de novo variant calling identifies cancer mutation profiles in the 1000 Genomes Project

Jeffrey Ng[1], Pankaj Vats[2], Elyn Fritz-Waters[3], Evin M. Padhi[1], Zachary L. Payne[1], Shawn Leonard[3], Stephanie Sarkar[1], Marc West[2], Chandler Prince[3], Lee Trani[4], Marshall Jansen[3], George Vacek[2], Mehrzad Samadi[2], Timothy T. Harkins[2], Craig Pohl[3], Tychele N. Turner[1,*]

1. Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA
2. NVIDIA
3. Research Infrastructure Services, Washington University School of Medicine, St. Louis, MO 63110, USA
4. McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63110, USA

*Correspondence to: tychele@wustl.edu

Tychele N. Turner, Ph.D.

Washington University School of Medicine

Department of Genetics

4523 Clayton Avenue

Campus Box 8232

St. Louis, MO 63110

ABSTRACT

Detection of *de novo* variants (DNVs) is critical for studies of disease-related variation and mutation rates. We developed a GPU-based workflow to call DNVs, using 602 trios from the 1000 Genomes Project as a control. We detected 445,711 DNVs, having a bimodal distribution, with peaks at 200 and 2000 DNVs. The excess DNVs are cell line artifacts that are increasing with cell passage. Reduction in DNVs at CpG sites and in percent of DNVs with a paternal parent-of-origin with increasing number of DNVs supports this finding. Detailed assessment of individual NA12878 across multiple genome datasets from 2012 to 2020 reveals increasing number of DNVs over time. Mutation signature analysis across the set revealed individuals had either 1) age-related, 2) B-cell lymphoma, or 3) no prominent signatures. Our approach provides an important advancement for DNV detection and shows cell line artifacts present in lymphoblastoid cell lines are not always random.

INTRODUCTION

The 1000 Genomes Project is a data resource for the study of genetic variation that includes individuals from diverse genetic ancestries (*1, 2*). In this study, we present the first ever assessment of new mutations, termed *de novo* variants (DNVs) that are only found in children and not their parents, represented in this collection. DNVs are important for assessing mutation rates (*3*) and have been shown to contribute to human disease (autism (*4-12*), epilepsy (*13, 14*), intellectual disability (*15-18*), congenital heart disorders (*19-21*)). This is a critical reason to assess these variants in the 18 populations represented in this data (Figure S1). Moreover, the 1000 Genomes Project has been utilized in many applications as a control resource for filtering of genetic variation by allele frequency and/or variant presence-absence in the dataset (*22*).

One complicating factor of DNV assessment in this resource is the fact that sequencing data is generated from DNA isolated from lymphoblastoid cell lines (LCLs) (*2*) as opposed to primary tissue. Epstein-Barr Virus is used to make these LCLs and passaging over time enables the accumulation of cell line artifacts. These artifacts can complicate variant filtration schemes and the utility of this data as a frequency control. As opposed to a random accumulation of mutations in each individual, we found that DNVs fit one of three possible mutation profiles: 1) age-related (similar to true DNVs); 2) B-cell lymphoma related; and 3) an absence of specific profile. The most problematic of these is the similarity to B-cell lymphomas and it would be imperative that this data not be used as a control in the context of the study of these and related cancers.

A second aspect of this paper is the development of a rapid approach for identifying DNVs that will be applicable to any short-read whole-genome sequencing data. Typically calling of DNVs from raw sequence data to final calls can take days to weeks. Multiple *de novo* workflows exist that primarily rely on central processing unit (CPU)-based approaches (*4-9, 11, 12, 14-17, 19, 23-34*). These DNV workflows employ different approaches including strict filtering, utilizing multiple variant callers as opposed to using only one, machine-learning, and incorporation of genotypic information at other sites around the genome. Overall, there is no community consensus on a standard method for detecting DNVs. It is imperative that this process be streamlined and flexible to enable broad adoption across the community. The approach we developed to accelerate this analysis was the utilization of graphics processing units (GPUs), enabling highly-parallelized analyses. Herein, we developed a DNV workflow and integrated it into NVIDIA Parabricks (*35*) software for a drastic acceleration and reduced run time from alignment files, for each trio, to final DNVs in less than one hour. This type of computational upgrade for specific workflows are essential for the future of genomics as projects become larger in scale.

RESULTS

Rapid DNV calling with GPUs

Our DNV workflow utilized the existing features of the NVIDIA Parabricks software including fast GATK (*36*) HaplotypeCaller and Google's DeepVariant (*37*) gvcf generation. The run times are ~40 minutes per sample on a 4 GPU node and can be run in parallel on all three family members in the parent-child trio. Genotyping of the trio is quick through the use of GLnexus (*38*). Finally, our DNV workflow runs in ~1 hour with speedups at all steps with parallelization providing a clear advantage over CPU-based approaches (Figure 1A).

To benchmark our DNV workflow, we tested it on a monozygotic twin pair with WGS data derived from blood DNA. These individuals should share the same DNVs from generation in the germline. However, they may differ at some sites if DNVs occur in a post-zygotic, somatic manner. The twins shared 75 DNVs and contained 83 and 81 DNVs, respectively (Figures 1B and 1C). The percent CpG was 19.3% and 17.2%, respectively and in line with estimates ~20% (Figures 1B and 1C). As this monozygotic twin pair was discordant for the phenotype of autism, we also tested whether there were any protein-coding DNV differences between the two twins. These would potentially be relevant for autism, but there were no such differences.

We next assessed DNVs with our workflow in four trios from the 1000 Genomes Project (Figure 1D). Two were chosen at random (i.e., HG00405, HG00408) and two were chosen because they were "famous" trios assessed in many other studies (i.e., NA12878 (*26, 39*), NA19240 (*26*)). One of these trios (HG00405) had 70 DNVs and a CpG percent of 21.4 as we would have expected from DNA derived from blood. To our surprise, the other trios had varying numbers of DNVs from 592 to 2,230 with NA12878 (arguably the most studied individual in the 1000 Genomes Project) having the most DNVs. With the increase in DNVs the CpG percent dropped considerably down to ~10%. We also assessed 3,598 of the DNVs from the four trios by visual inspection of the reads in each family member (Table S1) and found that 93.6% of the variants appeared to be true DNVs, 4.9% were inherited, and 1.5% were low confidence calls.

Excess of DNVs observed in trio dataset

We applied our DNV workflow to all 602 trios (Table S2) in the 1000 Genomes Project and detected 445,711 total DNVs with 740.4 ± 968.0 DNVs per individual (Tables S3 and S4). There was a clear bimodal distribution wherein some individuals contained an excess of DNVs (Figure 2A). We split the data into two groups: individuals having less than or equal to 100 DNVs (n = 123) and individuals had greater than 100 DNVs (n = 479). The individuals with less than or equal to 100 DNVs were in line with expectation for expected DNVs. While we expected there to be no difference by ancestry we sought to see if this could somehow be a factor (Figure 2B). The population with the most DNVs was the CEU having on average 1,688 DNVs. We hypothesized that this may be because the CEU is one of the oldest cohorts in the 1000 Genome Project dating back to the HapMap project (*40*) and these individuals may have cell lines that have been cultured more over time than other populations.

Since we thought the individuals with excess of DNVs represented cell line artifacts, we assessed two main features of typical DNVs. These were the percent of DNVs at CpG locations and the percent of DNVs arising on the paternal chromosome. It has been well-established that the percent of DNVs at CpG should be ~20% (*3, 8*) and the percent of DNVs arising on the paternal

chromosome should be ~80% (*41*). We saw that overall, 13.7 ± 4.4% of DNVs per individual occurred at CpG sites. In the individuals with less than or equal to 100 DNVs this rose to 17.4 ± 5.2% and in families with greater than 100 DNVs it fell to 12.7 ± 3.6% (Figure 2C). The percent of DNVs that were phase-able for parent-of-origin was 37.2 ± 7.5% (Figure S2). Of the phased variants, 61.3 ± 11.3% were on the chromosome of paternal origin (Table S5). In the families with less than or equal to 100 DNVs this rose to 72.0 ± 8.5% and in the families with greater than 100 DNVs it fell to 58.6 ± 10.3%. The drop leveled off to ~50% in the individuals with the most DNVs (Figure 2D). This showed that the individuals with DNVs at expectation for counts also behaved more like true DNVs in regard to CpG percentage and percent arising on the paternal chromosome.

DNVs increase over time

We utilized the fact that the 1000 Genomes Project individual NA12878 has been studied and sequenced multiple times over the past ten years by WGS (*2*) (SRA identifiers: SRR944138 and SRR952827). Presumably, across time, the utilization of NA12878 has required additional culturing of this cell line, and potentially even by different laboratories. We aggregated five Illumina WGS datasets from this individual, downsampled them to ~30x coverage, and assessed them by our DNV workflow. The data for this individual ranged from the year 2012 to the year 2020 and we found that the 2012 experiment had the least DNVs (n = 2,060) and the 2020 experiment had the most DNVs (n = 2,230) (Figure 3A). Overall, the five replicates had a large overlap of DNVs (n = 1,820) across all samples. These shared DNVs constitute what were present in the ancestor of all the cell line replicates. Mutations not shared by all five replicates are sometimes shared by a subset of the replicates and are sometimes unique to the replicate. To formally assess the ancestral state, we built a phylogenetic tree based only on the DNVs and saw that the farthest replicates from each other in the tree were the 2012 and 2020 replicates (Figure 3B).

Genomes with cancer mutation profiles

We used mutation profile analysis (*42*) (Table S6) to determine whether the DNVs identified in individuals from the 1000 Genomes project had any certain characteristics. For this analysis, we utilized a method that would enable comparisons to known mutational profiles that are either age-related (reminiscent of true DNVs) or are seen in cancers (Figure 4A and Figure 4B). There were 186 individuals that had a strong contribution of an age-related signature (Signature 1A, Signature 1B). To our surprise, the other contributing signatures in individuals were primarily those associated with B-cell lymphomas (Signature 5, Signature 9 and Signature 17) in 241 individuals. This was intriguing because lymphoblastoid cell lines are generated from B-cells that are infected with Epstein Barr Virus and demonstrates that new mutations are not arising in a random manner. Rather they are being generated in a manner consistent with the development of cancer in the same cell type.

We further sought to determine what the mechanism was for the generation of a B-cell lymphoma-like state. First, we determined whether there was high rate of aneuploidies in the cell lines. By digital karyotyping (Table S7) we found that 595 individuals (98.8%) had a typical chromosome complement (46,XX or 46,XY), four were missing a sex chromosome (45,X0), one was 47,XXY, one had three chromosome 12 (47,XY), and one had three chromosome 9 (47,XY). This demonstrated that while these aneuploidies are occurring in some cell lines, they are probably not the main driving factor. Next, we looked at DNVs in genes involved in DNA repair and found

17 individuals contained a missense or loss-of-function in one of these genes (Table S8). Individuals with B-cell lymphoma profiles and disruptive mutations in DNA repair genes included mutations in the following genes *FANCF* (HG01126), *MUS81* (NA10838), *POLB* (NA10838), *POLD1* (NA19677), *POLE* (HG01096), *RAD18* (NA12864) (Figure 4C), *RAD51* (HG02683), *RPA4* (HG02630), and two individuals with mutations in *FANCA* (HG02841, HG03200) and *WRN* (HG04115, NA19161), respectively (Table 1). Third, we looked at Epstein Barr Virus load in each of the genomes (Table S9) and found that there was a weak, yet significant, correlation with the number of DNVs ($p = 2.32 \times 10^{-5}$, $r = 0.17$) (Figure S3). By visual inspection of phased variation in all individuals we also identified individuals with clusters of mutations (e.g., NA07048, Figure 4D, Figure S4).

Excess of DNVs in IGLL5

We applied a multi-phase approach to determine if there were any genes with enrichment of protein-coding DNVs in individuals with greater than 100 DNVs. In the first phase, we tested whether there was genome-wide significance for enrichment of protein-coding DNVs (missense, loss-of-function) in any specific genes. By application of two methods (chimpanzee-human, denovolyzeR), we identified 29 significant genes (*ARMC3*, *BCL2*, *BCR*, *C6orf15*, *CCDC168*, *CSMD3*, *EGR3*, *EXO1*, *HLA-B*, *HLA-C*, *IGLL5*, *KMT2D*, *LINGO2*, *LTB*, *MEOX2*, *MUC16*, *MUC22*, *NPAP1*, *PCLO*, *PRPF40A*, *RUNX1T1*, *SGK1*, *STRAP*, *TMEM232*, *TNXB*, *TTN*, *WDFY4*, *XIRP2*, *ZNF488*) with excess of DNVs (Table S10). In the second phase, we tested these 29 genes to see whether there were significantly more protein-coding DNVs in individuals with greater than 100 DNVs in comparison to individuals with less than or equal to 100 DNVs. Only *IGLL5* was significant in this comparison (1.79×10^{-3}) (Table S10, Table S11, Figure 4E).

DNVs identified in clinically-relevant variants

We tested whether any of the DNVs detected were already known to be pathogenic or likely-pathogenic in the Clinvar (*43*) database (Table 1). There were 15 mutations meeting these criteria (Table S12). We rescored these variants using Franklin software to assess their pathogenicity and found that 13 were also pathogenic or likely-pathogenic by this approach. Twelve of these variants were associated with described phenotypes in Clinvar. These included a missense variant in *SOS1* involved in Noonan syndrome, a missense variant in *SCN2A* involved in seizures, a stop gained variant in *UNC80* involved in a syndrome with hypotonia, intellectual disability, and characteristic facies, a missense variant in *THRB* involved in thyroid hormone resistance, a missense variant in *PKHD1* involved in polycystic kidney disease, a stop-gained in *ERCC6* involved in Cockayne syndrome, a stop-gained in *ANO5* involved in gnathodiaphyseal dysplasia, a stop-gained in *PHF21A* involved in inborn genetic disease, a missense in MYO7A in Usher syndrome type 1, a stop-gained in *ROBO3* in Gaze palsy with progressive scoliosis, a missense in *COL4A1* involved in inborn genetic disease, and a missense in *POLG* involved in POLG-related disorder.

