verschillende eigenschappen van levende organismen onafhankelijk van elkaar kunnen worden doorgegeven via de genen. In 1944 werd ontdekt dat moleculen met de naam desoxyribonucleïnezuur (DNA) het erfelijk materiaal bevatten en coderen voor genen. In de tweede helft van de 20ste eeuw werden vervolgens honderden van de in totaal 20.000 humane genen geïdentificeerd en werden delen van de DNA code ontrafeld. In 2001 werd voor het eerst het volledige menselijk genoom inclusief alle genen, in kaart gebracht. Met deze belangrijke mijlpaal begon een nieuw tijdperk waarin het onderzoek naar genetische variatie in de mens en de effecten daarvan op bepaalde eigenschappen, zowel in gezondheid en ziekte, enorm toenam. Nu, in 2015, zijn bijna alle veelvoorkomende variaties in het menselijk genoom bekend. Bovendien zijn er duizenden genetische variaties en/of genen geassocieerd met eigenschappen en ziekten. Hoewel er grote stappen zijn gemaakt in de identificatie van zeldzame genetische variatie, blijkt het vóórkomen van deze variatie enorm te verschillen tussen regio's en populaties. Het vereist daarom grote inspanningen in grote aantallen individuen uit verschillende populaties om zeldzame varianten te identificeren. In dit proefschrift beschrijf ik het onderzoek naar de genetische variatie in de Nederlandse populatie. Hiervoor is het DNA van 250 ouder-kind trio's onderzocht met behulp van de nieuwe generatie sequentietechnieken. Met deze methode is het volledige DNA in kaart gebracht. Dit is het Genoom van Nederland (GoNL) Project.

In **hoofdstuk 2** beschrijf ik het volledige spectrum aan genetische variatie zoals deze voorkomt in de Nederlandse populatie. Ik laat zien dat kennis over de genetische architectuur van een enkele populatie van belang is voor het vinden van associaties tussen genetische variatie en eigenschappen zoals bijvoorbeeld ziekte. Bovendien geef ik inzicht in de geschiedenis van de Nederlandse populatie door het DNA van mensen uit al de Nederlandse provincies te

vergelijken en de genetische verschillen in kaart te brengen. Ik breng aan het licht dat op basis van de genetische variatie aanwezig in de mensen uit de verschillende provincies patronen te zien zijn die de volksverhuizingen binnen Nederland naar aanleiding van de strijd tegen het water in de afgelopen millennia weerspiegelen.

Met het in kaart brengen van de genomen van twee generaties (ouders en kinderen) voorziet het GoNL Project in de unieke mogelijkheid om *de novo* veranderingen te onderzoeken. *De novo* veranderingen zijn veranderingen in het DNA van een individu, welke niet aanwezig zijn in het DNA van de ouders. Deze veranderingen waren oorspronkelijk wel aanwezig in de sperma- of eicel van een van de ouders. De verandering werd vervolgens doorgegeven aan het kind en wordt teruggevonden in de cellen van het kind die stuk voor stuk afkomstig zijn van de gefuseerde spermacel en eicel. Dergelijke *de novo* veranderingen komen voor in iedere generatie en zijn de bron van nieuwe genetische variatie. Deze nieuw ge**ïntroduceerde** variatie is vervolgens onderhevig aan natuurlijke selectie en verspreiding door voortplanting.

In **hoofdstuk 3** laat ik een nieuwe gecomputeriseerde benadering zien voor het identificeren van *de novo* veranderingen in het DNA van ouder-kind trio's. Deze methode heb ik vervolgens toegepast in de **hoofdstukken 2, 4 en 5** om veranderingen te identificeren in de 250 ouder-kind trio's in het GoNL Project.

In de **hoofdstukken 2 en 4** richt ik me op de één-nucleotide veranderingen (SNV's). Dit zijn de kleinste en tegelijkertijd ook meest voorkomende veranderingen in het genoom waarbij een enkele base (A, T, C of G) is vervangen door een andere base. Ik laat zien dat deze varianten niet-willekeurig verdeeld zijn over het genoom en dat het eiwit-coderende gedeelte van het genoom meer veranderingen bevat, waarschijnlijk onder invloed van natuurlijke selectie. Ten slotte bewijs ik dat de leeftijd van de vader tijdens de bevruchting 1) het aantal en 2) de locatie van *de novo* veranderingen in het genoom van het kind beïnvloedt. Kinderen van oudere vaders zijn drager van meer DNA veranderingen en deze veranderingen lijken vaker voor te komen in het eiwit-coderende en potentieel functionele deel van het DNA.

