7236–7249 *Nucleic Acids Research, 2018, Vol. 46, No. 14* *Published online 21 June 2018*
doi: 10.1093/nar/gky538

Human copy number variants are enriched in regions of low mappability

Jean Monlong[1,2], Patrick Cossette[3], Caroline Meloche[3], Guy Rouleau[4], Simon L. Girard[1,3,5] and Guillaume Bourque[1,2,6,*]

[1]Department of Human Genetics, McGill University, Montréal H3A 1B1, Canada, [2]Canadian Center for Computational Genomics, Montréal H3A 1A4, Canada, [3]Centre de Recherche du Centre Hospitalier de l'Universite de Montréal, Montréal H2X 0A9, Canada, [4]Montreal Neurological Institute, McGill University, Montréal H3A 2B4, Canada, [5]Département des sciences fondamentales, Université du Québec à Chicoutimi, Chicoutimi G7H 2B1, Canada and [6]McGill University and Génome Québec Innovation Center, Montréal H3A 1A4, Canada

Received February 20, 2018; Revised May 04, 2018; Editorial Decision May 29, 2018; Accepted June 12, 2018

ABSTRACT

Copy number variants (CNVs) are known to affect a large portion of the human genome and have been implicated in many diseases. Although whole-genome sequencing (WGS) can help identify CNVs, most analytical methods suffer from limited sensitivity and specificity, especially in regions of low mappability. To address this, we use `PopSV`, a CNV caller that relies on multiple samples to control for technical variation. We demonstrate that our calls are stable across different types of repeat-rich regions and validate the accuracy of our predictions using orthogonal approaches. Applying `PopSV` to 640 human genomes, we find that low-mappability regions are approximately 5 times more likely to harbor germline CNVs, in stark contrast to the nearly uniform distribution observed for somatic CNVs in 95 cancer genomes. In addition to known enrichments in segmental duplication and near centromeres and telomeres, we also report that CNVs are enriched in specific types of satellite and in some of the most recent families of transposable elements. Finally, using this comprehensive approach, we identify 3455 regions with recurrent CNVs that were missing from existing catalogs. In particular, we identify 347 genes with a novel exonic CNV in low-mappability regions, including 29 genes previously associated with disease.

INTRODUCTION

Genomic variation of 50 bp or more are collectively known as structural variants (SVs) and can take several forms including deletions, duplications, novel insertions, translocations and inversions (1). Copy number variants (CNVs) are unbalanced SVs, i.e. affecting DNA copy number, and include deletions and any type of duplications (tandem duplications, triplications and other amplifications). A wide range of mechanisms can produce SVs and is responsible for the diverse SV distribution across the genome, both in term of location and size (1–3). In healthy individuals, SVs are estimated to cumulatively affect a higher proportion of the genome as compared to single nucleotide polymorphisms (SNPs) (4). SVs have been associated with numerous diseases including Crohn's Disease (5), schizophrenia (6), obesity (7), epilepsy (8), autism (9), cancer (10) and other inherited diseases (11,12), and many SVs have a demonstrated detrimental effect.

While large SVs have been first studied using cytogenetic approaches and array-based technologies, whole-genome sequencing (WGS) is in theory capable of detecting SVs of any type and size (13). Numerous methods have been implemented to detect SVs from WGS data using either paired-end information (14,15), read-depth (RD) variation (16–18), breakpoints detection through split-read approach (19) or de novo assembly (20). CNVs, potentially the most impactful SVs, can be detected by any of these strategies but are often resolved with a RD approach as it directly looks for signs of copy number changes. However, several features of WGS experiments result in technical bias and continue to be a major challenge. For example, GC content (21), mappability (22,23), replication timing (24), DNA quality and library preparation (25) have a detrimental impact on the uniformity of the RD (26). Unfortunately, this variability is difficult to fully correct for as it involves different factors, some of which are unknown, that vary from one experiment to another. This issue particularly impairs the detection of CNV with weaker signal, which is inevitable in regions of low-mappability that represent around 10% of the human genome (27), for smaller CNVs or in cancer samples with cell heterogeneity or stromal contamination. As a result, existing approaches suffer from limited sensitivity and specificity (3,13), especially in regions of low-

[*]To whom correspondence should be addressed. Tel: +1 514 398 7245; Fax: +1 514 398 1790; Email: guil.bourque@mcgill.ca

complexity and low-mappability (22,23). Even when problematic regions were masked and state-of-the-art bias correction (21,28) were applied, we showed that technical variation in RD could still be found across three WGS datasets studied (29).

