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

Authors and affiliations

Wigard P. Kloosterman[1,18], Laurent C. Francioli[1,18], Fereydoun Hormozdiari[2], Tobias Marschall[3], Jayne Y. Hehir-Kwa[4], Abdel Abdellaoui[5], Eric-Wubbo Lameijer[6], Matthijs H. Moed[6], Vyacheslav Koval[7], Ivo Renkens[1], Markus J. van Roosmalen[1], Pascal Arp[7], Lennart C. Karssen[8], Bradley P. Coe[2], Robert E. Handsaker[9], Eka D. Suchiman[6], Edwin Cuppen[1], Djie T. Thung[4], Mitch McVey[10], Michael C. Wendl[11,12], Genome of the Netherlands Consortium, Andre Uitterlinden[7], Cornelia M. van Duijn[8], Morris Swertz[13,14], Cisca Wijmenga[13,14], Gertjan van Ommen[15], P. Eline Slagboom[6], Dorret I. Boomsma[5], Alexander Schönhuth[3], Evan E. Eichler[2], Paul I. W. de Bakker[1,16], Kai Ye[11,*] and Victor Guryev[17,*]

[1] Department of Medical Genetics, University Medical Center Utrecht, Utrecht, 3584CG, The Netherlands

[2] Department of Genome Sciences, University of Washington, Seattle, 98105, USA

[3] Life Sciences Group, Centrum voor Wiskunde en Informatica, Amsterdam, 1098XG, The Netherlands

[4] Department of Human Genetics, Radboud University Medical Center, Nijmegen, 6525GA, The Netherlands.

[5] Department of Biological Psychology, VU University Amsterdam, Amsterdam, 1081BT, The Netherlands

[6] Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, 2300RC, The Netherlands

[7] Department of Internal Medicine, Erasmus Medical Center, Rotterdam, 3000CA, The Netherlands

[8] Department of Epidemiology, Erasmus Medical Center, Rotterdam, 3000CA, The Netherlands

[9] Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

[10] Department of Biology, Tufts University, Medford, MA 02115, USA

[11] The Genome Institute, Washington University, St. Louis, MO 63108, USA

[12] Department of Mathematics, Washington University, St. Louis, MO 63108, USA

[13] Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, 9700RB, The Netherlands

[14] Genomics Coordination Center, University of Groningen, University Medical Center Groningen, Groningen, 9700RB, The Netherlands

[15] Department of Human Genetics, Leiden University Medical Center, Leiden, 2300RC, The Netherlands

[16] Department of Epidemiology, University Medical Center Utrecht, Utrecht, 3584CG, The Netherlands

[17] European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, Groningen, 9713AD, The Netherlands

[18] These authors contributed equally to this work

* Correspondence:

Kai Ye: The Genome Institute, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Ave, St. Louis, MO 63108, USA. Phone: +1 (314) 286-1800. Fax: +1 (314) 286-1810. E-mail: kye@genome.wustl.edu

Victor Guryev: European Research Institute for the Biology of Ageing, University Medical Center Groningen, Antonius Deusinglaan 1, Building 3226, Int. Zip FA50, 9700AD, Groningen, The Netherlands. Phone: +31 (6) 5272 4873. Fax: +31 (50) 361 7310. Email: v.guryev@umcg.nl

Running title: *De novo* structural changes in the human genome

Keywords: whole genome sequencing; *de novo* mutations; structural variants; indels; general human population.

Manuscript type: Research Article

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 2.94 indels (1-20bp) and 0.16 SVs (>20bp) per generation. *De novo* 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. 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 single nucleotide variations (SNVs) (Cheng et al. 2005). 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 (Scally and Durbin 2012). While several projects have started to investigate the rates and characteristics of *de novo* SNVs (Michaelson et al. 2012; Francioli et al. 2014; Kong et al. 2012; Besenbacher et al. 2015), those of *de novo* short insertions and deletions (indels) and large structural variants (SVs) have been much less studied (Campbell and Eichler 2013).

Copy number variations (CNVs) and SVs contribute substantially to human genetic variation (Korbel et al. 2007; Sebat et al. 2004; Iafrate et al. 2004; Tuzun et al. 2005) and the phenotypic impact of CNVs may be larger than of SNVs (Conrad et al. 2010; Stranger et al. 2007; Redon et al. 2006). The impact of novel changes in genome structure is further illustrated by their role in human genetic disease (Cooper et al. 2011; Stankiewicz and Lupski 2010). Copy number variations (CNVs) are widely studied and have been implicated in a variety of neurological disorders, such as autism (Sebat et al. 2007), schizophrenia (Walsh et al. 2008) and intellectual disability (Cooper et al. 2011). 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 (Veltman and Brunner 2012).

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 (Boomsma et al. 2014; Francioli et al. 2014).

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 (Londin et al. 2011). 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 (Francioli et al. 2014).

Indels (1-20bp) were called using three different tools using information from gapped reads and split-reads (**Fig. 1, Methods**). We focused exclusively on variants that were detected only in a single child by at least one tool with high confidence (**Supplemental**

Table 1). We performed experimental validation assays for all 1,169 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 968 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* indels 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 pair, 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 parent. 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 deep-coverage whole-exomes of 24 parents. This revealed 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 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, Methods**). We identified a total of 601 *de novo* SV candidates in the 258 GoNL offspring based on permissive call settings and visual inspection using the Integrated Genome Viewer (IGV) (Robinson et al. 2011) (**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**). The *de novo* SV 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 (Stewart et al. 2011). 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 might be recurring in the population (**Supplemental Table 2**).

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 (Campbell and Eichler 2013; Kondrashov 2003; Lynch 2010; Besenbacher et al. 2015; Ramu et al. 2013). The mutation rate for copy number variants was estimated to be 0.03 CNVs larger than 500bp (Conrad et al. 2010) and 0.012 CNVs larger than 100kbp (Itsara et al. 2010) 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 and spectrum of structural changes (**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 *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 (Belancio et al. 2008; Burns and Boeke 2012).