Hoofdstuk 5 behandelt de grote en meer complexe vormen van *de novo* variatie. In totaal werden 332 *de novo* veranderingen geïdentificeerd in het DNA van de kinderen in het GoNL project. Hiermee geef ik de complexiteit van herschikkingen tussen chromosomen overduidelijk weer. Specifiek wil ik hier ingaan op een klasse van complexe variaties (inserties en deleties) die ik tegenkwam in bijna 5% van de onderzochte populatie terwijl deze varianten vrijwel volledig ontbreken in de huidige databases die genetische variatie documenteren. Ik concludeer hieruit

dat dit type variatie relatief veel voorkomt, maar nog niet veel bestudeerd is. Uit onze onderzoeksresultaten blijkt dat de voetafdruk van grote, *de novo* veranderingen in het DNA veel groter is dan de voetafdruk van kleine, *de novo* SNV's. Dit is het geval in zowel het eiwit-coderende als het niet-coderende deel van het DNA. Dit contrasteert met niet-*de novo* variatie, waarbij de ratio grote:kleine veranderingen minder groot is. Dit suggereert dat varianten van grote omvang in hogere mate onderhevig zijn aan natuurlijke selectie dan kleinere veranderingen. Daarnaast zagen we dat de meeste grote, *de novo* veranderingen, net als kleine, *de novo* SNV's, vaker afkomstig zijn van de spermacel dan van de eicel.

Samenvattend laat ik in mijn proefschrift de resultaten zien van een grootschalig DNA onderzoek naar de genetische variatie in de Nederlandse populatie. Het bestuderen van ouder-kind trio's is een bewezen gunstige methode waarmee de veelvoorkomende, de zeldzame en ook *de novo* veranderingen in het DNA in detail kunnen worden onderzocht. Met dit onderzoek zijn nieuwe inzichten verkregen over de geërfde en *de novo* variatie in het menselijk genoom.

List of publications

- **L.C. Francioli***, Mircea Cretu-Stancu*, Kiran V. Garimella, Kaitlin E. Samocha, Benjamin M. Neale, Mark J. Daly, Eric Banks, Mark A. DePristo and Paul I.W. de Bakker. A framework for the detection of *de novo* mutations in trio sequencing data. Manuscript in preparation.
- **L.C. Francioli***, P.P. Polak*, A. Koren*, A. Menelaou, S. Chun, I. Renkens, Genome of the Netherlands Consortium, C.M. van Duijn, M.A. Swertz, C. Wijmenga, G.B. van Ommen, P.E. Slagboom, D.I. Boomsma, K. Ye, V. Guryev, P.F. Arndt, W.P. Kloosterman, P.I.W. de Bakker*, S.R. Sunyaev*. Genome-wide patterns and properties of *de novo* mutations in humans. Accepted at Nature Genetics.
- W.P. Kloosterman*, **L.C. Francioli***, F. Hormozdiari, T. Marschall, J.Y. Hehir-Kwa, A. Abdellaoui, E.W. Lameijer, M.H. Moed, V. Koval, I. Renkens, M.J. van Roosmalen, P. Arp, L.C. Karssen, B.P. Coe, R.E. Handsaker, E.D. Suchiman, E. Cuppen, D.T. Thung, M. McVey, M.C. Wendl, Genome of the Netherlands Consortium, A. Uitterlinden, C.M. van Duijn, M.A. Swertz, C. Wijmenga, G.B. van Ommen, P.E. Slagboom, D.I. Boomsma, A. Schönhuth, E.E. Eichler, P.I.W. de Bakker, K. Ye* and V. Guryev*. Characteristics of *de novo* structural changes in the human genome. Revised manuscript under review at Genome Research.
- C. van Duijn, E. van Leeuwen, L. Karssen, J. Deelen, A. Isaacs, C. Medina-Gomez, H. Mbarek, A. Kanterakis, S. Trompet, I. Postmus, N. Verweij, D. van Enckevort, J. Huffman, C. White, M. Feitosa, T. Bartz, A. Manichaikul, P. Joshi, G. Peloso, P Deelen, F. van Dijk, G. Willemsen, E. de Geus, Y. Milaneschi, B. Penninx, **L.C. Francioli**, A. Menelaou, S.L. Pulit, F. Rivadeneira, A. Hofman, B. Oostra, O. Franco, I. Mateo Leach, M. Beekman, A. de Craen, H.W. Uh, H. Trochet, L. Hocking, D. porteous, N. Sattar, C. Packard, B. Buckley, J. Brody, J. Bis, J. Rotter, J. Mychaleckyj, H. Campbell, Q. Duan, L. Lange, J. Wilson, C. Hayward, O. Polasek, V. Vitart, I. Rudan, A. Wright, S. Rich, B. Psaty, I. Borecki, P. Kearney, D. Stott, L. Cupples, Generation Scotland, CHARGE Lipids Working Group, The Genome of the Netherlands consortium, J.W. Jukema, P. van der Harst, E. Sijbrands, J.J. Hottenga, A. Uitterlinden, M.A.