DISCUSSION

While the 1000 Genomes Project data has been extensively studied in the past, there has been no previous cross-cohort assessment of DNVs. This limitation is primarily because family-based sequencing was not available until 2020 when this cohort was sequenced by high-coverage short-read whole-genome sequencing ten years after the initial ground-breaking publication on the 1000 Genomes Project (*44*). Determining DNV profiles across this dataset of diverse individuals is critical for assessment of mutation rates in the human population, while also providing a more complete catalog of all genetic variants within these individuals. The decision to sequence these individuals using DNA derived from lymphoblastoid cell lines was a practical one. However, it opened the door to the possibility of cell line artifacts, while simultaneously introducing a dynamic aspect to this extensive set of controls. As control samples, the cell lines that were used as the inputs for the 1000 Genome Project are still actively used across laboratories, acting as matched controls for workflows to known sets of variants. The large distribution of DNVs across the 1000 Genomes Project suggest that a subset of the control source inputs are dynamic, and in some cases, harbor a spectrum of genetic variants associated with B-cell lymphomas or named clinical syndromes. Laboratories using control samples from the 1000 Genomes Project should account for both the presence and dynamic nature of the reported DNVs and in some cases may consider changing which control samples to use within the laboratory to avoid any of the associated issues with the presence of DNVs. Additionally, other public efforts to establish reference data sets using cell lines should consider the impacts of DNVs on their project design.

We utilized a novel and accelerated analysis workflow to detect DNVs from short-read, whole-genome sequencing data. In total, we identified 445,711 DNVs in the 602 children assessed in this study. We provide family-level VCFs, DNV calls, and phased DNV results for the 602 trios in this study as a public community resource (Globus endpoint: "Turner Lab at WashU - DNV in 1000 Genomes Paper", direct link: https://app.globus.org/file-manager?origin_id=3eff453a-88f4-11eb-954f-752ba7b88ebe&origin_path=%2F). Originally, it was assumed that the DNV's across the 1000 Genomes Project would have been random and minimal, and yet only 20% of the offspring (123 children) have a number of DNVs around expectation (< 100) and the remainder have an excess of DNVs with the most extreme case being an individual (HG02683) having 11,219 DNVs. We hypothesized that the excess DNVs were cell line artifacts and found multiple lines of evidence to support this hypothesis, including a reduction in the percent of DNVs at CpG as well as the reduction in percent phased to the paternal parent-of-origin chromosome with increasing DNVs, respectively. A detailed analysis of individual NA12878, who has been studied various times over the years, revealed increasing DNVs in the more recently sequenced samples also supporting this hypothesis. The changes in the DNVs for NA12878 suggest the dynamic nature of the DNVs, demonstrating that the number is increasing over time.

When mutational signature analysis was performed on this new set of DNVs, the most common mutation signatures were those seen in B-cell lymphomas. This is important as the lymphoblastoid cell lines are generated from B-cells and points to a non-random accumulation of mutations that are in line with the development of cancer in this cell type. In particular, we identified mutations in key DNA repair genes as well as a statistically significant excess of DNVs in *IGLL5 (45, 46)*. This gene is found to be mutated in B-cell lymphomas and protein-coding DNVs are identified in 27 individuals in this cohort; all of which have >100 overall DNVs. From our work, we identify two contributing factors causing these higher levels of DNVs, one is the

mutation of DNA repair genes while the second is an excess of Epstein-Barr Viral load. Future work using long-read sequencing and *de novo* assemblies will be imperative to identify complete viral integration in these genomes as integration sites can have impacts on cell line stability. One unexpected consequence of B-cell lymphoma mutation signatures in some individuals from the 1000 Genomes Project would be a new pathway to study the mechanisms and biology of the development of this cancer.

In addition to the DNA repair gene DNVs, we identified fifteen pathogenic or likely-pathogenic DNVs that had already been implicated in a database of clinical variation (Clinvar). This calls into question the use of the 1000 Genomes Project data as a control for both B-cell lymphomas and more generally for DNVs identified in clinical patients. More importantly, the extensive spectrum of DNVs that can appear in a cell line call into question the use of control samples derived from lymphoblastoid cell lines. Currently, to our knowledge the Genome in a Bottle and Human Pangenome Reference Consortium (HPRC) are building reference databases and pangenomes using DNA from lymphoblastoid cell lines. Although it does seem that the use of blood for some samples was at least initially discussed for the HPRC (https://www.genome.gov/Pages/Research/Sequencing/Meetings/HGR_Webinar_Summary_March1_2018.pdf), it does appear the project has defaulted to using lymphoblastoid cell lines. We find it is imperative that these efforts consider utilizing native DNA isolated from blood as the source or utilize a family-based design to identify and remove DNVs. In this way, the highest quality references can be built that will stand the test of time. Finally, we recommend that much like the Simons Simplex Collection, that studies assessing DNVs in individuals with a particular phenotype of interest, also sequence DNA from blood cells and not DNA post-culturing of lymphoblastoid cell lines.

Online Methods

Trio dataset

A total of 602 trios from the 1000 Genomes Project were sequenced at the New York Genome Center as described previously (*2*). The aligned data files (crams) are located at http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/1000G_2504_high_coverage.sequence.index and http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/1000G_698_related_high_coverage.sequence.index . Description of the 602 trios is found in Supplemental Table S2.

Single-nucleotide variant and insertion/deletion calling

The NVIDIA Parabricks program version 3.0.0 was utilized to call single-nucleotide variants (SNVs) and small insertions/deletions (indels) with GATK (*36*) version 4.1.0 and Google's DeepVariant (*37*) version 0.10. For each individual, a GVCF was generated for these two variant callers. The GVCFs were then genotyped, on a per trio basis, using the GLnexus (*38*) version 1.2.6 genotyper. Post-calling, we checked the counts of all variants and heterozygous variants per chromosome in each individual (Figure S5).

de novo variant calling

DNVs were called by identifying all DNVs in GATK and DeepVariant based on the parent and child genotypes, respectively. The intersection of these DNVs was then identified from the two callers and filtering was carried out as follows: genotype quality greater than 20, depth of >= 10, an allele balance > 0.25, and no presence of the DNV in any reads in the parents. DNVs in low complexity regions, centromeres, and recent repeats were removed from further analysis. To assess the DNVs, we manually scored 3980 sites with SAMtools version 1.9 tview. To score these sites, we looked at the first column (variant location in tview images) of both parents and the proband sample to see what mutations were present. If there was any mutation in the first column of the mother or father, regardless of quality, that matched the main mutation in the proband's first column, then we denoted the mutation as maternal, paternal, or both depending on whether it was the mother's mutation that matched the proband or the fathers or both parents. If the main mutation in the first column of the parental samples did not match the proband's mutation, then we knew this sample would be de novo, thus verifying our results.

Phasing of de novo variants

We utilized Unfazed v1.0.2 (https://github.com/jbelyeu/unfazed) (*47*) to phase the *de novo* variants in our study with regard to the parent-of-origin chromosome. First, a bed file containing *de novo* variants was generated for each individual. Second, the *de novo* bed file, DeepVariant full genome trio VCF, and the alignment files for all trio members were run through Unfazed. Since Unfazed uses different approaches to phasing on the X chromosome in males and females, we only focus on phased variants on the autosomes in this study.

NA12878 additional datasets

We identified additional high-coverage whole-genome sequencing data from NA12878 from the SRA (https://www.ncbi.nlm.nih.gov/sra) and other sources. These included SRA data SRR944138 and SRR952827 both from 2013, McDonnell Genome Institute data gerald_HFKWMDSXX and H_IJ-NA12878 both from 2018, and the high-coverage data from

2020. To avoid differences due to coverage, we downsampled all datasets to 30x using SAMtools. All data was re-mapped to build 38 using SpeedSeq (*48*) version 0.1.2 and run through the de novo workflow using the NA12891 and NA12892 parental WGS data from 2020. We again did a count check for total and heterozygous variants per chromosome (Figure S6).

Phylogenetic tree of de novo variants

To assess the differences between different NA12878 replicates we built a multi-sequence fasta file where each fasta represents the aggregate of all possible DNVs identified in this individual. The specific steps to build the tree were as follows: 1) we first merged the samples together and converted the genotypes for each de novo mutation from 0/0 or 0/1 to the nucleotide counterparts (e.g., AA, CG, TC) for all of the NA12878 samples; 2) next we converted these genotype symbols to their IUPAC code; 3) we then collapsed the IUPAC symbols into a sequence per sample and placed them into a fasta file. We also included a reference "sample", which was just the reference allele at each de novo mutation and 4) we used MEGAX (*49*) version 10.2.4 to create a maximum likelihood phylogenetic tree.

Mutation profile assessment

We utilized the deconstructSig (*42*) software version-1.9.0 inside of Parabricks to perform mutation signature analysis. The prominent signature was chosen for an individual and if there was not one prominent signature than the weights of two signature was equal to or greater than (>= 0.31) both signatures were represented in the tables and figures.

Karyotype analysis

Read-depth based karyotypes were generated by assessment of the aligned sequence data. First, the number of reads per chromosome was calculated using SAMtools (*50*) in each individual. Second, the size of each chromosome was generated using the reference genome data and by removing locations of gaps from the reference. Third, the copy number of each of the chromosomes was calculated as follows: ((fold coverage per chromosome) / (fold coverage of chromosome 1))*2.

Viral analysis

We ran SAMtools idxstats on all individuals to determine the number of mapped reads to each chromosome. We then calculated the copy number of EBV in each individual as follows: EBV copy number = ((mapped reads to EBV * 150 base pairs per read) / length of EBV) / ((mapped reads to chromosome 1 * 150 base pairs per read) / length of chromosome 1)

DNV enrichment in genes

To test for DNV enrichment in genes we utilized two methods: chimpanzee-human and denovolyzeR. These were run as previously described (*10, 51*).

Annotation of protein-coding DNVs

We uploaded the DNV calls to the open-cravat program (https://opencravat.org/) and specifically identified Clinvar as one of the annotation categories. Rescoring of DNVs in Franklin was performed using Franklin (https://franklin.genoox.com).

ACKNOWLEDGMENTS

Funding: This work was supported by grants from the National Institutes of Health (R00MH117165 to T.N.T.). Author Contributions: J.N., T.H., C.P., and T.N.T. designed the study; J.N., P.V., E.F., E.M.P., Z.P., S.L., S.S., C.P., L.T., M.J., M.S., C.P. and T.N.T. performed analyses; J.N. and T.N.T. wrote the paper; and all authors reviewed and edited the paper. Competing interests: P.V., M.W., G.V. And T.T.H are full time employees of NVIDIA; and Data and materials availability: We provide family-level VCFs, DNV calls, and phased DNV results for the 602 trios in this study as a public community resource (Globus endpoint: "Turner Lab at WashU - DNV in 1000 Genomes Paper", direct link: https://app.globus.org/file-manager?origin_id=3eff453a-88f4-11eb-954f-752ba7b88ebe&origin_path=%2F). The raw alignment files for the high-coverage 1000 Genomes Project data are available at the following website: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/. The following cell lines/DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: NA06984, NA06985, NA06986, NA06989, NA06994, NA07000, NA07037, NA07048, NA07051, NA07056, NA07347, NA07357, NA10847, NA10851, NA11829, NA11830, NA11831, NA11832, NA11840, NA11843, NA11881, NA11892, NA11893, NA11894, NA11918, NA11919, NA11920, NA11930. NA11931, NA11932, NA11933, NA11992, NA11994, NA11995, NA12003, NA12004, NA12005, NA12006, NA12043, NA12044, NA12045, NA12046, NA12058, NA12144, NA12154, NA12155, NA12156, NA12234, NA12249, NA12272, NA12273, NA12275, NA12282, NA12283, NA12286, NA12287, NA12340, NA12341, NA12342, NA12347, NA12348, NA12383, NA12399, NA12400, NA12413, NA12414, NA12489, NA12546, NA12716, NA12717, NA12718, NA12748, NA12749, NA12750, NA12751, NA12760, NA12761, NA12762, NA12763, NA12775, NA12776, NA12777, NA12778, NA12812, NA12813, NA12814, NA12815, NA12827, NA12828, NA12829, NA12830, NA12842, NA12843, NA12872, NA12873, NA12874, NA12878, NA12889, NA12890, NA12376, NA10838, NA12329, NA10852, NA10840, NA12386, NA12864, NA12801, NA12344, NA10861, NA07029, NA12753, NA12832, NA12485, NA12802, NA12739, NA10856, NA10845, NA12818, NA10831, NA12766, NA10864, NA10843, NA12877, NA12335, NA12817, NA12752, NA12767, NA10855, NA12707, NA10857, NA10839, NA12740, NA10837, NA10836, NA07348, NA11993, NA12057, NA11839, NA06993, NA07014, NA06995, NA12146, NA12865, NA10859, NA06991, NA12336, NA10860, NA12145, NA07045, NA07349, NA07031, NA07345, NA12891, NA07055, NA07435, NA10835, NA12274, NA12875, NA10842, NA12239, NA10830, NA12056, NA11917, NA12892, NA06997, NA07022, NA12264, NA11891, NA07034, NA12248, NA10865, NA10863, NA10854, NA11882, NA07346, NA07019, NA12343, NA10846. These data were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1. We are grateful to all of the families at the participating SSC sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, 25D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, and E. Wijsman). We appreciate obtaining access to phenotypic and genetic data for the monozygotic twin pair on SFARI Base. Approved researchers can obtain the SSC population dataset described in this study (https://www.sfari.org/resource/simons-simplex-collection/) by applying at https://base.sfari.org.