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 x 10^{-4}). Together, these data suggest possible differences in the effects of environmental factors or the vulnerability for acquiring *de*

novo SVs per family (Conrad et al. 2011). 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 (Kong et al. 2012; Michaelson et al. 2012; Francioli et al. 2014; Jiang et al. 2013; Conrad et al. 2011). In addition, the majority of sporadic *de novo* CNVs and cytogenetically balanced genomic rearrangements in patients with congenital disorders are paternal in origin (Hehir-Kwa et al. 2011; Batista et al. 1994). 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 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 (Montgomery et al. 2013; Bhangale et al. 2005), 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, $p < 2.2 \times 10^{-16}$) (Montgomery et al. 2013), 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 *de novo* indels occurred in non-repeat regions and did not lead to a copy count change (non-CCC). These likely result from imperfect double-stranded DNA break repairs by non-homologous end-joining (NHEJ) which can create indels at the repair junction. Their very high deletion to insertion ratio of 12.9:1 supports their occurrence through NHEJ (Hastings et al. 2009). 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 (Hastings et al. 2009; Montgomery et al. 2013). Another five non-CCC indels presented microhomologies of at least 4bp, possibly indicating emergence through microhomology-mediated end joining (MMEJ) (McVey and Lee 2008).

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 (Levy et al. 2007), 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) (Yu and McVey 2010) and theta-mediated end joining (TMEJ) (Roerink et al. 2014). 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**) (Lam et al. 2010). 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 homology-driven repair and these are classified as mediated 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 *Alu*Y family (**Supplemental Fig. 2, Supplemental Table 2**). The sequences of the breakpoint junctions of the *de novo Alu*Y 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 (Burns and Boeke 2012).

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 *Alu*Y 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 (Mills et al. 2011; Pang et al. 2010, 2013; Kidd et al. 2010) (**Supplemental Table 4**) and found substantial differences between studies, which probably reflect methodological differences (Pang et al. 2013).

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. Examination of these genes in the exome sequencing database from the Exome Aggregation Consortium (ExAC, http://exac.broadinstitute.org) revealed that all of them contain heterozygous loss-of-function mutations in the population. This indicates that heterozygous changes in these genes possibly have no early developmental consequences. Mutations in two of the affected genes - *BANK1* and *PROC* - are associated with systemic lupus erythematosus (Kozyrev et al. 2008) and thrombophilia (Romeo et al. 1987), 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 (**Supplemental Methods**). Consistent with recent studies involving families with disorders (Kong et al. 2012; Michaelson et al. 2012; Jiang et al. 2013; Gilissen et al. 2014), an average of 45 *de novo* SNVs per child were detected in the GoNL Project (Francioli et al. 2014). 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 28.6 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 18 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.

Discussion

The human genome continuously evolves as a result of mutation and selection. Because of the relatively low rate of SV and indel formation, large numbers of parent-offspring families are required to capture the full spectrum of *de novo* changes that alter genome structure every generation (Campbell and Eichler 2013). Moreover, the detection and genotyping of these variants remain challenging given their diversity in both size and type (Alkan et al. 2011). Although limited by the short size of reads and the relatively low coverage depth used in this study, we have provided a representative picture of the landscape of *de novo* SVs and indels in the human genome based on whole-genome sequencing of 250 families by leveraging multiple calling approaches.

Our work demonstrates that both *de novo* indels and SVs originate primarily in the paternal germline, complementing recent findings on *de novo* SNVs (Kong et al. 2012; Michaelson et al. 2012; Francioli et al. 2014). 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 (Stankiewicz and Lupski 2010). Roughly 15% of patients with intellectual disability or congenital abnormalities harbor an apparently causative CNV, most of which occur *de novo* (Hochstenbach et al. 2011). 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 greater 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**) (Mills et al. 2011; Conrad et al. 2010). 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 (Stankiewicz and Lupski 2010). In this study we demonstrate that a broad range of *de novo* indels and structural variations is also characteristic for individuals obtained from a general human population.

Methods

Whole genome sequencing and alignment

Genomic DNA from nucleated blood cells was obtained from 250 Dutch families (231 trios, 8 quartets with dizygotic twins and 11 quartets with monozygotic twins), which were selected without phenotypic ascertainment. Library construction and whole-genome sequencing was performed 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 (Li and Durbin 2009). We expect that alignment to GRCh38 would not significantly alter our findings, given that *de novo* variation is dependent on differences between parental and offspring genomes. Aligned data were processed following the Genome Analysis Toolkit (GATK) best practices v2 (DePristo et al. 2011): 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 (Francioli et al. 2014).

Detection of *de novo* variants

Indels were called using GATK UnifiedGenotyper (DePristo et al. 2011) and Pindel (Ye et al. 2009) 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 mutation prior of 10^{-4} per base per generation. We kept 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 in the child, (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 *de novo* candidates. 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 SVs were called and filtered independently by 11 algorithms based on the following approaches: gapped/split read mapping (Pindel (Ye et al. 2009), GATK UnifiedGenotyper (DePristo et al. 2011), GATK HaplotypeCaller), analysis of discordant pairs (BreakDancer (Chen et al. 2009), 1-2-3-SV (http://tools.genomes.nl/123sv.html) (Kloosterman et al. 2011), Genome STRiP (Handsaker et al. 2011), MATE-CLEVER (Marschall et al. 2013)), read depth analysis (CNVnator (Abyzov et al. 2011), DWAC-seq (http://tools.genomes.nl/dwac-seq.html), FACADE (Coe et al. 2010)). In addition, Mobster was used to call *de novo* mobile element insertions (MEIs) (Thung et al. 2014). 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 (Robinson et al. 2011) to discard false positives due to alignment artifacts. 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 (Luo et al. 2012)) was used for breakpoint fine-mapping for SVs greater than 100bp. A detailed

description of the tools, settings, filtering and variant calls, including sensitivity analysis is provided in the **Supplemental Methods**.