To control for technical variation, we recently developed a CNV detection method, PopSV, which uses a set of reference samples to detect abnormal RD (29). In each genome tested, the RD in a region is compared to the same region in the reference samples. PopSV differs from most previous RD methods, such as RDXplorer (30) or CNVnator (17), that scan the genome horizontally and look for regions that diverge from the expected global average. Even when approaches rely on a ratio between an aberrant sample and a control, such as FREEC (16) or BIC-seq (31), we showed that they do not sufficiently control for experiment-specific noise as compared to PopSV (29). Glusman *et al.* (32) does go further and normalize the RD with pre-computed RD profiles that fit the GC-fingerprint of a sample but this approach excludes regions with extreme RD and does not integrate the variance observed in individual regions. PopSV is also different from approaches such as cn.MOPS (18) and Genome STRiP (33) that scan simultaneously the genome of several samples and fit a Bayesian or Gaussian mixture model in each region. Those methods have more power to detect CNVs present in several samples but may miss sample-specific events. Moreover, their basic normalization of the RD and fully parametric models forces them to conceal a sizable portion of the genome and variants with weaker signal. Finally, another strategy to improve the accuracy of CNV detection has been to use an ensemble approach that combines information from different methods relying on different types of reads. Large resequencing projects such as the 1000 Genome Project (3,34) and the Genomes of Netherlands (GoNL) project (35,36) have adopted this strategy and have successfully identified many CNVs using an extensive panel of detection methods combined with low-throughput validation. Such a strategy increases the specificity of the calls at the cost of sensitivity.

Notably, with most of the tools and approaches described above, repeat-rich regions and other problematic regions of the genome are often removed or smoothed at some step of the analysis, to improve the accuracy of the calls. Although some methods (37,38) try to model ambiguous mapping and repeat structure, only particular situations are addressed and, as a consequence, low-mappability regions are just scarcely covered in the most recent CNV catalogs (34). This is unfortunate given that CNVs in such regions have already been associated with various diseases (12,39–42) and that these regions are also more likely variable. Indeed, different types of genomic repeats are likely to contribute to CNV formation. For example, CNVs are known to be enriched in segmental duplications (2) and short and long tandem repeats are also known to be highly polymorphic (43,44). Moreover, repeat templates, like segmental duplications or transposable elements, can facilitate the formation of CNV through non-allelic homologous recombination and other mechanisms (45).

Given these facts and the growing realization of the importance of repetitive regions in the genome (46,47), we wanted to investigate the performance of PopSV in low-mappability regions and explore the comprehensive CNV distribution across a large cohort of healthy individuals. After showing that population-based RD measures are better than existing mappability estimates to correct for variable coverage, we apply PopSV to 640 WGS individuals from three human cohorts. We compare the performance of PopSV on these datasets with existing CNV detection methods in regions of low-mappability and validate the quality of the predictions across different repeat profiles using PCR validation. Additionally, using publicly available long-read sequencing data and assemblies, we show that PopSV is able to detect some highly ambiguous CNVs. Next, having demonstrated the quality of the PopSV calls, we characterize the patterns of CNVs across the human genome and produce a CNV catalog where variants of different types are better represented compared to existing catalogs. We further find that CNVs are significantly enriched in regions of low-mappability and in different classes of repeats. Finally, we identify novel CNV regions in low-mappability regions that were absent from previous CNV catalogs and describe their impact on protein-coding genes.