1

FIGURES

Figure 1

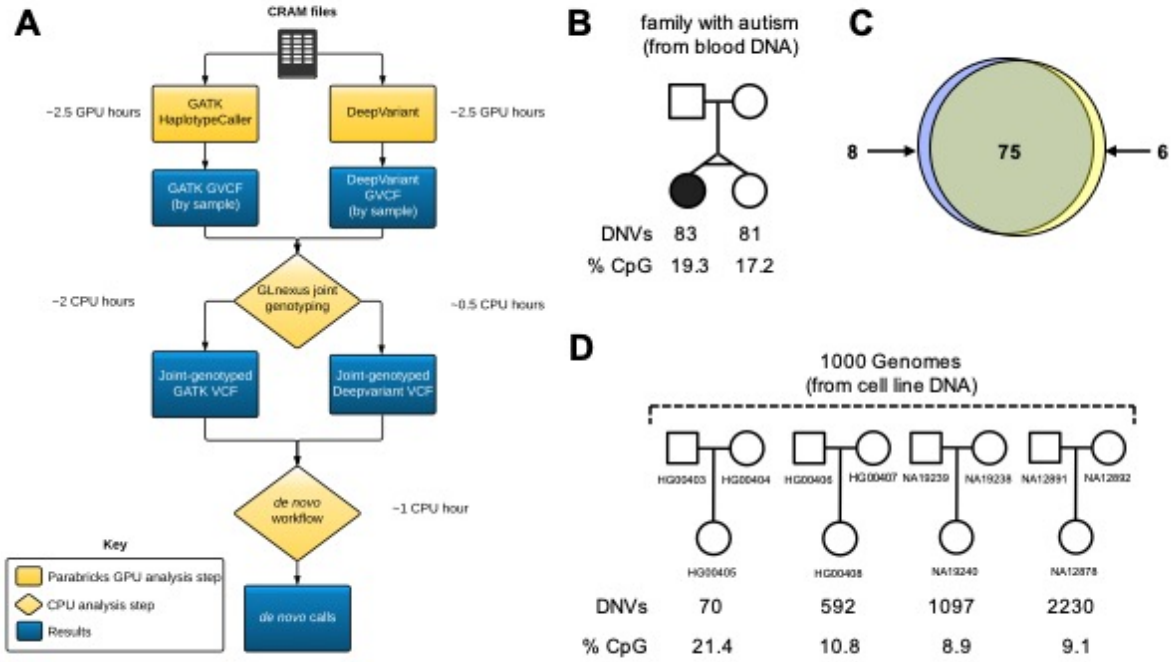


Figure 1: DNV calling in short-read whole-genome sequencing data. A) *de novo* workflow for detection of DNVs from aligned read files (crams); B and C) Benchmarking DNV workflow in a monozygotic twin pair sequenced from DNA derived from blood; D) DNV detecting in four trios in the 1000 Genomes Project.

Figure 2

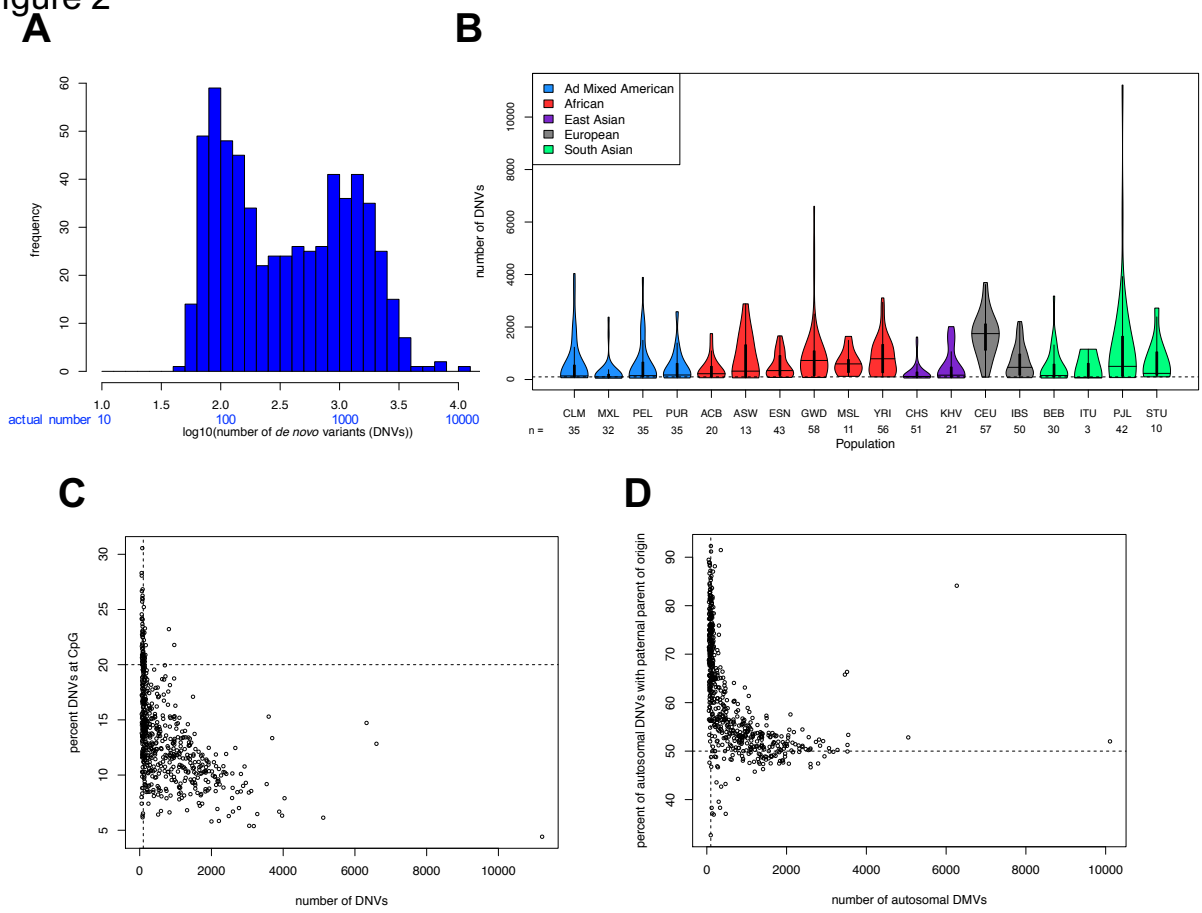


Figure 2: Characteristics of DNVs detected in 602 trios. A)Histogram of DNV counts in 602 trios; B) DNV counts by population; C) Percent of DNVs at CpG sites versus the total number of DNVs; D) Percent of DNVs phased to have a paternal parent-of-origin versus the total number of DNVs.

Figure 3

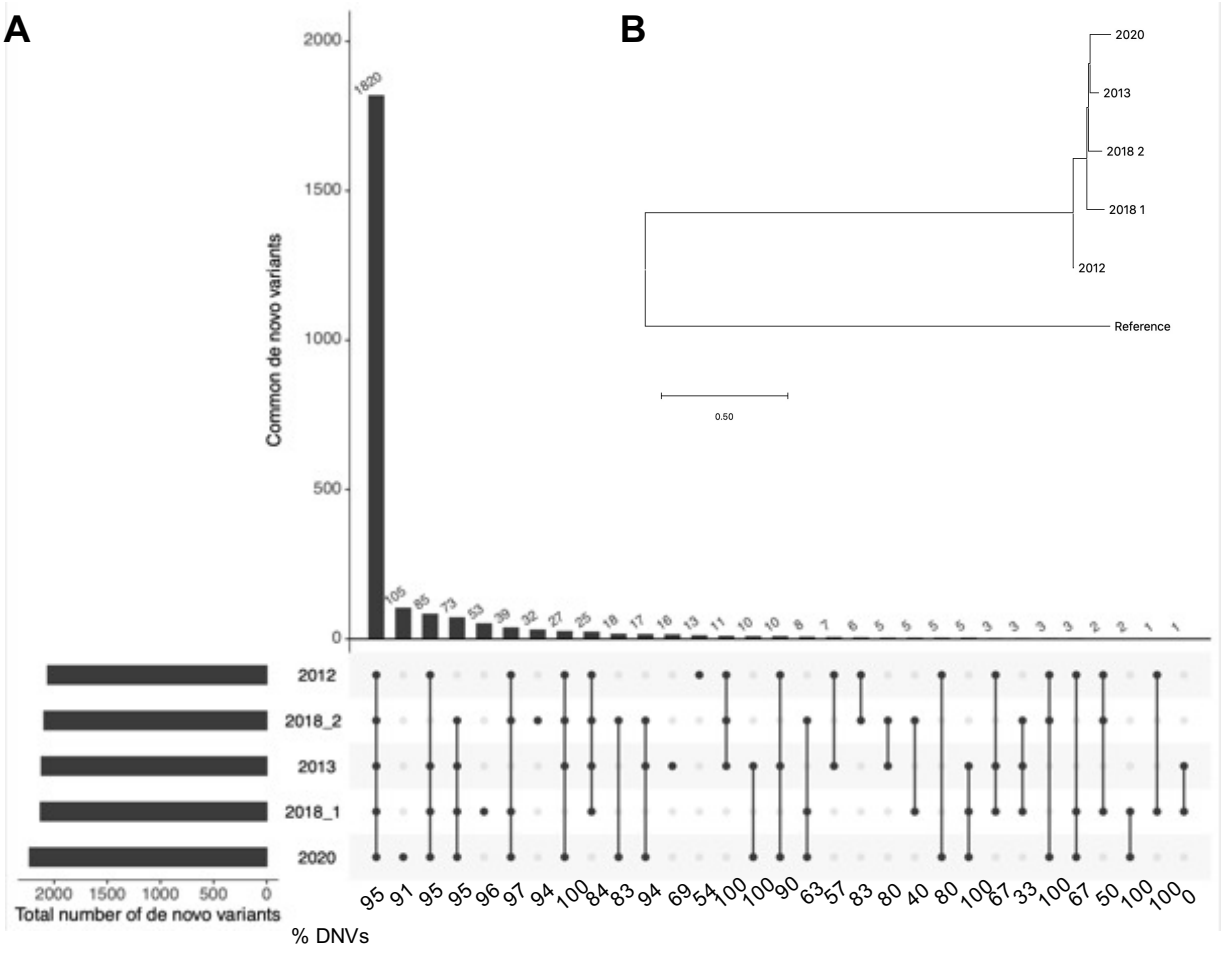


Figure 3: Assessment of five replicates of NA12878. A) UpSet plot demonstrating the number of variants detected in the replicates (at the bottom of the plot the percent of true DNVs is listed for each category); B) Phylogenetic tree of the five replicates

Figure 4

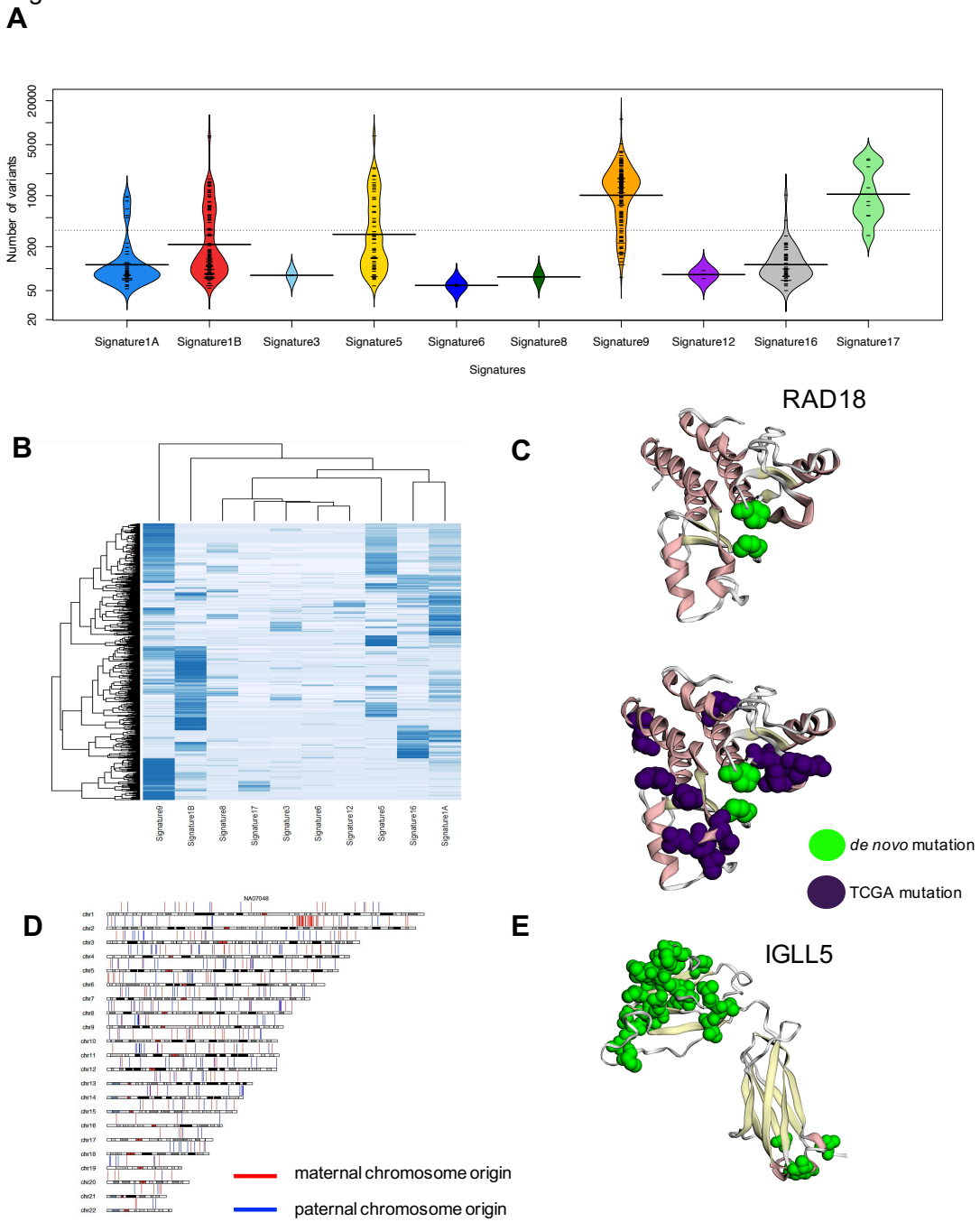


Figure 4: Mutational properties of DNVs. A) Mutation signature analysis showing the total number of DNVs and the individuals with each signature type; B) Heatmap of individuals based on their mutational signatures; C) Mutations in the DNA repair gene RAD18 shown on their 3D structure (and modeled using mupit). Also, shown are known cancer mutations from The Cancer Genome Atlas (TCGA); D) Location of DNVs based on their phased parent-of-origin in

NA07048. Most notable there are a cluster of mutations on the maternal chromosome on chromosome 2; E) DNVs in IGLL5 shown on their 3D structure (and modeled using mupit).