Experimental validation

Oligonucleotide primers for amplification of a genomic segment containing the variant (for mutations smaller than 100bp) or variant breakpoints (for larger SVs) were designed using Primer3 software (**Supplementary Table 2**). PCR products were resequenced with Sanger, IonTorrent or MiSeq (2x250 bp) technologies. Genotyping of the resequenced variants is described in the **Supplemental Methods**.

Parental origin

We used genotypes from phased haplotypes (Francioli et al. 2014) 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.

Parental haplotypes for SVs were determined from allele ratios at overlapping SNPs. Assignment to the paternal or maternal haplotype was made if: a) one or more homozygous alleles in the offspring are located inside a *de novo* deletion and could only be inherited from one parent; b) 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; c) a SNP in the offspring was located within a 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 = 0.0092) 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).

We used a multinomial model with equal probability for each child to receive a *de novo* variant to test for uniform distribution of variants across children (goodness-of-fit p-value obtained using 100,000 Monte Carlo replicates).

In one sample, we observed 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, the second the number of placements that would result in the collapse of both events. If D is the observed distance in bp between the two events, then the number of the total placements that are significant 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. Using permutations, we ruled out any correlation between siblings from a DZ pair ($p = 0.59$). The rate was 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 all 250 families. Only one child was considered for each of the monozygotic twin pairs and siblings from dizygotic twin pairs were considered as genetically independent with respect to *de novo* SVs. The rate was calculated by dividing the number of *de novo* SVs (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*. (Montgomery et al. 2013), except for Predicted Hotspots (PR) that we did not use since they were not readily available and complex indels that are new in our data.

Analysis of mutation formation mechanisms of SVs was performed using BreakSeq software v. 1.3 (Lam et al. 2010). See the **Supplemental Methods** for a full description of indel and SV classification.

Data access

Sequence data have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI), and are available via https://www.ebi.ac.uk/ega/studies/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).

Author Contributions

WPK, LCF, VG and KY planned and directed the research. CW, CD, GO, PES, DB and PdB supervised the GoNL data generation. MS supervised read mapping. JHK, AS, TM, FH, BPC, AA, EL, DTT, MM, REH, VG, KY, LCF and VK performed SV and indel calling. IR, EC, PA, WPK and LK performed validation assays and analysis. EDS, CW, AU and DB provided DNA samples for validation. MR, LCF and VG performed phasing of indels and SVs. MW, TM and AS provided input for statistical testing. MMcV provided input on indel formation mechanisms. WPK, LCF, KY and VG wrote the manuscript. FH, CW, TM, AS, JHK, EC, MCW, CMD, DB, AA, EE and PIWB commented on the manuscript.

Figure legends

Figure 1. Overview of study design. A total of 250 parent-offspring families were sequenced at 14.5x 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 (1,169 in 99 children) and SVs (601 in 258 children) were subjected to experimental validation leading to 291 validated *de novo* indels and 41 *de novo* SVs.

Figure 2. Frequency of *de novo* indels and SVs. (**A**) Size-frequency distribution of 332 validated *de novo* indels and SVs identified in this study. In addition, the frequency of *de novo* SNVs is shown (Francioli et al. 2014). 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.

Figure 3. Overview of *de novo* and inherited 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.

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.

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. **(E)** Duplication of the entire *PROC1* gene. **(F)** Duplication of 3 entire genes (*GCNT3*, *GTF2A2*, *BNIP2*). 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 the ENCODE project (The ENCODE Project Consortium, 2007). The y-axis shows the ratio of the average number of affected bases per offspring relative to SNVs.

References

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

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363–376.

Batista DAS, Pai GS, Stetten G. 1994. Molecular analysis of a complex chromosomal rearrangement and a review of familial cases. *Am J Med Genet* **53**: 255–263.

Belancio VP, Hedges DJ, Deininger P. 2008. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Res* **18**: 343–358.

Besenbacher S, Liu S, Izarzugaza JMG, Grove J, Belling K, Bork-Jensen J, Huang S, Als TD, Li S, Yadav R, et al. 2015. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* **6**: 5969.

Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. 2005. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum Mol Genet* **14**: 59–69.

Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, Ye K, Guryev V, Vermaat M, van Dijk F, et al. 2014. The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* **22**: 221–7.

Burns KH, Boeke JD. 2012. Human transposon tectonics. *Cell* **149**: 740–752.

Campbell CD, Eichler EE. 2013. Properties and rates of germline mutations in humans. *Trends Genet* **29**: 575–584.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**: 677–681.

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Pääbo S, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**: 88–93.

Coe BP, Chari R, MacAulay C, Lam WL. 2010. FACADE: A fast and sensitive algorithm for the segmentation and calling of high resolution array CGH data. *Nucleic Acids Res* **38**: e157.

Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella K V, et al. 2011. Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**: 712–714.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704–712.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, et al. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* **43**: 838–846.

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Francioli LC, Menelaou A, Pulit SL, van Dijk F, Palamara PF, Elbers CC, Neerincx PBT, Ye K, Guryev V, Kloosterman WP, et al. 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**: 818–25.

Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A, et al. 2014. Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**: 344–347.

Handsaker RE, Korn JM, Nemesh J, McCarroll SA. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**: 269–276.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet* **10**: 551–64.

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

Hochstenbach R, Buizer-Voskamp JE, Vorstman JAS, Ophoff RA. 2011. 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.

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

Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE. 2010. De novo rates and selection of large copy number variation. *Genome Res* **20**: 1469–81.

Jiang Y, Yuen RKC, Jin X, Wang M, Chen N, Wu X, Ju J, Mei J, Shi Y, He M, et al. 2013. Detection of Clinically Relevant Genetic Variants in Autism Spectrum Disorder by Whole-Genome Sequencing. *Am J Hum Genet* **93**: 249–263.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**: 837–847.

Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, Bakker SCM, Letteboer T, van Nesselrooij B, Hochstenbach R, Poot M, et al. 2011. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum Mol Genet* **20**: 1916–1924.