Swertz, G.B. van Ommen, P.I.W. de Bakker, P.E. Slagboom, D.I. Boomsma, and C. Wijmenga. Population-specific imputations with the Genome of the Netherlands identify a ABCA6 variant associated with cholesterol levels. Accepted at Nature Communications.

- **L.C. Francioli***, A. Menelaou*, S.L. Pulit*, F. van Dijk*, P. Francesco Palamara, C.C. Elbers, P.B.T. Neerincx, K. Ye, V. Guryev, W.P. Kloosterman, P. Deelen, A. Abdellaoui, E.M. van Leeuwen, M. van Oven, M. Vermaat, M. Li, J.F.J. Laros, L.C. Karssen, A. Kanterakis, N. Amin, J.J. Hottenga, E.W. Lameijer, M. Kattenberg, M. Dijkstra, H. Byelas, J. van Setten, B. D.C. van Schaik, J. Bot, I.J. Nijman, I. Renkens, T. Marschall, A. Schönhuth, J.Y. Hehir-Kwa, R.E. Handsaker, P.P. Polak, M. Sohail, D. Vuzman, F. Hormozdiari, D. van Enckevort, H. Mei, V. Koval, M.H. Moed, K.J. van der Velde, F. Rivadeneira, K. Estrada, C. Medina-Gomez, A. Isaacs, S.A. McCarroll, M. Beekman, A.J.M. de Craen, H.E.D. Suchiman, A. Hofman, B. Oostra, A.G. Uitterlinden, G. Willemsen, LifeLines Cohort Study, M. Platteel, J.H. Veldink, L.H. van den Berg, S.J. Pitts, P. Potluri, P. Sundar, D.R. Cox, S.R. Sunyaev, J.T. den Dunnen, M. Stoneking, P. de Knijff, M. Kayser, Q. Li, Y. Li, Y. Du, R. Chen, H. Cao, N. Li, S. Cao, J. Wang, J.A. Bovenberg, I. Pe'er, P.E. Slagboom, C.M. van Duijn, D.I. Boomsma, G.B. van Ommen, P.I.W. de Bakker, M.A. Swertz, C. Wijmenga. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nature Genetics 2014; 46, 818–25.
- P. Deelen, A. Menelaou, E.M. van Leeuwen, A. Kanterakis, F. van Dijk, C. Medina-Gomez, **L.C. Francioli**, J.J. Hottenga, L.C. Karssen, K. Estrada, E. Kreiner-Møller, F. Rivadeneir, J. van Setten, J. Gutierrez-Achury, H.J. Westra, L. Franke, D. van Enckevort, M. Dijkstra, H. Byelas, C.M. van Duijn, Genome of the Netherlands Consortium, P.I.W. de Bakker, C. Wijmenga, M.A. Swertz. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands. Eur. J. Hum. Genet. 2014; 22,1321–6.
- J.J. Kuiper, J. Van Setten, S. Ripke, R. Van 'T Slot, F. Mulder, T. Missotten, G.S. Baarsma, **L.C. Francioli**, S.L. Pulit, C.G. De Kovel, N. Ten Dam-Van, A.I. Den Hollander, P. Huis in het Veld, C.B. Hoyng, M. Cordero-Coma, J. Martín, V. Llorenç, B. Arya, D. Thomas, S.C. Bakker, R.A. Ophoff, A. Rothova, P.I.W. de Bakker, T. Mutis, B.P. Koeleman. A genome-wide association study identifies a functional ERAP2 haplotype associated with birdshot chorioretinopathy. Hum Mol Genet. 2014; 23(22):6081-7.
- C.A. Brownstein, A.H. Beggs, N. Homer, B. Merriman, T.W. Yu, K.C. Flannery, E.T. DeChene, M.C. Towne, S.K. Savage, E.N. Price, I.A. Holm, L.J. Luquette, E. Lyon, J. Majzoub, P. Neupert, McCallie D Jr,P. Szolovits, H.F. Willard, N.J. Mendelsohn, R. Temme, R.S. Finkel, S.W. Yum, L. Medne, S.R. Sunyaev, I. Adzhubey, C.A. Cassa, P.I. de Bakker, H. Duzkale, P. Dworzyński, W. Fairbrother,