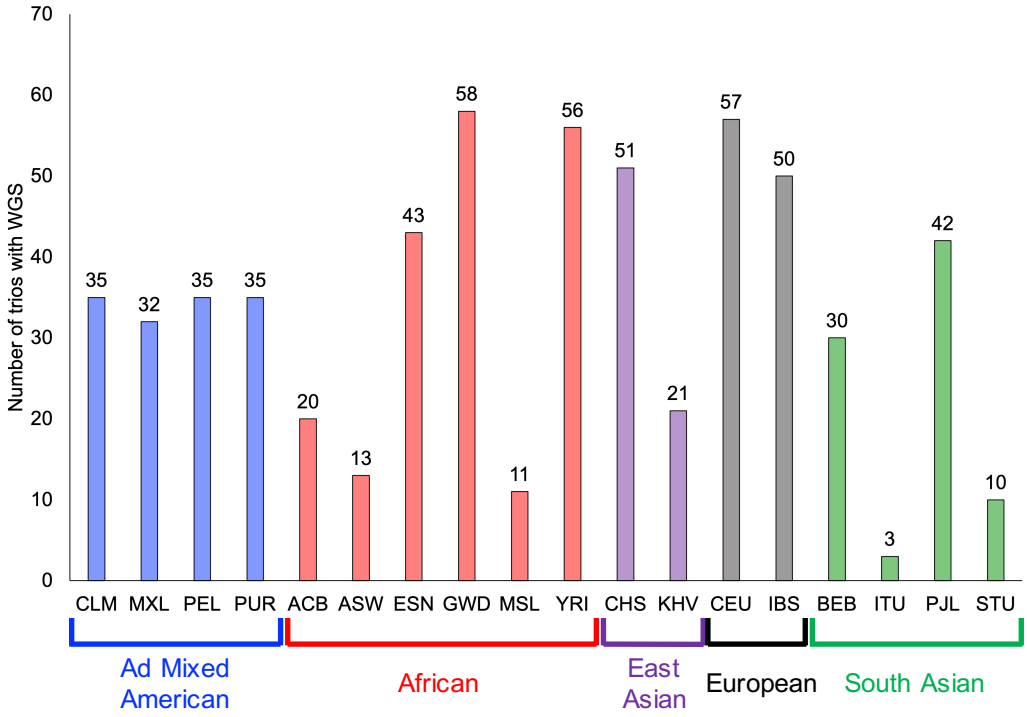
TABLES

Table 1. DNVs in DNA damage repair genes and clinically relevant variants

Category	individual	*de novo* variant	variant type	gene
DNA damage repair gene	HG01074	chr3_48447050_C_G	missense	*ATRIP*
	HG02841	chr16_89799603_A_G	splice_donor	*FANCA*
	HG03200	chr16_89762010_C_T	missense	*FANCA*
	HG01126	chr11_22625482_T_G	missense	*FANCF*
	NA18875	chr5_80654794_G_A	missense	*MSH3*
	HG02650	chr6_31759121_C_T	missense	*MSH5*
	NA10838	chr11_65865247_C_T	missense	*MUS81*
	NA10838	chr8_42357362_AT_A	frameshift	*POLB*
	NA19677	chr19_50407375_G_A	missense	*POLD1*
	HG01096	chr12_132634327_C_T	missense	*POLE*
	HG01755	chr15_89321792_C_T	missense	*POLG*
	NA12864	chr3_8958938_G_T	missense	*RAD18*
	HG02683	chr15_40729853_T_C	missense	*RAD51*
	HG02630	chrX_96884884_G_A	missense	*RPA4*
	NA19919	chr3_133644039_A_G	missense	*TOPBP1*
	HG04115	chr8_31120294_C_T	missense	*WRN*
	NA19161	chr8_31124967_G_T	missense	*WRN*
Clinvar pathogenic / likely pathogenic	HG03795	chr11_22274728_C_T	stop_gained	*ANO5*
	NA10854	chr13_110179298_C_T	missense	*COL4A1*
	NA10842	chr10_49530737_G_A	stop_gained	*ERCC6*
	HG02668	chr1_111787063_C_T	missense	*KCND3*
	HG02466	chr1_39485559_G_A	missense	*MACF1*
	HG02129	chr11_77206108_G_A	missense	*MYO7A*
	HG03122	chr11_45949458_G_A	stop_gained	*PHF21A*
	NA12707	chr6_52058438_C_T	missense	*PKHD1*
	HG01755	chr15_89321792_C_T	missense	*POLG*
	HG02892	chr11_124875581_C_T	stop_gained	*ROBO3*
	HG03635	chr2_165310406_G_A	missense	*SCN2A*
	NA10830	chr2_39023106_C_T	missense	*SOS1*
	NA10831	chr3_24143512_G_A	missense	*THRB*
	HG01629	chr2_209775898_C_T	stop_gained	*UNC80*
	HG00558	chr16_88435401_G_A	missense	*ZNF469*

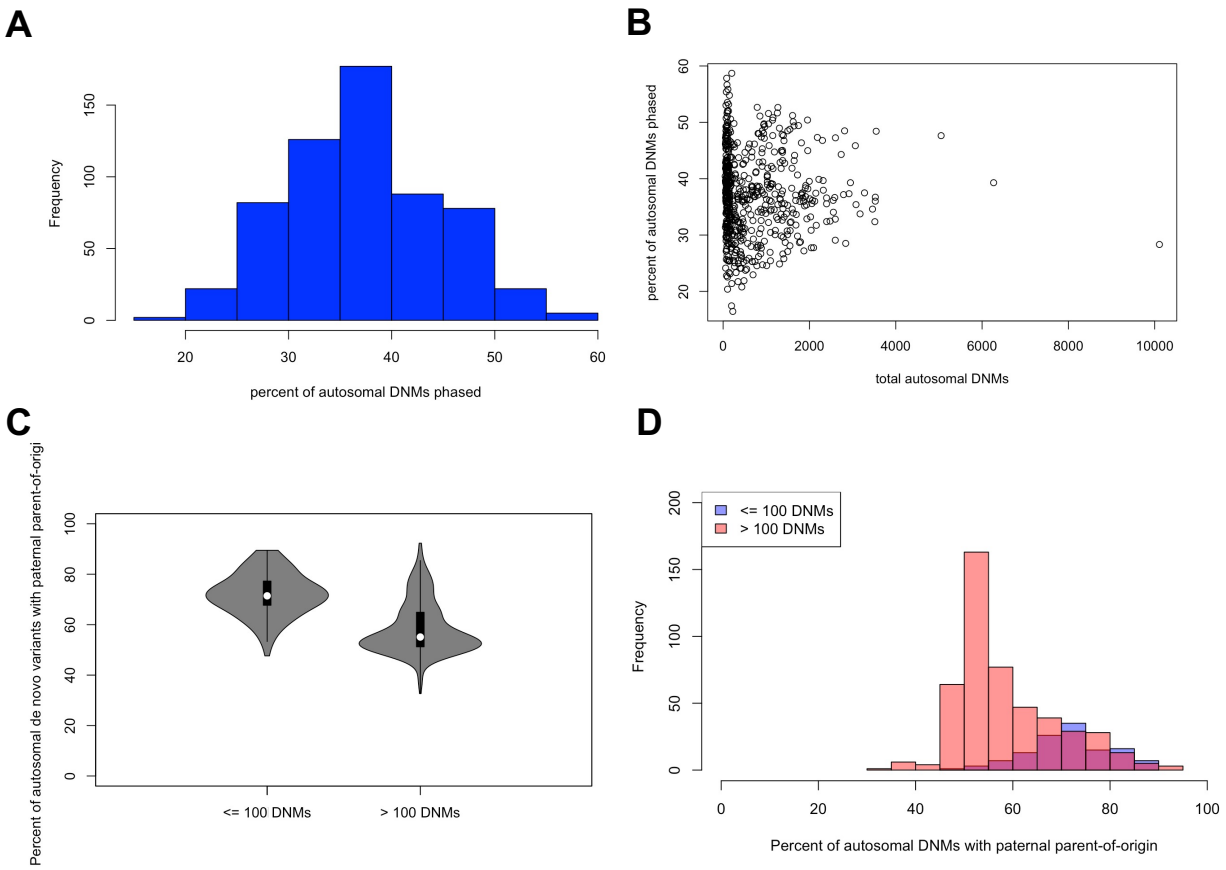
SUPPLEMENTARY FIGURES

Figure S1



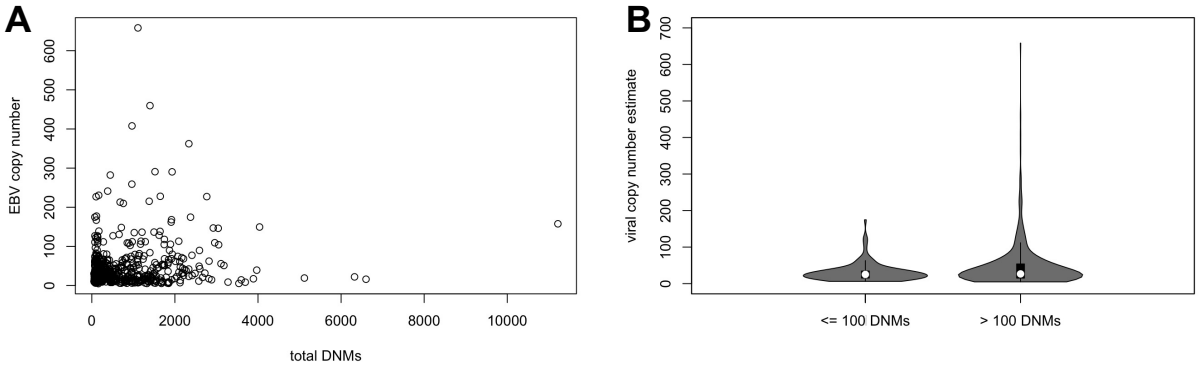
S1 Sample population distribution. A distribution of the populations by super and sub populations defined by the 1000 Genomes Project. Blue represents Ad Mixed American super population, red represents African super population, purple is East Asian super population, grey represents European super population, and green represents South Asian super population.

Figure S2



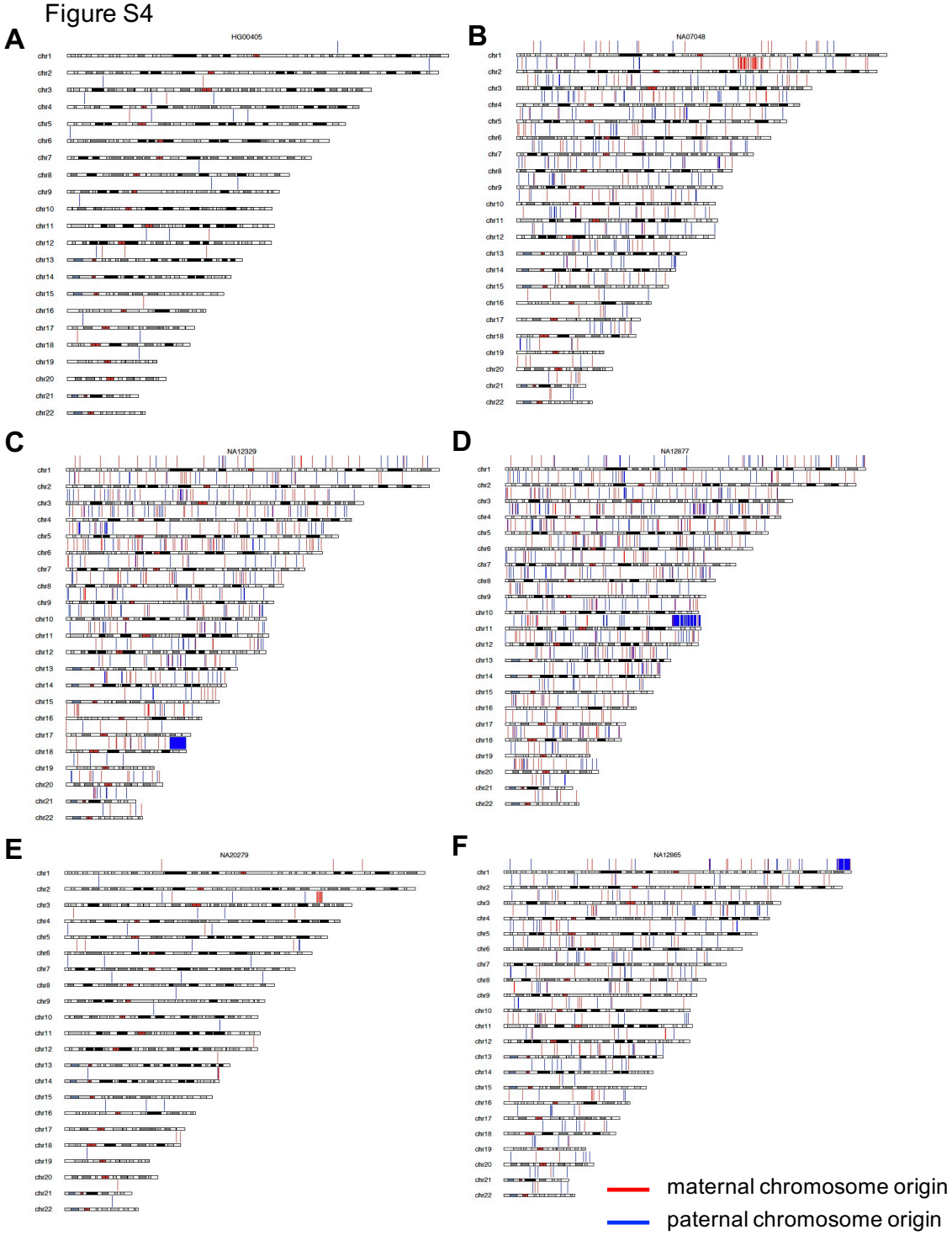
S2 Phase distribution of DNM. Histogram of the percent autosomal DNMs that was fixed A) distribution of percent autosomal DNMs phased vs. total autosomal DNMs.; B) Violin plot of percent of autosomal DNMs that with paternal parent-of-origin.; C) Distribution of DNMs with paternal parent-of-origin. The pink graph represents samples that had greater than 100 DNMs. The blue graph represents samples that had less than or equal to 100 DNMs. There is a trend of higher percent of DNMs with paternal parent-of-origin compared in the group that had less than or equal 100 DNMs compared to those with greater than 100 DNMs, which would be expected if the DNMs are real.

Figure S3



S3 EBV Copy Number Estimation. A) distribution of estimated EBV copy number vs total number of DNMs. There was a minor correlation between EBV copy number and total DNMs ($p = 2.32e\text{-}05$, $r = 0.17$).; B) violin plot comparing estimated EBV copy number and groups of less than or equal to 100 DNMs and greater than 100 DNMs.