Kondrashov AS. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Hum Mutat* **21**: 12–27.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**: 471–475.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science* **318**: 420–6.

Kozyrev S V, Abelson A-K, Wojcik J, Zaghlool A, Linga Reddy MVP, Sanchez E, Gunnarsson I, Svenungsson E, Sturfelt G, Jönsen A, et al. 2008. Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. *Nat Genet* **40**: 211–216.

Lam HYK, Mu XJ, Stütz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. 2010. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* **28**: 47–55.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5**: 2113–2144.

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

Londin ER, Keller MA, D'Andrea MR, Delgrosso K, Ertel A, Surrey S, Fortina P. 2011. Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. *BMC Genomics* **12**: 464.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**: 18.

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

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

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

Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, et al. 2012. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**: 1431–1442.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**: 59–65.

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al. 2013. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* **23**: 749–61.

Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC, et al. 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* **11**: R52.

Pang AWC, Migita O, Macdonald JR, Feuk L, Scherer SW. 2013. Mechanisms of Formation of Structural Variation in a Fully Sequenced Human Genome. *Hum Mutat* **34**: 345–354.

Ramu A, Noordam MJ, Schwartz RS, Wuster A, Hurles ME, Cartwright RA, Conrad DF. 2013. DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat Methods* **10**: 985–7.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444**: 444–454.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Roerink SF, van Schendel R, Tijsterman M. 2014. Polymerase theta-mediated end joining of replication-associated DNA breaks in C. elegans. *Genome Res* **24**: 954–62.

Romeo G, Hassan HJ, Staempfli S, Roncuzzi L, Cianetti L, Leonardi A, Vicente V, Mannucci PM, Bertina R, Peschle C. 1987. Hereditary thrombophilia: identification of nonsense and missense mutations in the protein C gene. *Proc Natl Acad Sci U S A* **84**: 2829–2832.

Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet* **13**: 824–824.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. 2007. Strong association of de novo copy number mutations with autism. *Science* **316**: 445–449.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61**: 437–455.

Stewart C, Kural D, Strömberg MP, Walker JA, Konkel MK, Stütz AM, Urban AE, Grubert F, Lam HYK, Lee WP, et al. 2011. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* **7**: e1002236.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**: 848–853.

The Encode Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.

Thung DT, de Ligt J, Vissers LE, Steehouwer M, Kroon M, de Vries P, Slagboom EP, Ye K, Veltman JA, Hehir-Kwa JY. 2014. Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biol* **15**: 488.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37**: 727–732.

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

Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, et al. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**: 539–543.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**: 2865–71.

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

Supplemental Materials

Supplemental Figures 1 and 2.

Supplemental Tables 1-5.

Supplemental Methods.

Table 1. Indel classes and mechanisms

Class	Example[1]	Observations	Possible Mechanisms	Sequence Features
Homopolymer Run	Ref. CTGAGGAAGAG**T**TTTTTTTACA De Novo CTGAGGAAGAG-TTTTTTTACA	21 insertions 7 deletions	Polymerase slippage	Repeat context
Tandem Repeats	Ref. CTACCCCAGGC**AGAG**AGAGAAA De Novo CTACCCCAGGC----AGAGAAA	8 insertions 19 deletions	Polymerase slippage	Repeat context
Copy Count Changing	Ref. CAGAAGG----TAGCTAGTCAG De Novo CAGAAGG**TAGC**TAGCTAGTCAG	37 insertions 74 deletions	Polymerase slippage	Local copy count change
Non Copy Count Changing	Ref. CTAAAGGGCAG**TC**TTGCAAAAG De Novo CTAAAGGGCAG--TTGCAAAAG	8 insertions 90 deletions	NHEJ[2]	Blunt or microhomolgy at breakpoints
	Ref. AGTCAAAAACCA**AAG**TTTTGAA De Novo AGTCAAAAACCA---TTTTGAA	8 deletions	NHEJ[2] / hairpin loop	Palindrome (≥6bp) in surrounding context (≤20bp)
	Ref. GGGGAGAA**TTGAGAC**TTGATCA De Novo GGGGAGAA-------TTGATCA	5 deletions	NHEJ[2] / MMEJ[3] / replication slippage	Microhomology ≥4bp at breakpoints
Complex	Ref. ACTCACAAAAA**AA**TTTTTTTCC De Novo ACTCACAAAAA-**T**TTTTTTTCC	2 variants	Polymerase slippage	Repeat context
	Ref. CACATGGGCT**TCC**-----TGTC De Novo CACATGGGCT**GGAGCCCA**TGTC	8 variants	SD-MMEJ[4] TMEJ[5]	Palindromic or templated insertion
	Ref. CCAAAGTGCT**GG**GATTACAGGC De Novo CCAAAGTGCT**C**-GATTACAGGC	4 variants	Unknown	None