L.C. Francioli, B.H. Funke, M.A. Giovanni, R.E. Handsaker, K. Lage, M.S. Lebo, M. Lek, I. Leshchiner, D.G. MacArthur, H.M. McLaughlin, M.F. Murray, T.H. Pers, P.P. Polak, S. Raychaudhuri, H.L. Rehm, R. Soemedi, N.O. Stitziel, S. Vestecka, J. Supper, C. Gugenmus, B. Klocke, A. Hahn, M. Schubach, M. Menzel, S. Biskup, P. Freisinger, M. Deng, M. Braun, S. Perner, R.J. Smith, J.L. Andorf, J. Huang, K. Ryckman, V.C. Sheffield, E.M. Stone, T. Bair, E.A. Black-Ziegelbein, T.A. Braun, B. Darbro, A.P. DeLuca, D.L. Kolbe, T.E. Scheetz, A.E. Shearer, R. Sompallae, K. Wang, A.G. Bassuk, E. Edens, K. Mathews, S.A. Moore, O.A. Shchelochkov, P. Trapane, A. Bossler, C.A. Campbell, J.W. Heusel, A. Kwitek, T. Maga, K. Panzer, T. Wassink, D. Van Daele, H. Azaiez, K. Booth, N. Meyer, M.M. Segal, M.S. Williams, G. Tromp, P. White, D. Corsmeier, S. Fitzgerald-Butt, G. Herman, D. Lamb-Thrush, K.L. McBride, D. Newsom, C.R. Pierson, A.T. Rakowsky, A. Maver, L. Lovrečić, A. Palandačić, B. Peterlin, A. Torkamani, A. Wedell, M. Huss, A. Alexeyenko, J.M. Lindvall, M. Magnusson, D. Nilsson, H. Stranneheim, F. Taylan, C. Gilissen, A. Hoischen, B. van Bon, H. Yntema, M. Nelen, W. Zhang, J. Sager, L. Zhang, K. Blair, D. Kural, M. Cariaso, G.G. Lennon, A. Javed, S. Agrawal, P.C. Ng, K.S. Sandhu, S. Krishna, V. Veeramachaneni, O. Isakov, E. Halperin, E. Friedman, N. Shomron, G. Glusman, J.C. Roach, J. Caballero, H.C. Cox, D. Mauldin, S.A. Ament, L. Rowen, D.R. Richards, F.A. San Lucas, M.L. Gonzalez-Garay, C.T. Caskey, Y. Bai, Y. Huang, F. Fang, Y. Zhang, Z. Wang, J. Barrera, J.M. Garcia-Lobo, D. González-Lamuño, J. Llorca, M.C. Rodriguez, I. Varela, M.G. Reese, F.M. De La Vega, E. Kiruluta, M. Cargill, R.K. Hart, J.M. Sorenson, G.J. Lyon, D.A. Stevenson, B.E. Bray, B.M. Moore, K. Eilbeck, M. Yandell, H. Zhao, L. Hou, X. Chen, X. Yan, M. Chen, C. Li, C. Yang, M. Gunel, P. Li, Y. Kong, A.C. Alexander, Z.I. Albertyn, K.M. Boycott, D.E. Bulman, P.M. Gordon, A.M. Innes, B.M. Knoppers, J. Majewski, C.R. Marshall, J.S. Parboosingh, S.L. Sawyer, M.E. Samuels, J. Schwartzentruber, I.S. Kohane, D.M. Margulies. 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(3):R53