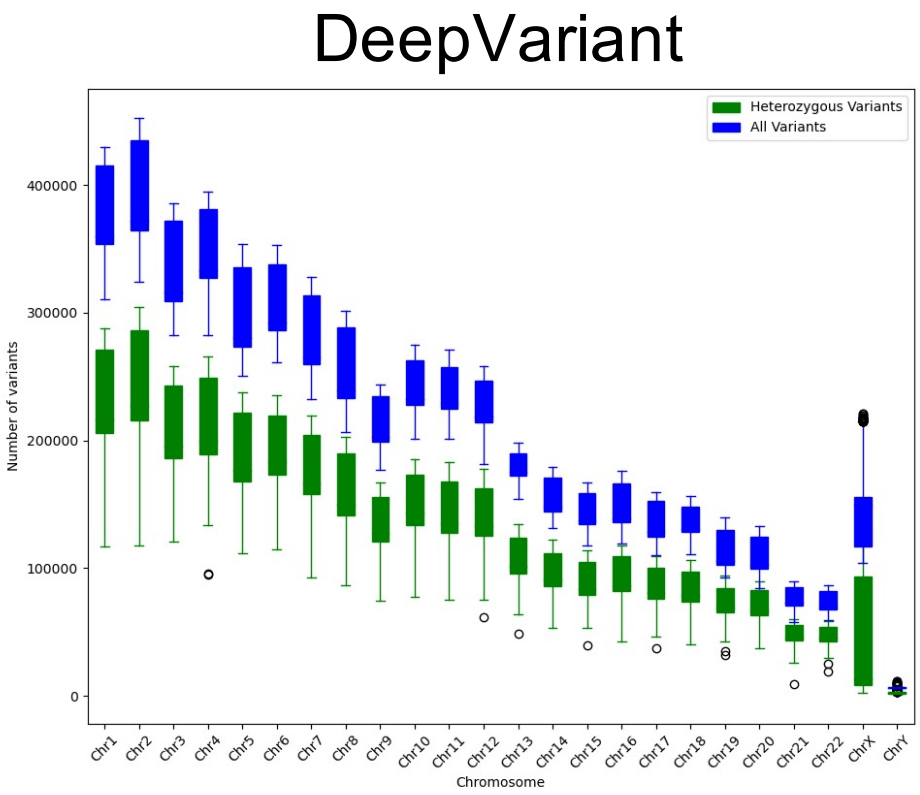
Figure S4



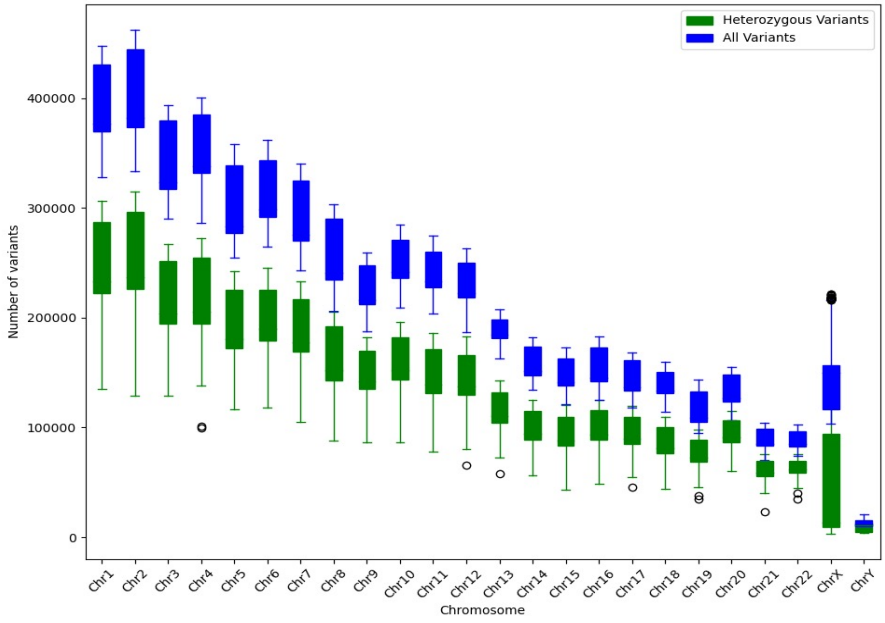
S4 Phase distributions across autosomal chromosomes. Distribution of phased DNMs across 6 different samples. The blue marks represent paternal parent-of-origin and the red mark represents the maternal parent-of-origin. A) represents a normal distribution, B) represents a clustering of maternal mutations, C), D) and F) represent paternal clustering of mutations, and E) represents a clustering of maternal mutations possible event that could lead to further DNVs.

Figure S5

A



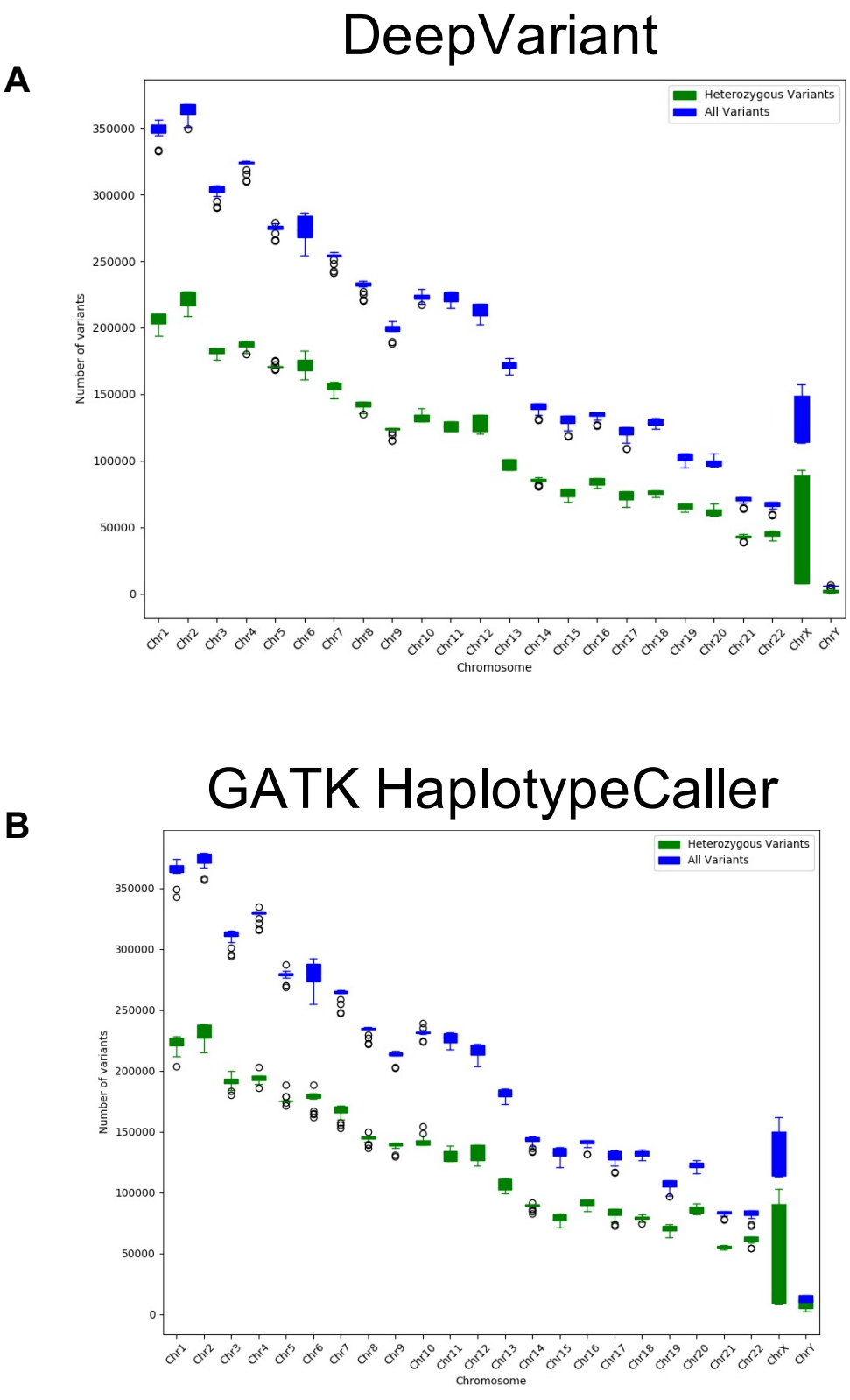
B



1

S5 Total counts of variants for DeepVariant and Haplotyepcaller trio calls. Distribution of variants by chromosome found in the trios from the 602 1000 Genomes Project families. The total number of variants in blue, and heterozygous variants, in green.

Figure S6



1

S6 Total counts of variants for DeepVariant and Haplotyepcaller downsampled 30x coverage NA12878 trio calls. Distribution of variants by chromosome found in the trios from the various NA12878 samples, downsampled to 30x coverage, as well as samples NA12891 and NA12892. The total number of variants in blue, and heterozygous variants, in green.

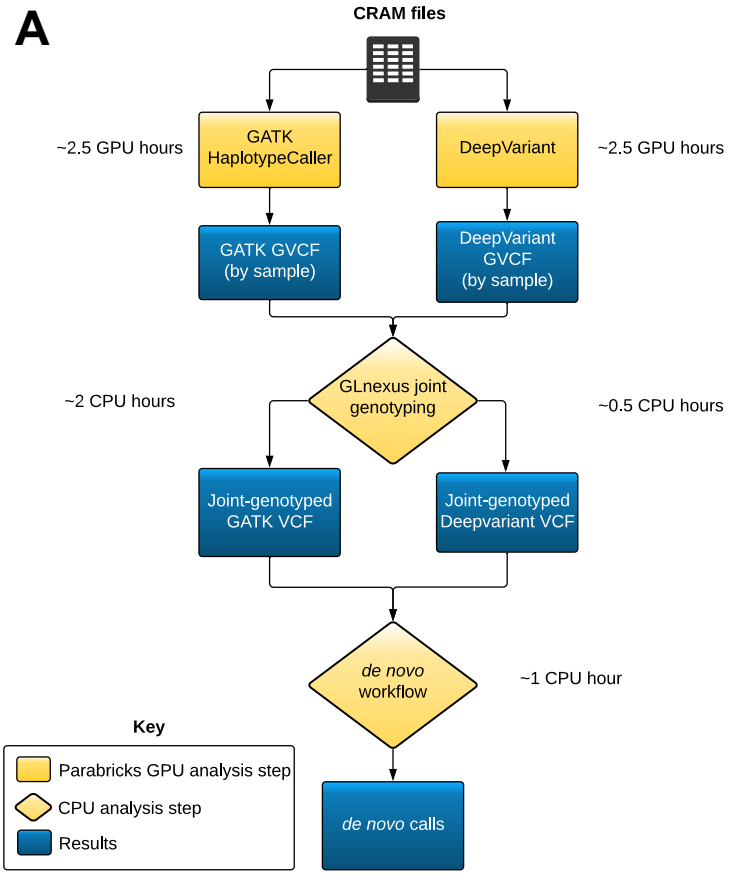
REFERENCES

1. A. Auton *et al.*, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
2. M. Byrska-Bishop *et al.*, High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv*, 2021.2002.2006.430068 (2021).
3. L. Ségurel, M. J. Wyman, M. Przeworski, Determinants of mutation rate variation in the human germline. *Annual review of genomics and human genetics* **15**, 47-70 (2014).
4. P. Feliciano *et al.*, Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *NPJ genomic medicine* **4**, 19 (2019).
5. I. Iossifov *et al.*, The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, (2014).
6. B. J. O'Roak *et al.*, Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature genetics* **43**, 585-589 (2011).
7. B. J. O'Roak *et al.*, Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250 (2012).
8. T. N. Turner *et al.*, Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**, 710-722.e712 (2017).
9. T. N. Turner *et al.*, Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *American journal of human genetics* **98**, 58-74 (2016).
10. T. N. Turner *et al.*, Sex-Based Analysis of De Novo Variants in Neurodevelopmental Disorders. *American journal of human genetics* **105**, 1274-1285 (2019).
11. S. J. Sanders *et al.*, De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012).
12. S. De Rubeis *et al.*, Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).
13. K. L. Helbig *et al.*, Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy. *Genetics in medicine : official journal of the American College of Medical Genetics*, (2016).
14. A. S. Allen *et al.*, De novo mutations in epileptic encephalopathies. *Nature* **501**, 217-221 (2013).
15. J. Kaplanis *et al.*, Integrating healthcare and research genetic data empowers the discovery of 28 novel developmental disorders. *bioRxiv*, 797787 (2020).
16. DDD, Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438 (2017).
17. J. de Ligt *et al.*, Diagnostic exome sequencing in persons with severe intellectual disability. *The New England journal of medicine* **367**, 1921-1929 (2012).
18. C. Gilissen *et al.*, Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344-347 (2014).
19. J. Homsy *et al.*, De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science (New York, N.Y.)* **350**, 1262-1266 (2015).
20. S. Zaidi *et al.*, De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-223 (2013).
21. A. Sifrim *et al.*, Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nature genetics* **48**, 1060-1065 (2016).
22. H. L. Rehm *et al.*, ACMG clinical laboratory standards for next-generation sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics* **15**, 733-747 (2013).
23. S. Besenbacher *et al.*, Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nature communications* **6**, 5969 (2015).
24. D. I. Boomsma *et al.*, The Genome of the Netherlands: design, and project goals. *European journal of human genetics : EJHG* **22**, 221-227 (2014).
25. A. Chesi *et al.*, Exome sequencing to identify de novo mutations in sporadic ALS trios. *Nature neuroscience* **16**, 851-855 (2013).