[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

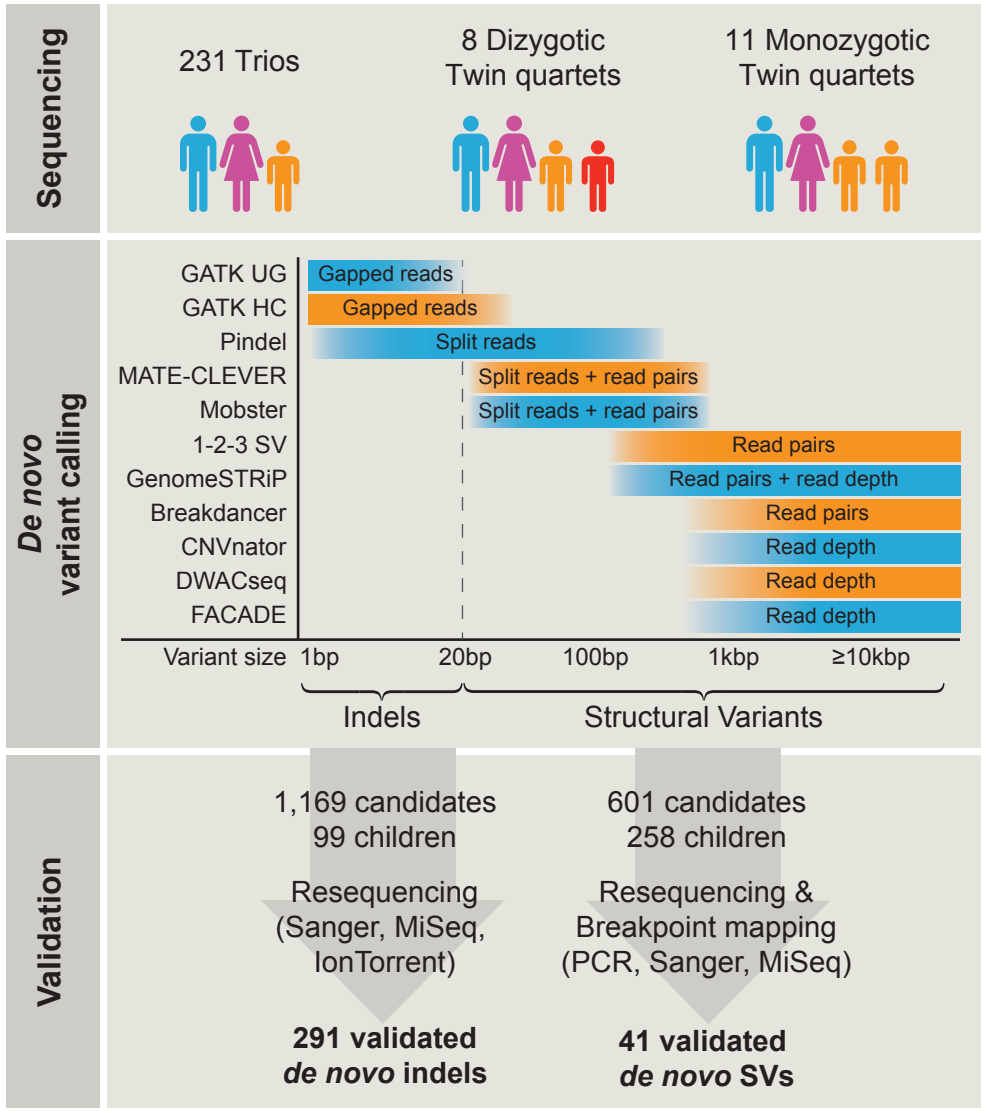

Sequencing
231 Trios
8 Dizygotic Twin quartets
11 Monozygotic Twin quartets
De novo variant calling
GATK UG
GATK HC
Pindel
MATE-CLEVER
Mobster
1-2-3 SV
GenomeSTRiP
Breakdancer
CNVnator
DWACseq
FACADE
Gapped reads
Gapped reads
Split reads
Split reads + read pairs
Split reads + read pairs
Read pairs
Read pairs + read depth
Read pairs
Read depth
Read depth
Read depth
Variant size
1bp
20bp
100bp
1kbp
≥10kbp
Indels
Structural Variants
Validation
1,169 candidates
99 children
Resequencing
(Sanger, MiSeq,
IonTorrent)
291 validated
de novo indels
601 candidates
258 children
Resequencing &
Breakpoint mapping
(PCR, Sanger, MiSeq)
41 validated
de novo SVs

Figure 1

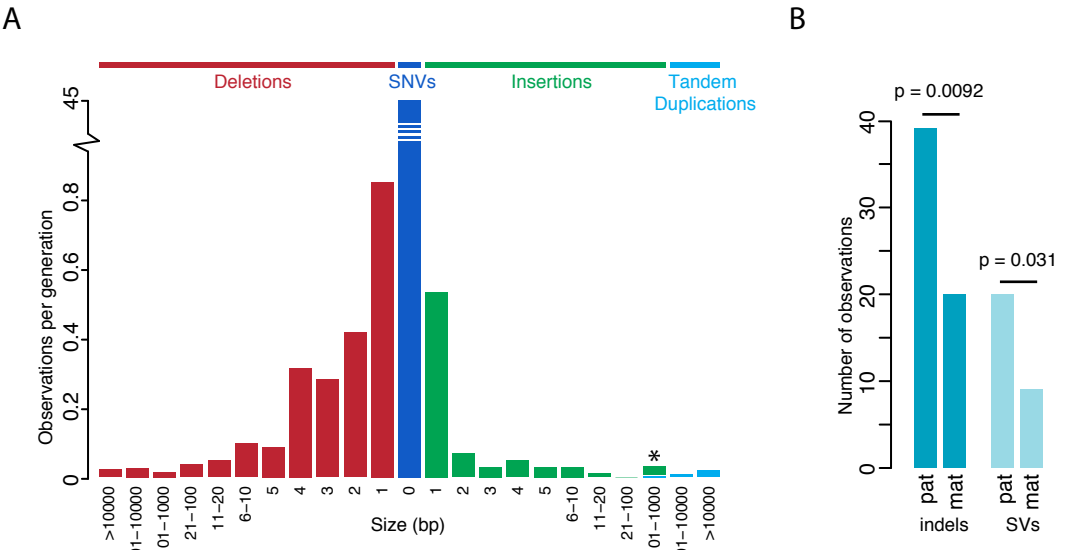

A
Deletions
SNVs
Insertions
Tandem Duplications
Observations per generation
0
0.2
0.4
0.6
0.8
45
>10000
1001−10000
101−1000
21−100
11−20
6−10
5
4
3
2
1
0
1
2
3
4
5
6−10
11−20
21−100
101−1000
1001−10000
>10000
*
Size (bp)
B
p = 0.0092
p = 0.031
Number of observations
0
10
20
30
40
pat
mat
pat
mat
indels
SVs