- D.I. Boomsma, C. Wijmenga, E.P. Slagboom, M.A. Swertz, L.C. Karssen, A. Abdellaoui, K. Ye, V. Guryev, M. Vermaat, F. van Dijk, **L.C. Francioli**, J.J. Hottenga, J.F. Laros, Q. Li, Y. Li, H. Cao, R. Chen, Y. Du, N. Li, S. Cao, J. van Setten, A. Menelaou, S.L. Pulit, J.Y. Hehir-Kwa, M. Beekman, C.C. Elbers, H. Byelas, A.J. de Craen, P. Deelen, M. Dijkstra, J.T. den Dunnen, P. de Knijff, J. Houwing-Duistermaat, V. Koval, K. Estrada, A. Hofman, A. Kanterakis, Enckevort Dv,H. Mai, M. Kattenberg, E.M. van Leeuwen, P.B. Neerincx, B. Oostra, F. Rivadeneira, E.H. Suchiman, A.G. Uitterlinden, G. Willemsen, B.H. Wolffenbuttel, J. Wang, P.I. de Bakker, G.J. van Ommen, C.M. van Duijn. The

Genome of the Netherlands: design, and project goals. Eur J Hum Genet. 2014; 22(2):221-7

- A. Kiezun, S.L. Pulit, **L.C. Francioli**, F. van Dijk, M. Swertz, D.I. Boomsma, C.M. van Duijn, P.E. Slagboom, G.J. van Ommen, Wijmenga C; Genome of the Netherlands Consortium, P.I.W. de Bakker, S.R. Sunyaev. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. PLoS Genet. 2013; 9(2):e1003301.
- D. Vernez, A. Milon, **L.C. Francioli**, J.L. Bulliard, L. Vuilleumier, L. Moccozet. A numeric model to simulate solar individual ultraviolet exposure. Photochem Photobiol. 2011; 87(3):721-8.

Conference Presentations

- **L.C. Francioli**, P.P. Polak, W.P. Kloosterman, S.S. Sunyaev, P.I.W. de Bakker, Genome of the Netherlands Consortium. *De novo* mutations in the Genome of the Netherlands Project. Platform presentation at the Gordon Research Seminar on Human Genetics & Genomics, Smithfield (RI), USA, 2013
- **L.C. Francioli**, P.P. Polak, W.P. Kloosterman, S.S. Sunyaev, P.I.W. de Bakker, Genome of the Netherlands Consortium. *De novo* mutations in the Genome of the Netherlands Project. Poster presentation at the Gordon Research Conference on Human Genetics & Genomics, Smithfield (RI), USA, 2013
- **L.C. Francioli**, P.P. Polak, W.P. Kloosterman, S.S. Sunyaev, P.I.W. de Bakker, Genome of the Netherlands Consortium. *De novo* mutations in the Genome of the Netherlands Project. Platform presentation at the European Society for Human Genetics meeting, Paris, France, 2013.
- **L. C. Francioli**, K. V. Garimella, K. E. Samocha, F. Van Dijk, B. M. Neal, M. J. Daly, E. Banks, M. Swertz, M. A. DePristo, P. I. W. de Bakker. Trio-aware variant calling for accurate genotyping and de novo mutation detection. Platform presentation at the European Society for Human Genetics meeting, Nuremberg, Germany, 2012.

Acknowledgements

This thesis would not have been possible without the support of many people to whom I am greatly indebted.

To my supervisor, Prof. **Paul de Bakker**, for giving me the chance to join your lab and steering my career in a different direction, for teaching me so much about genetics and statistics and for being such an inspiring supervisor. I can still vividly feel my astonishment when we ended our first Skype call after you invited me to visit your lab in Boston all the way from Damascus after a short 15 minutes chat and despite my unusual path. In these last few years working with you, I appreciated that this was one of the many examples where you valued motivation and personal contact over protocol and normality. In fact, my somewhat unusual path continued throughout my PhD, first when I started by working in Groningen while you were in Boston, to then move shortly to Boston when you came back to the Netherlands. Even after we were finally both in the same country, I started commuting first with Paris and then with Aberdeen to spend time with my wife and son. I cannot thank you enough for your flexibility with respect to where and when I was working; it has enabled me to discover different labs, grow my professional network and spend countless happy moments with my family. Even more remarkable is that although it sometimes felt like we would not meet for a month at a time, you gave me the best guidance I could have expected from a supervisor. You always questioned everything while keeping an open mind and bubbling with out-of-the-box suggestions, and I thoroughly enjoyed our sometimes long and often impromptus meetings going through an idea or a problem in all its depth. I hope we will keep in touch in the future and continue working together on many projects!