26. D. F. Conrad *et al.*, Variation in genome-wide mutation rates within and between human families. *Nature genetics* **43**, 712-714 (2011).
27. S. Dimassi *et al.*, Whole-exome sequencing improves the diagnosis yield in sporadic infantile spasm syndrome. *Clinical genetics* **89**, 198-204 (2016).
28. S. Dong *et al.*, De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell reports* **9**, 16-23 (2014).
29. M. Fromer *et al.*, De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184 (2014).
30. I. Iossifov *et al.*, De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-299 (2012).
31. H. Jonsson *et al.*, Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519-522 (2017).
32. S. E. McCarthy *et al.*, De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular psychiatry* **19**, 652-658 (2014).
33. G. Narzisi *et al.*, Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nature methods* **11**, 1033-1036 (2014).
34. A. Ramu *et al.*, DeNovoGear: de novo indel and point mutation discovery and phasing. *Nature methods* **10**, 985-987 (2013).
35. K. R. Franke, E. L. Crowgey, Accelerating next generation sequencing data analysis: an evaluation of optimized best practices for Genome Analysis Toolkit algorithms. *Genomics Inform* **18**, e10 (2020).
36. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).
37. R. Poplin *et al.*, A universal SNP and small-indel variant caller using deep neural networks. *Nature biotechnology* **36**, 983-987 (2018).
38. T. Yun *et al.*, Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *bioRxiv*, 2020.2002.2010.942086 (2020).
39. J. M. Zook *et al.*, Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. (2014).
40. R. A. Gibbs *et al.*, The International HapMap Project. *Nature* **426**, 789-796 (2003).
41. T. A. Sasani *et al.*, Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife* **8**, (2019).
42. R. Rosenthal, N. McGranahan, J. Herrero, B. S. Taylor, C. Swanton, DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome biology* **17**, 31 (2016).
43. M. J. Landrum *et al.*, ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research* **46**, D1062-d1067 (2018).
44. G. R. Abecasis *et al.*, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
45. F. Chen, Y. Zhang, C. J. Creighton, Systematic identification of non-coding somatic single nucleotide variants associated with altered transcription and DNA methylation in adult and pediatric cancers. *NAR Cancer* **3**, zcab001 (2021).
46. L. Pedrosa *et al.*, Proposal and validation of a method to classify genetic subtypes of diffuse large B cell lymphoma. *Scientific reports* **11**, 1886 (2021).
47. J. R. Belyeu, T. A. Sasani, B. S. Pedersen, A. R. Quinlan, Unfazed: parent-of-origin detection for large and small de novo variants. *bioRxiv*, 2021.2002.2003.429658 (2021).
48. C. Chiang *et al.*, SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature methods* **12**, 966-968 (2015).
49. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular biology and evolution* **35**, 1547-1549 (2018).

50. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078-2079 (2009).
51. B. P. Coe *et al.*, Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nature genetics* **51**, 106-116 (2019).

Figure 1


A
CRAM files
~2.5 GPU hours
GATK HaplotypeCaller
DeepVariant
~2.5 GPU hours
GATK GVCF (by sample)
DeepVariant GVCF (by sample)
~2 CPU hours
GLnexus joint genotyping
~0.5 CPU hours
Joint-genotyped GATK VCF
Joint-genotyped Deepvariant VCF
de novo workflow
~1 CPU hour
Key
Parabricks GPU analysis step
CPU analysis step
Results
de novo calls

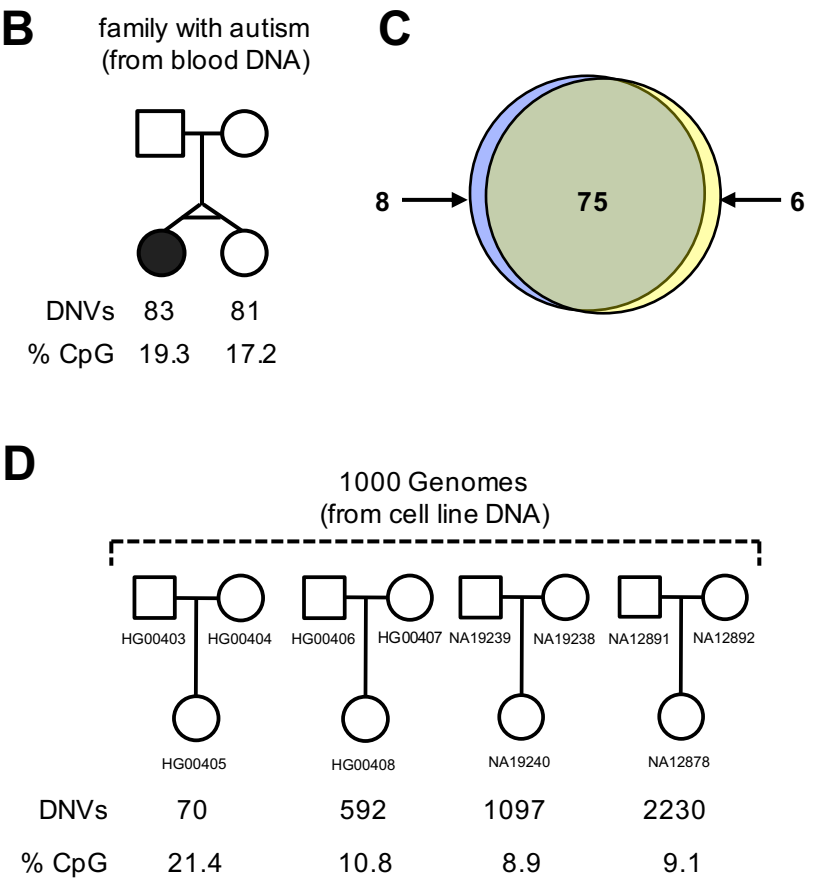

B
family with autism (from blood DNA)
DNVs 83 81
% CpG 19.3 17.2
C
8
75
6
D
1000 Genomes (from cell line DNA)
HG00403
HG00404
HG00406
HG00407
NA19239
NA19238
NA12891
NA12892
HG00405
HG00408
NA19240
NA12878
DNVs 70 592 1097 2230
% CpG 21.4 10.8 8.9 9.1

Figure 2

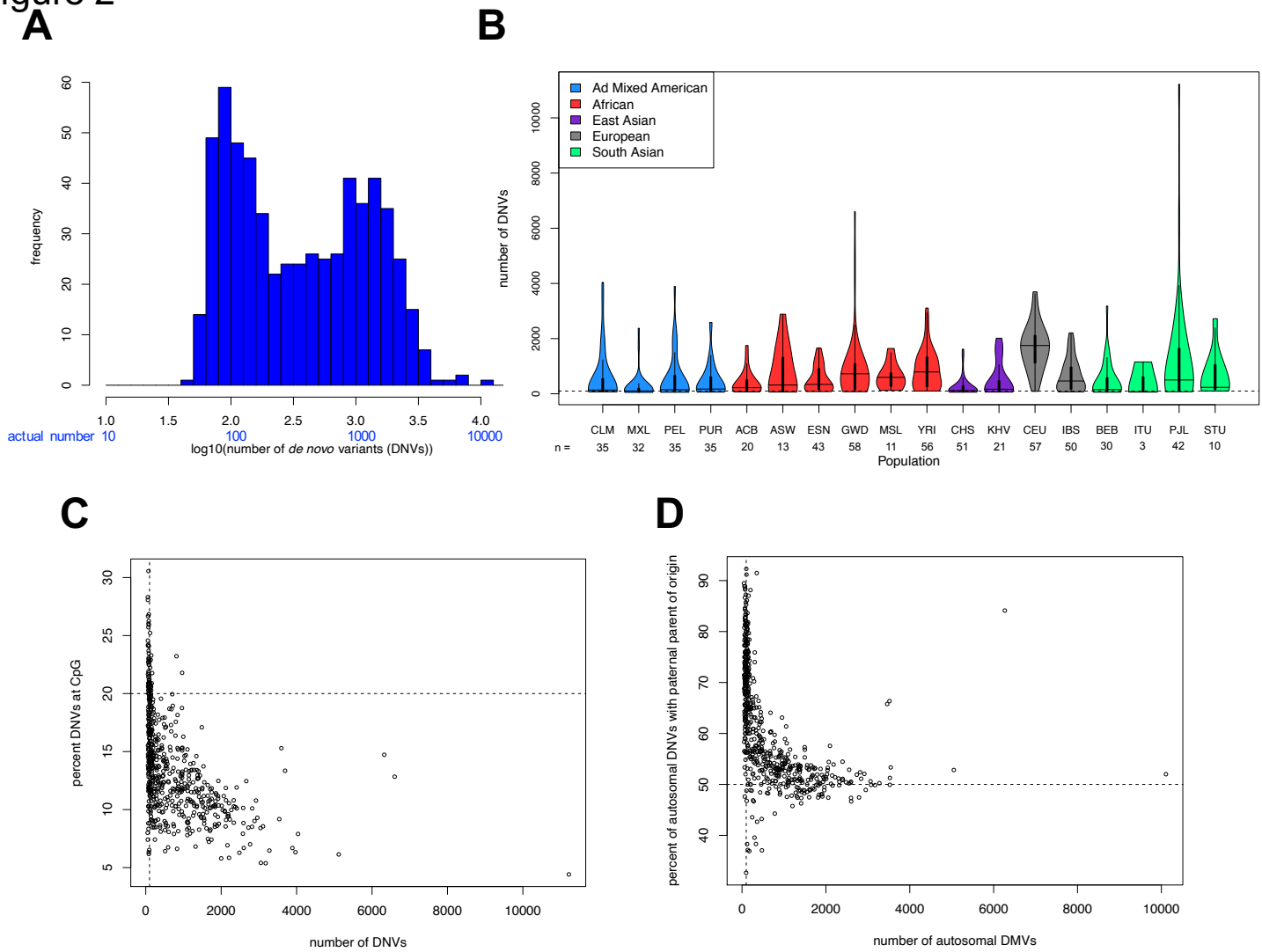

A
frequency
log10(number of de novo variants (DNVs))
actual number 10
100
1000
10000
B
Ad Mixed American
African
East Asian
European
South Asian
number of DNVs
CLM MXL PEL PUR ACB ASW ESN GWD MSL YRI CHS KHV CEU IBS BEB ITU PJL STU
n = 35 32 35 35 20 13 43 58 11 56 51 21 57 50 30 3 42 10
Population
C
percent DNVs at CpG
number of DNVs
D
percent of autosomal DNVs with paternal parent of origin
number of autosomal DMVs

Figure 3

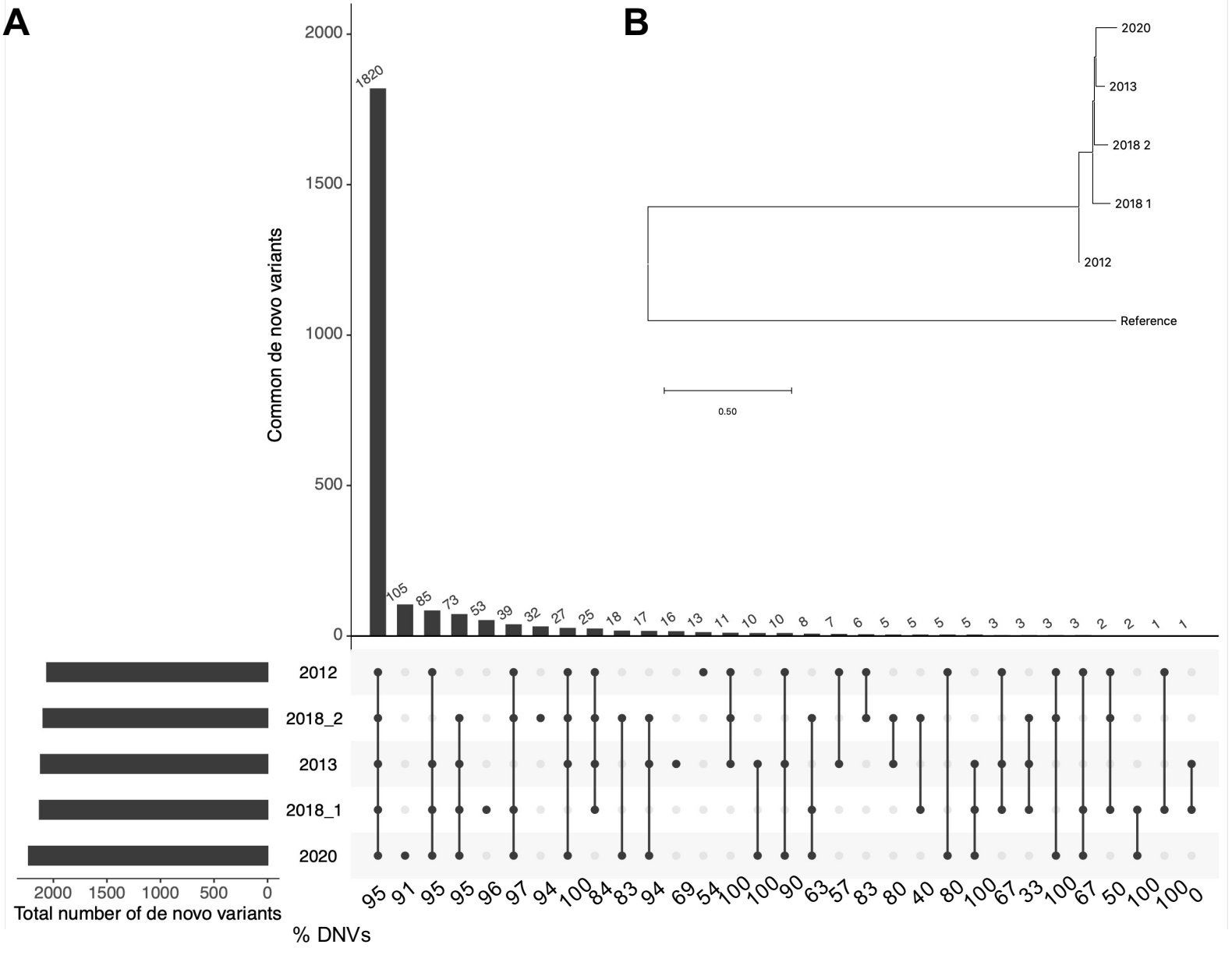
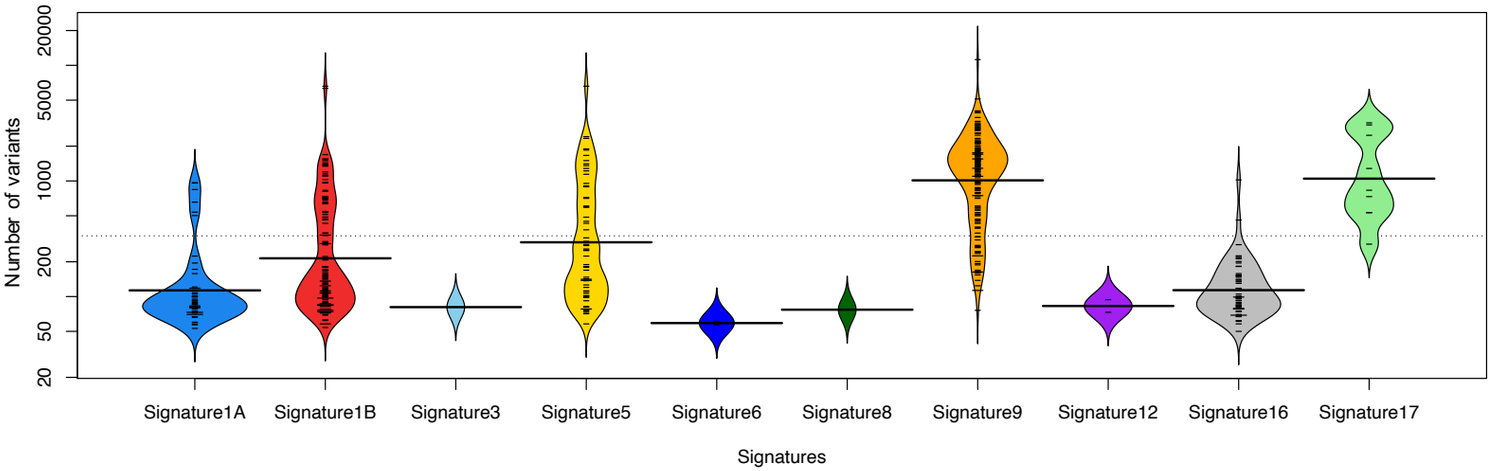
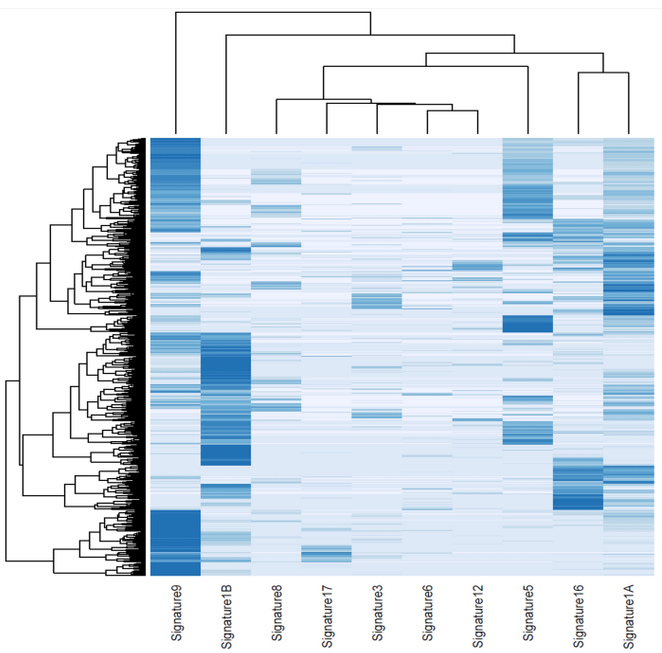