Figure 2

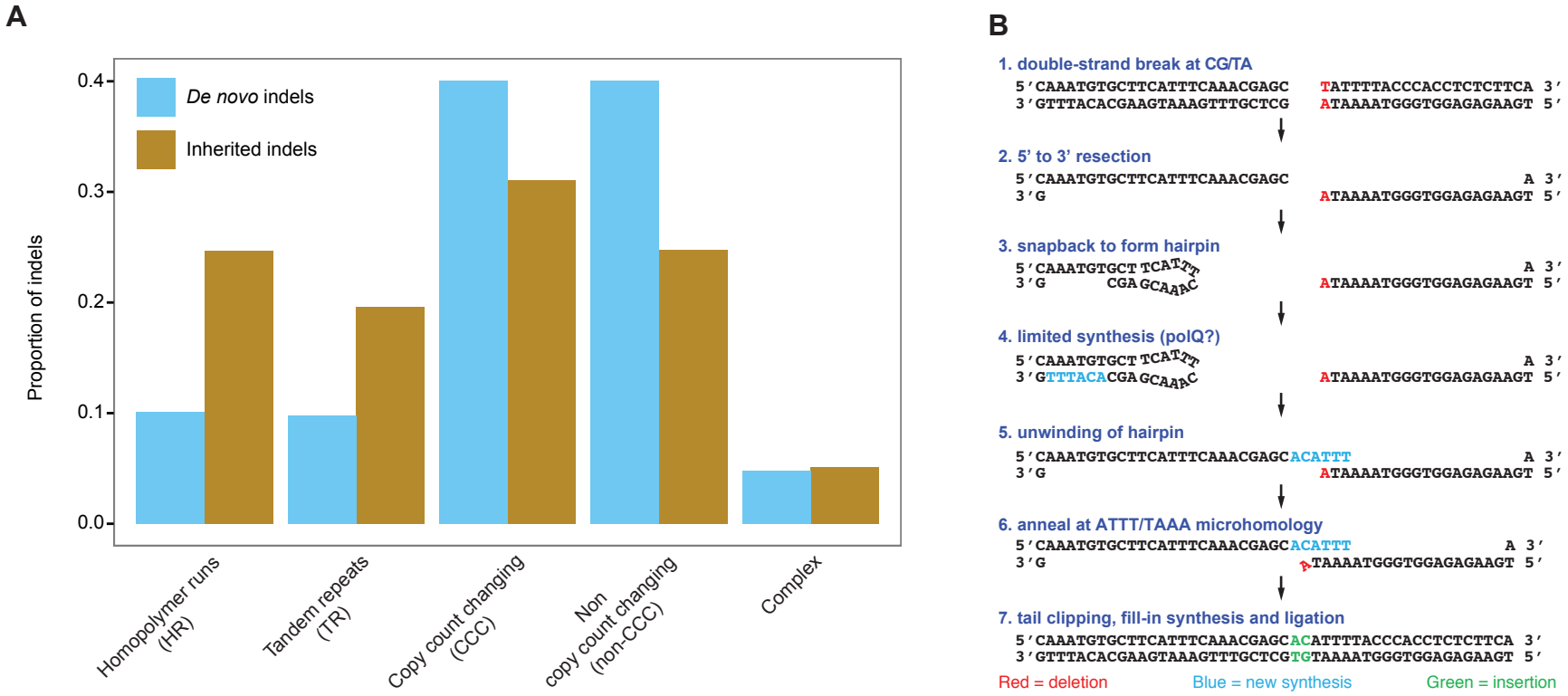
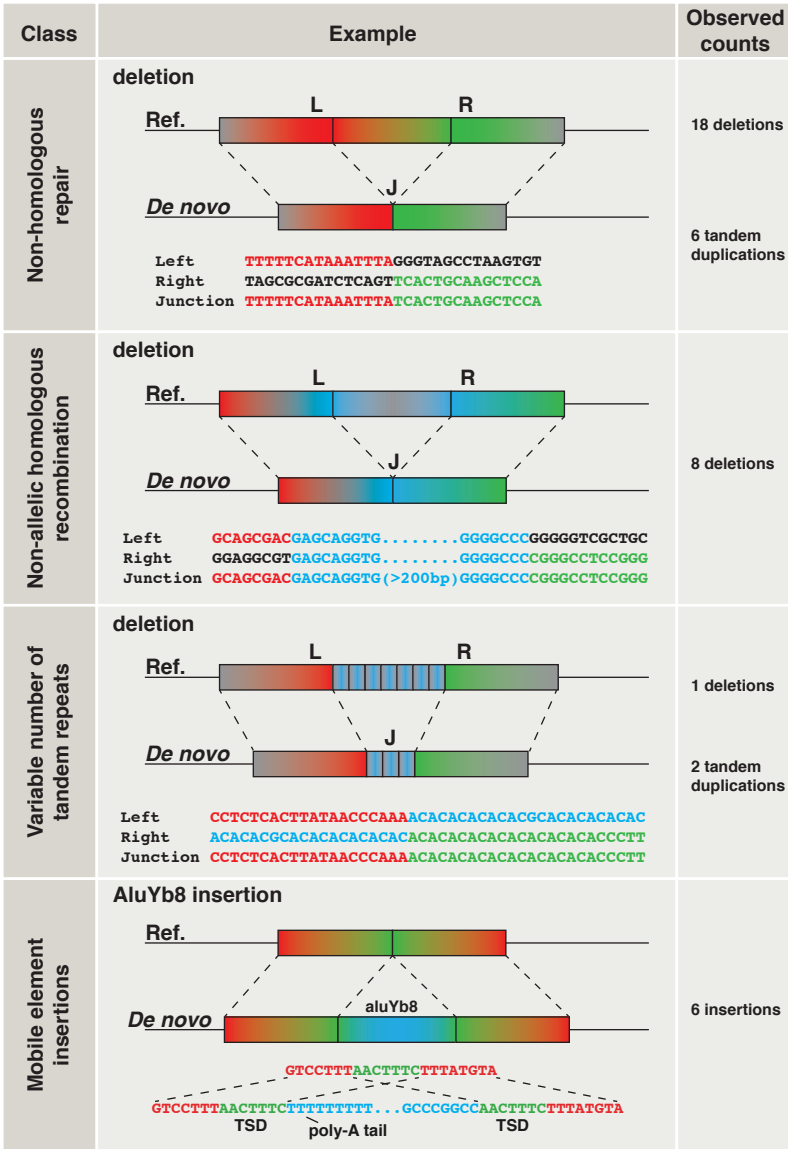