To Prof. **Cisca Wijmenga** and Dr. **Morris Swertz** for hosting me at the University Medical Center Groningen (UMCG) during the first 6 months of my thesis and for training me in genetics data analysis and in cluster computing. I spent a fantastic few months at UMCG and am very grateful to you for this opportunity. Of course,

all the people in the lab were the source of my daily fun and learning! In particular, **Freerk Van Dijk**, **Alexandros Kanterakis**, **Pieter Neerincx**, **George Byelas** and **Martijn Dijkstra** for the nice atmosphere in our shared office and the punctual early Dutch lunches! But also to **Despoina Antonakaki**, **Jihane Romanos**, **Barbara Hrdlickova**, **Céline Martineau**, **Isis Ricano Ponce**, **Gosia Trynka**, **Javier Gutierrez-Achury**, **Sabyasachi Senapati**, **Joeri der Velde**, **Patrick Deelen** and everyone else in the department for all the good times in and out of the office.

To Prof. **Shamil Sunyaev**, for teaching me about mutations, mutagenesis and population genetics. I have enjoyed our collaborations immensely and it goes without saying that my genetics knowledge and this thesis would not be the same without your involvement. At first, chatting with you was very challenging due to the numerous concepts that were natural to you but entirely foreign to me, and every time we hung up I was left with the strange feeling that we had elaborated a plan which details escaped me. But thanks to your patience, and with the help of **Paz Polak**, I now feel like we speak the same language and can reason together! My thanks of course extend to everyone in your lab, in particular **Paz Polak** for the long (and often quite fun) work sessions and **Dana Vuzman** for our ongoing collaboration on transmission disequilibrium, **Mashaal Sohail** for finding all types of biases in my data and **Sung Chun** for his extraordinary work on our mutation map.

To the members of my thesis reading committee, Prof. **Edwin Cuppen**, Prof. **Cisca Wijmenga**, Prof. **Danielle Posthuma**, Prof. **Jan Veldink** and Prof. **Frank Holstege** for taking the time to read and comment on this work.

To everyone in the Genome of the Netherlands Project (GoNL), for the incredible collaborations and the interesting meetings all over the Netherlands. In particular, thanks to **Victor Guryev**, **Kai Ye**, **Alexander Schönhuth**, **Tobias Marschall**, **Jayne Hehir-Kwa**, **Bob Handsacker** and the rest of the GoNL Structural Variation group for all the interesting discussions we had and for our collaborations on analyzing the most difficult and exotic genetic variation in the GoNL. I would of course like to thank the steering committee for putting this project together and the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL) for funding it.

To **Mark DePristo**, **Eric Banks**, **Ryan Poplin**, **Mauricio Carneiro** and **Kiran Garimella** at the Genome Sequencing and Analysis (GSA) group at the Broad Institute for welcoming me to their meetings when I was in Boston and for the numerous and fruitful discussions we had on how to analyze our sequencing data.

To Prof. **Suzanne Leal**, for inviting me to teach at the Max Delbrück Center in Berlin and at the Rockefeller University in New-York. I had a great time during these courses, both in and outside the classroom, and am looking forward to our next teaching events already!

To **Lennart Karssen**, **Amin Najaf**, **Jasper Saris** for inviting me to teach at the Erasmus Medical Center in Rotterdam. These were my first experiences teaching about sequencing data analysis and I thoroughly enjoyed them.

To **Amnon Koren** for all our discussions on replication timing and our collaboration on de novo mutations.

To **Amalio Telenti** for introducing me to Paul in the first place and making this all possible.

To the de Bakker lab, for….well for so much! It saddens me to think that soon I won't be a part of this lab anymore and won't enjoy your company on a daily basis! I have enjoyed so much the light and positive lab atmosphere and all our (very regular) weekly meetings! **Androniki** and **Sara**, thanks so much for helping me understand statistics and actually making them cool! Thanks as well for the long evening work sessions we had on the GoNL main paper… as much as it was crazy hours and work load – sometimes in a no-GoNL moods -, it was also a lot of fun to be plumbing all the pieces of this massive paper together (and picking colors for the plots)! **Vinicius**, **Mircea**, **Balder** and **Daniel**, thanks for all the nice lunch breaks we took and all the discussions we had then and in the office! **Jessica**, **Jytte**, **Maarten**, **Femke**, **Sander** and **Daiane**, thanks for all the good times in the office but also at meetings, conferences and retreats!