A
B
Common de novo variants
2000
1500
1000
500
0
1820
105
85
73
53
39
32
27
25
18
17
16
13
11
10
10
8
7
6
5
5
5
5
3
3
3
3
2
2
1
1
2020
2013
2018 2
2018 1
2012
Reference
0.50
2012
2018_2
2013
2018_1
2020
2000 1500 1000 500 0
Total number of de novo variants
95 91 95 95 96 97 94 100 84 83 94 69 54 100 100 90 63 57 83 80 40 80 100 67 33 100 67 50 100 100 0
% DNVs

Figure 4

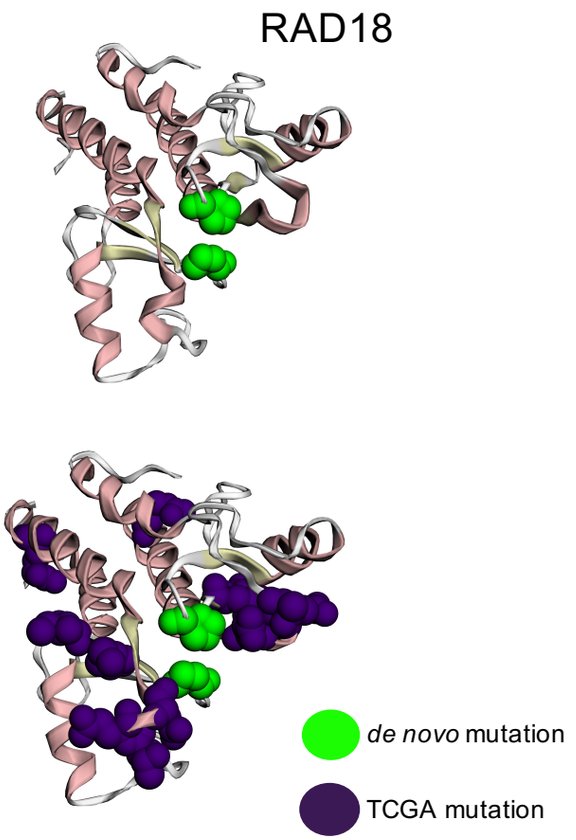
A



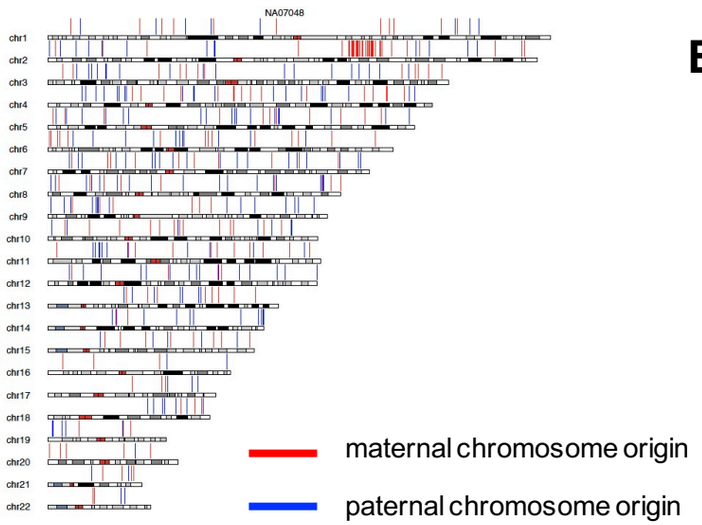
B



C



D



E

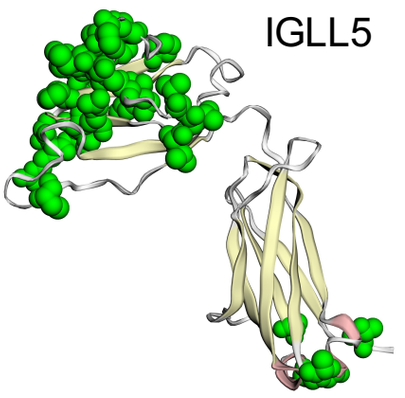


Figure S1

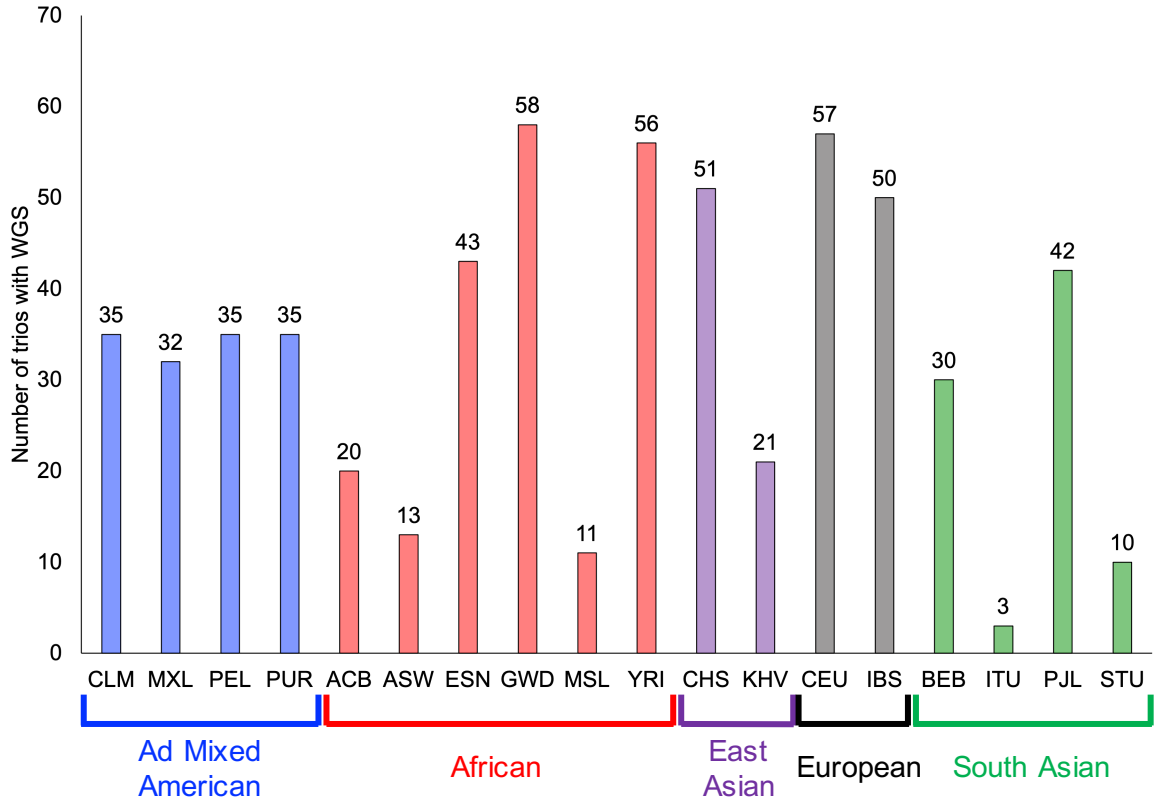

Number of trios with WGS
70
60
50
40
30
20
10
0
35
32
35
35
20
13
43
58
11
56
51
21
57
50
30
3
42
10
CLM MXL PEL PUR ACB ASW ESN GWD MSL YRI CHS KHV CEU IBS BEB ITU PJL STU
Ad Mixed American
African
East Asian
European
South Asian

Figure S2

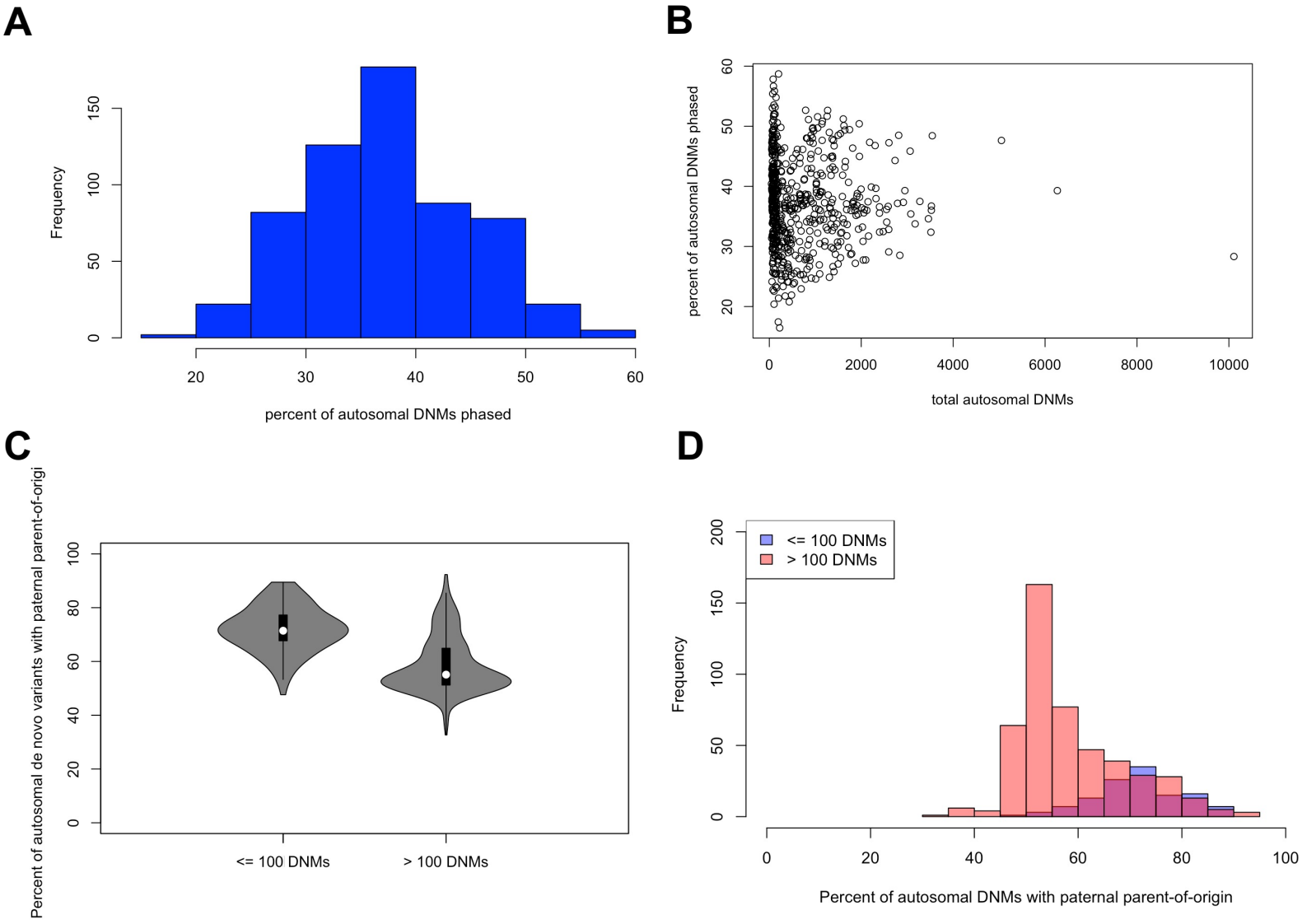

A
Frequency
0
50
100
150
20
30
40
50
60
percent of autosomal DNMs phased
B
percent of autosomal DNMs phased
20
30
40
50
60
0
2000
4000
6000
8000
10000
total autosomal DNMs
C
Percent of autosomal de novo variants with paternal parent-of-origi
0
20
40
60
80
100
<= 100 DNMs
> 100 DNMs
D
<= 100 DNMs
> 100 DNMs
Frequency
0
50
100
150
200
0
20
40
60
80
100
Percent of autosomal DNMs with paternal parent-of-origin