A
0.4
0.3
0.2
0.1
0.0
Proportion of indels
De novo indels
Inherited indels
Homopolymer runs (HR)
Tandem repeats (TR)
Copy count changing (CCC)
Non copy count changing (non-CCC)
Complex
B
1. double-strand break at CG/TA
5′CAAATGTGCTTCATTTCAAACGAGC TATTTTACCCACCTCTCTTCA 3′
3′GTTTACACGAAGTAAAGTTTGCTCG ATAAAATGGGTGGAGAGAAGT 5′
2. 5′ to 3′ resection
5′CAAATGTGCTTCATTTCAAACGAGC A 3′
3′G ATAAAATGGGTGGAGAGAAGT 5′
3. snapback to form hairpin
5′CAAATGTGCT TCATTT A 3′
3′G CGA GCAAAC ATAAAATGGGTGGAGAGAAGT 5′
4. limited synthesis (polQ?)
5′CAAATGTGCT TCATTT A 3′
3′GTTTACACGA GCAAAC ATAAAATGGGTGGAGAGAAGT 5′
5. unwinding of hairpin
5′CAAATGTGCTTCATTTCAAACGAGCACATTT A 3′
3′G ATAAAATGGGTGGAGAGAAGT 5′
6. anneal at ATTT/TAAA microhomology
5′CAAATGTGCTTCATTTCAAACGAGCACATTT A 3′
3′G pTAAAATGGGTGGAGAGAAGT 5′
7. tail clipping, fill-in synthesis and ligation
5′CAAATGTGCTTCATTTCAAACGAGCACATTTTACCCACCTCTCTTCA 3′
3′GTTTACACGAAGTAAAGTTTGCTCGTGTAAAATGGGTGGAGAGAAGT 5′
Red = deletion
Blue = new synthesis
Green = insertion

Figure 3

A


Class
Example
Observed counts
Non-homologous repair
deletion
Ref.
L
R
De novo
J
Left TTTTTCATAAATTTAGGGTAGCCTAAGTGT
Right TAGCGCGATCTCAGTTCACTGCAAGCTCCA
Junction TTTTTCATAAATTTATCACTGCAAGCTCCA
18 deletions
6 tandem duplications
Non-allelic homologous recombination
deletion
Ref.
L
R
De novo
J
Left GCAGCGACGAGCAGGTG.....GGGGCCCGGGGGTCGCTGC
Right GGAGGCGTGAGCAGGTG.....GGGGCCCCGGGCCTCCGGG
Junction GCAGCGACGAGCAGGTG(>200bp)GGGGCCCCGGGCCTCCGGG
8 deletions
Variable number of tandem repeats
deletion
Ref.
L
R
De novo
J
Left CCTCTCACTTATAACCCAAAACACACACACACGCACACACACAC
Right ACACACGCACACACACACACACACACACACACACACACCCTT
Junction CCTCTCACTTATAACCCAAAACACACACACACACACACCCTT
1 deletions
2 tandem duplications
Mobile element insertions
AluYb8 insertion
Ref.
De novo
aluYb8
GTCCTTTAACTTTCTTTATGTA
GTCCTTTAACTTTCTTTTTTTTT...GCCCGGCCAACTTTCTTTATGTA
TSD
poly-A tail
TSD
6 insertions

B

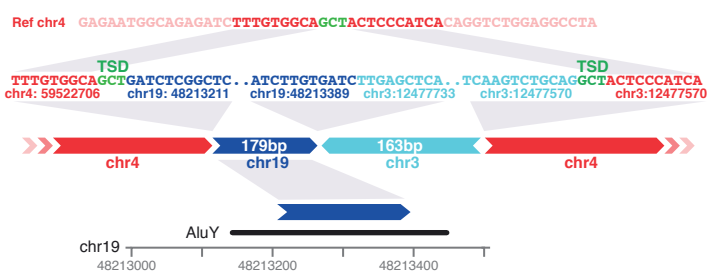

Ref chr4 GAGAATGGCAGAGATCTTTGTGGCAGCTACTCCCATCACAGGTCTGGAGGCCTA
TSD
TSD
TTTGTGGCAGCTGATCTCGGCTC..ATCTTGTGATCTTGAGCTCA..TCAAGTCTGCAGGCTACTCCCATCA
chr4: 59522706
chr19: 48213211
chr19:48213389
chr3:12477733
chr3:12477570
chr3:12477570
179bp
163bp
chr4
chr19
chr3
chr4
AluY
chr19
48213000
48213200
48213400