To everyone in the UMCU Medical Genetics departments, for the relaxed work atmosphere, the interesting and critical discussions and of course the regular birthday cakes and borrels! I'd like especially to thank **Wigard** for all our fun discussions and our fruitful collaboration, **Ies** for being patient and always happy to answer my many sequencing questions and **Monique** for helping me so much with the admin stuff I hate! **Nayia** and **Kirsten**, thanks so much for always finding the time to discuss issues and help with the most random things I was confronted with during my PhD, and for providing an "outsider" view on my work when I needed one. I'd also like to thank the whole lunch crew, **Glen**, **Glenn**, **Marijn**, **Ruben**, **Irena**, **Maartje** for all these good relaxing moments we had! Of course, my life would not have been the same without the fun people in the Flexroom: **Flip**, **Kirsten**, **Eric** and everyone else!

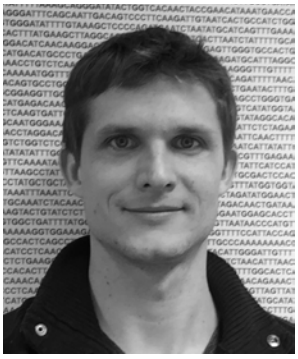
To **Gijs**, for all the great time we had making Paul's oratie movie, for the epic kiteboarding sessions, for the coffee breaks, lunches and dinners and of course for being my one of my paranymphs.

To Masskiting and all the kiters in the Boston area for making my stay in Boston such a memorable experience! In particular, **Leo**, **Ytzen**, **Jean**, **Kirk**, **Terrance**, **Johnny**, **Kyle** and **Eric** for "adopting me" from day 1 and for all the sessions and afer-sessions we shared together.

To my parents, **Patrick** and **Martine,** for always being supportive of my moves around the world and for taking such great interest in my work! To all my siblings, **Olivier**, **Cyril**, **Jade**, **Yannick** and **Charlotte** whom I wish I could see much more often.

Finally, but most importantly, to my wonderful wife **Lydia** and my son **Rayan**. Lydia, you have accepted to live in different countries so that I could pursue my PhD in genetics, and even after Rayan was born you were fine with my international weekly commute. This took extraordinary strength and I am forever grateful to you for doing it. In addition, you have also supported me when I needed to work long hours, encouraged when I was having doubts and provided me with helpful critical feedback. Most importantly, Rayan and you make my life fun and lovable everyday and I am so happy that we now finally all live together again!

Curriculum vitae



Laurent C. Francioli was born on March 23rd 1983 in Lausanne, Switzerland. In 2001, he started studying Computer Science at the Swiss Federal Institute of Technology in Lausanne (EPFL). In 2003, he organized and participated in the first student exchange between EPFL and the University of New South Whales (UNSW) in Sydney (Australia) where he spent one year and was awarded a study abroad scholarship for his academic results. In 2005, he spent 6 months as an algorithm development intern at Roche Molecular Systems in Alameda (California, USA) where he developed a novel algorithm for polymerase chain reaction (PCR) data analysis and co-authored two international patents. Laurent obtained his Master's degree in Computer Science from EPFL in 2006 after completing his dissertation on "The Blue Brain Project Database" under the supervision of Dr. Fabio Porto, Dr. Asif Jan and Prof. Henry Markram. He then joined an EPFL spin-off, routeRANK, as the first software engineer and developed its travel search engine. He stayed with routeRANK for two years, during which time routeRANK was incorporated, got numerous awards and grew to 8 employees. In 2009, Laurent started working as a freelancer and provided software development and IT consulting services to clients including the World Health Organization (WHO), the Institute of Work and Health (IST) and the University Hospital of Lausanne (CHUV). Although his business was based in Switzerland, he lived in Damascus (Syria) where he taught programming at the Syrian Virtual University (SVU). In 2011, Laurent started his PhD in human genetics at the University Medical Center Utrecht (UMCU), the Netherlands, under the supervision of Prof. Paul de Bakker. During his graduate studies, he spent 6 months in the lab of Prof. Cisca Wijmenga at the University Medical Center Groningen (UMCG) and 3 months at the Broad Institute of Harvard and MIT in Cambridge, Massachusetts. In addition to his research, Laurent also taught on sequencing data analysis at various courses in Rotterdam, Berlin and New York. Laurent currently lives in Amsterdam with his wife Lydia, his son Rayan and his daughter Ayla.