Figure S3

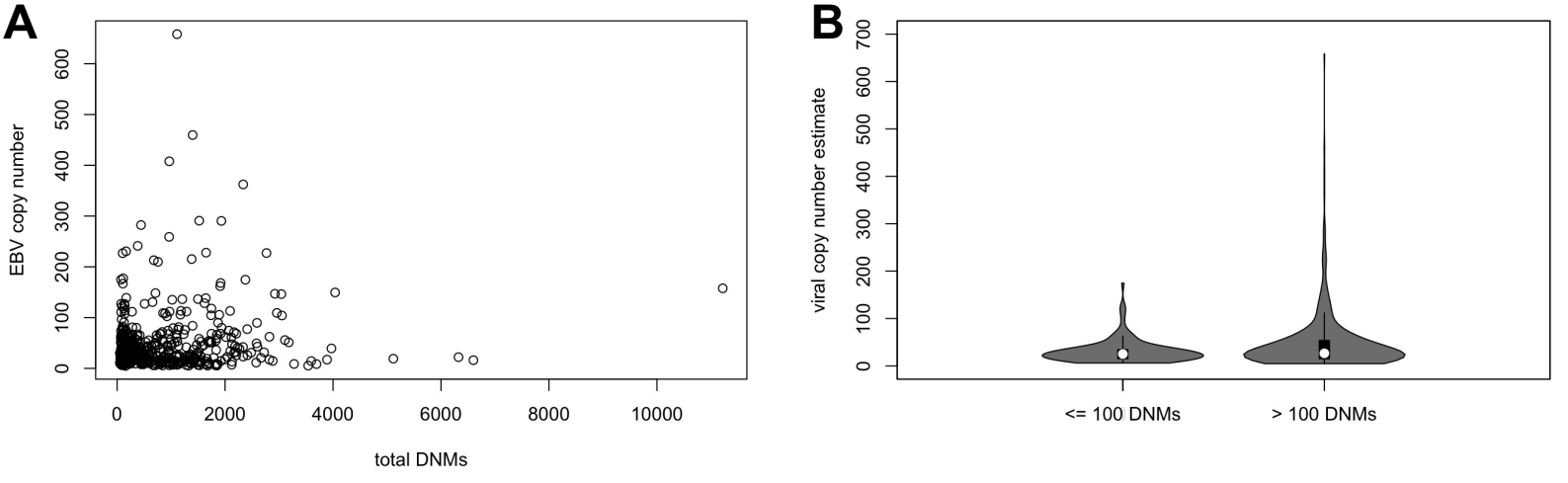

A
EBV copy number
0
100
200
300
400
500
600
0
2000
4000
6000
8000
10000
total DNMs
B
viral copy number estimate
0
100
200
300
400
500
600
700
<= 100 DNMs
> 100 DNMs

Figure S4

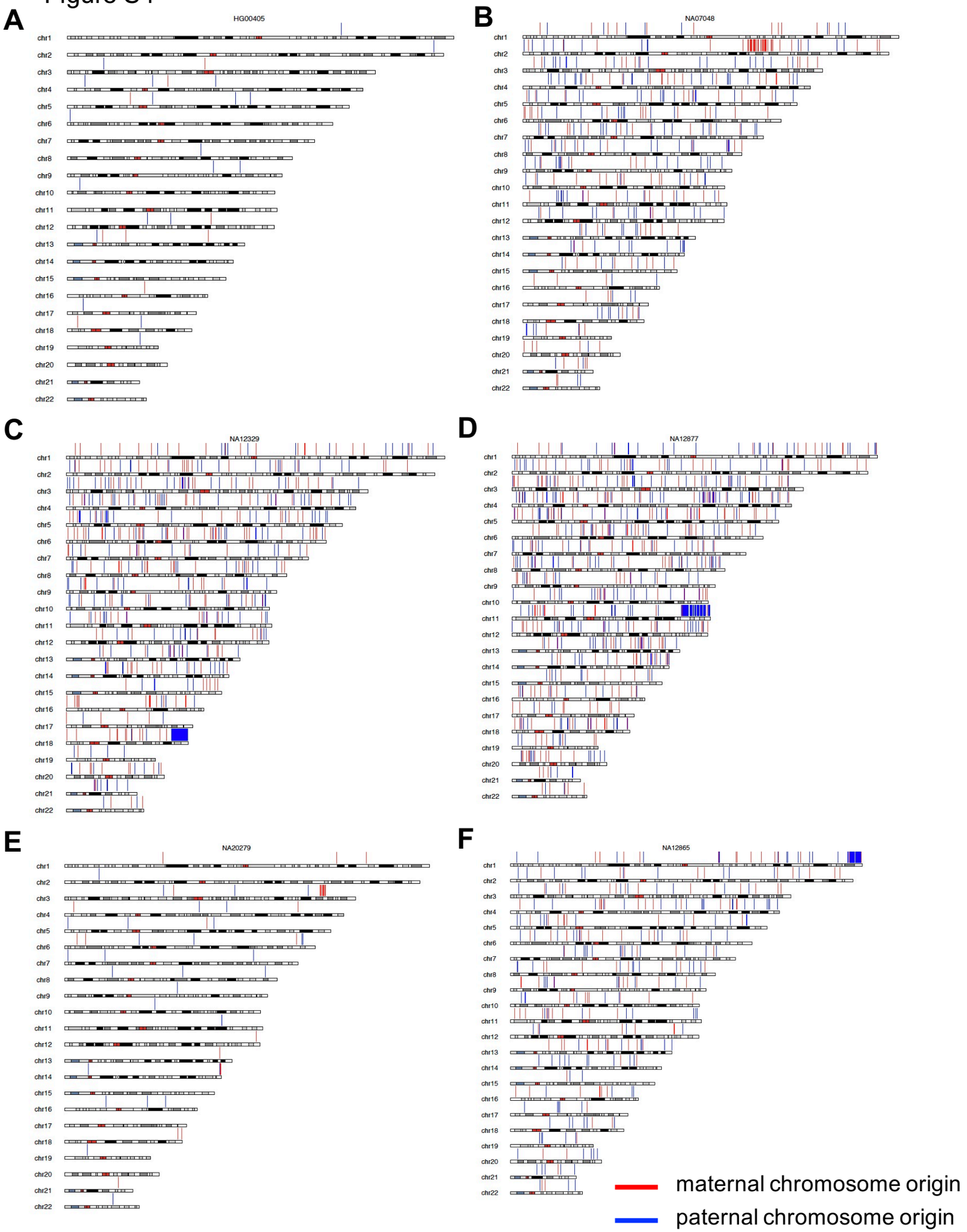

A
HG00405
B
NA07048
C
NA12329
D
NA12877
E
NA20279
F
NA12865
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
maternal chromosome origin
paternal chromosome origin

Figure S5

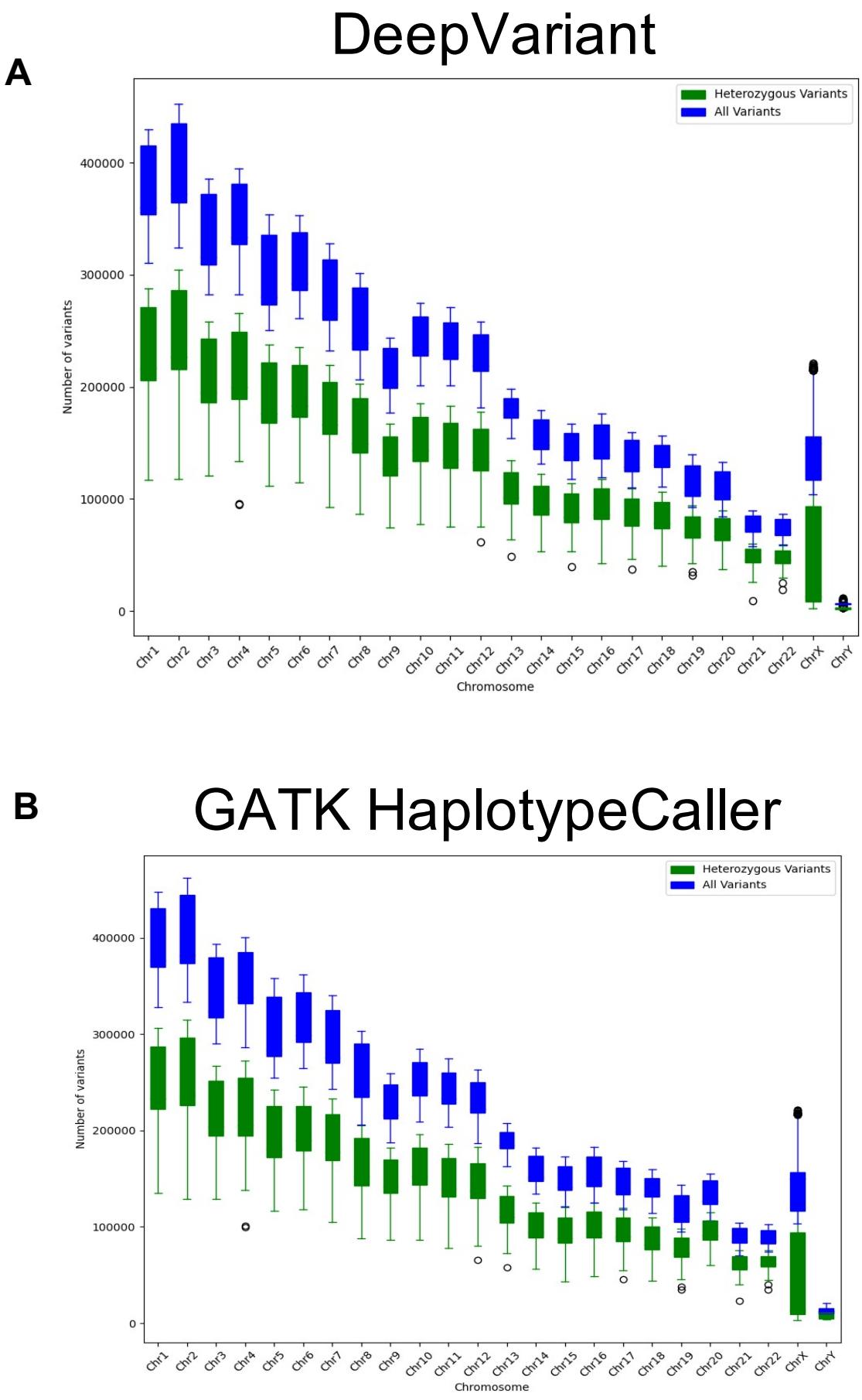
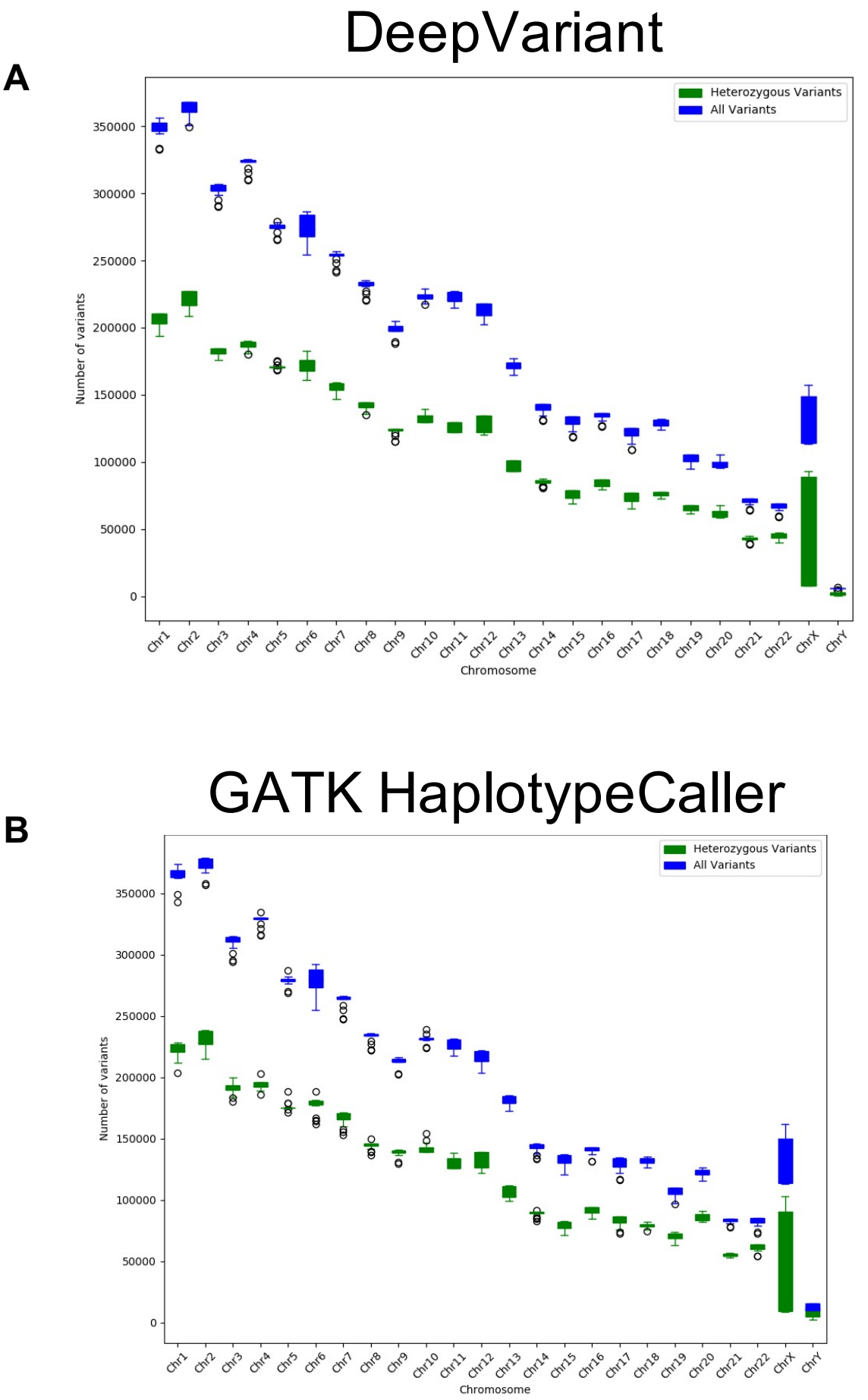

A
DeepVariant
Heterozygous Variants
All Variants
Number of variants
0
100000
200000
300000
400000
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
ChrY
Chromosome
B
GATK HaplotypeCaller
Heterozygous Variants
All Variants
Number of variants
0
100000
200000
300000
400000
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
ChrY
Chromosome

Figure S6


A
DeepVariant
Heterozygous Variants
All Variants
Number of variants
0
50000
100000
150000
200000
250000
300000
350000
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
ChrY
Chromosome
B
GATK HaplotypeCaller
Heterozygous Variants
All Variants
Number of variants
0
50000
100000
150000
200000
250000
300000
350000
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
ChrY
Chromosome