Figure 4

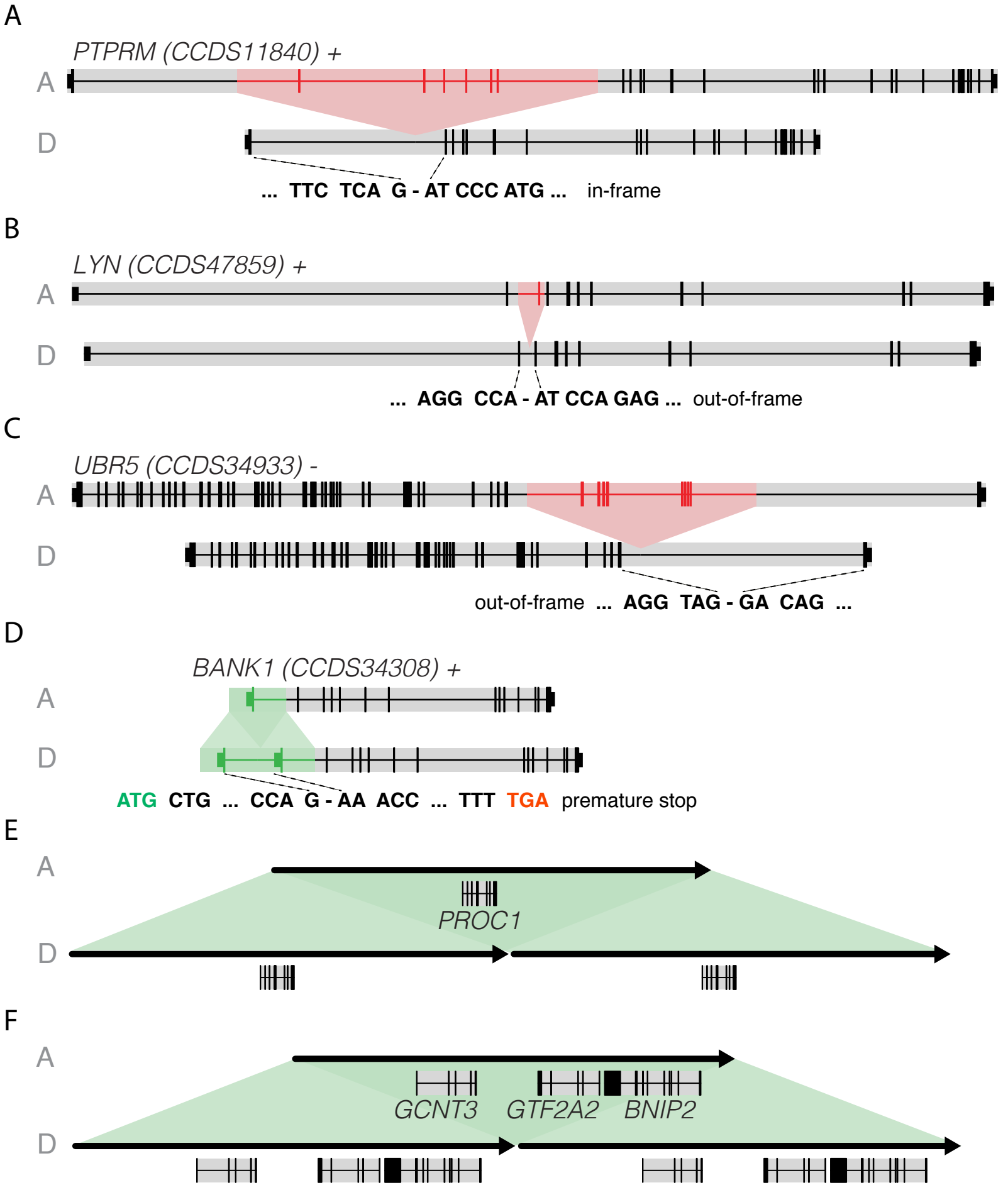

A
PTPRM (CCDS11840) +
A
D
... TTC TCA G - AT CCC ATG ... in-frame
B
LYN (CCDS47859) +
A
D
... AGG CCA - AT CCA GAG ... out-of-frame
C
UBR5 (CCDS34933) -
A
D
out-of-frame ... AGG TAG - GA CAG ...
D
BANK1 (CCDS34308) +
A
D
ATG CTG ... CCA G - AA ACC ... TTT TGA premature stop
E
A
PROC1
D
F
A
GCNT3
GTF2A2
BNIP2
D

Figure 5

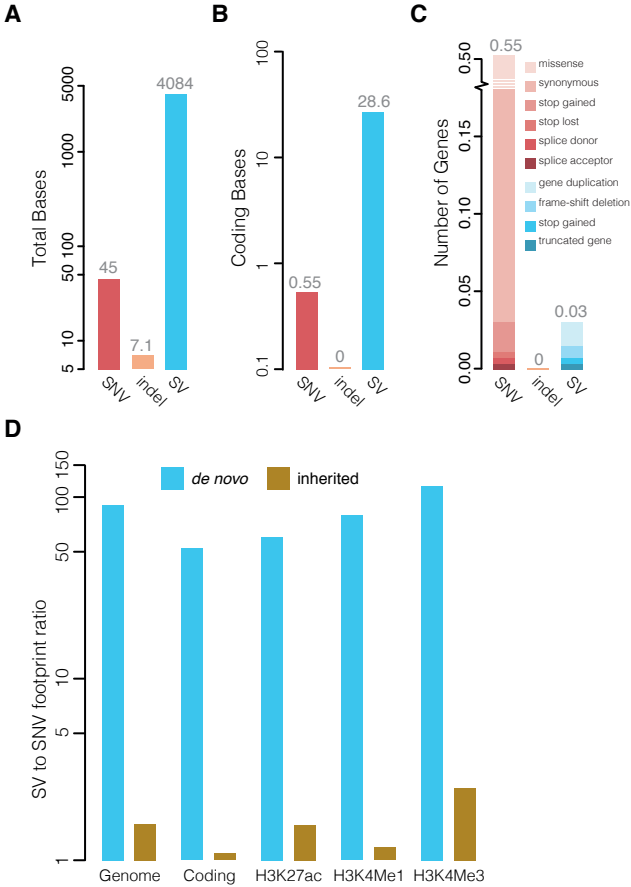

A
Total Bases
5
10
50
100
1000
5000
45
7.1
4084
SNV
indel
SV
B
Coding Bases
0.1
1
10
100
0.55
0
28.6
SNV
indel
SV
C
Number of Genes
0.00
0.05
0.10
0.15
0.50
0.55
0
0.03
SNV
indel
SV
missense
synonymous
stop gained
stop lost
splice donor
splice acceptor
gene duplication
frame-shift deletion
stop gained
truncated gene
D
SV to SNV footprint ratio
1
5
10
50
100
150
de novo
inherited
Genome
Coding
H3K27ac
H3K4Me1
H3K4Me3

Figure 6