MATERIALS AND METHODS

Data

Three publicly available WGS datasets were used. The first is a twin study (48) with an average depth of 40× across 45 French-Canadian individuals, including 10 families of parents and monozygotic twins. The second is a renal cell carcinoma dataset (49) (CageKid) with 95 tumor/normal pairs from four European countries and an average depth of 54×. The third contains 500 unrelated Dutch individuals from the GoNL (35) dataset with an average depth of 14×. In each study, the sequenced reads had been aligned using bwa (50). See Supplementary Information for more details on access and read processing.

Read count across the genome

The genome was fragmented in non-overlapping bins of fixed size. As a RD measure we used the number of properly mapped reads, defined as read pairs with correct orientation and insert size, and a mapping quality of 30 (Phred score) or more. In each sample, GC bias was corrected by fitting a LOESS model between the bin's RD and the bin's GC content. We used a bin size of 5 Kbp for most of the analysis. When specified, we used smaller bin sizes of 500 bp or 2 kb.

RD and mappability estimates

To compare RD and mappability estimates in the Twin study, we first removed bins with extremely high RD if deviating from the median RD by more than 5 standard deviation. The RD across the different samples were then combined and quantile normalized. For each bin, we computed the average RD and standard deviation across the samples. We downloaded the mappability track for hg19 (27) and computed the average mappability in each bin. We compared the RD in one randomly selected sample with the mappability estimates and with the inter-sample RD average. To correct for the variation explained by the mappability estimates we fitted a generalized additive model using a cubic regression spline between the mappability estimates and RD in the sample (see Supplementary Information). With these estimations and the global standard deviation we computed a Z-score for each bin. A similar set of Z-scores was computed using the inter-sample average and standard deviation. The normality of these two Z-score distributions were compared in term of excess kurtosis and skewness. The Z-score distributions were also compared in different mappability intervals. Finally, 45 samples of each cohort were combined and their RD quantile normalized. The inter-sample RD mean and standard deviation were then computed separately in each cohort and compared with the mappability estimates and RD in the selected sample.

`PopSV` approach for CNV detection

`PopSV` was first described and applied in a CNV analysis of epilepsy patients (29). Briefly, a set of samples are chosen as reference and used to guide the normalization of each bin. After normalization the average RD and standard deviation in each bin are saved and used to transform the RD in all samples into Z-scores. CNVs are called in each sample when the RD is significantly higher or lower than in the reference samples. The Z-scores can be segmented using the circular binary segmentation (51) or after statistical testing at the bin level. As recommended, `PopSV` was run separately on each dataset to avoid false positives due to potential variation in sequencing protocols. More details are available in the original publication (29) and in the Supplementary Information. With `PopSV` there is no filtering, masking, smoothing or altering of repeat-rich regions: all the regions with properly mapped reads are analyzed.

Coverage track and low-mappability regions

The average RD in the reference samples, a feature used during CNV calling, was used as a coverage track. Bins with a RD lower than 4 standard deviation from the median were classified as *low-mappability* (or *low coverage*). To highlight the most challenging region, we also defined *extremely low coverage* regions if the average RD was lower than 100 reads. We overlapped these regions with protein-coding genes and segmental duplications (see Supplementary Information), and computed the distance to the nearest centromere, telomere or assembly gap. We also counted the number of protein-coding genes overlapping at least one low-coverage region.

CNV detection using other methods

`FREEC` (16) and `CNVnator` (17) were run on each sample separately starting from the BAM files and using the same bin size as for `PopSV` (5 kb). `cn.MOPS` (18) was run on the same GC-corrected bin counts than for `PopSV` and samples from the same dataset were jointly analyzed. After retrieving split reads using `YAHA` (52), `LUMPY` (53) was run and we kept all the deletions and duplications larger than 300 bp. `BND` variants with both ends more than 300 bp apart in the same chromosome were also included as they could be CNVs lacking support to characterize their type properly. See Supplementary Information for more details.

Clustering samples using the CNV calls

The similarity between two samples is defined by the amount of sequence called in both divided by the average amount of sequence called (see Supplementary Information). This distance is used for hierarchical clustering of the samples in the Twin study using different linkage criteria (*average*, *complete* and *Ward*). The clustering was performed using calls in regions with extremely low coverage ($\leq$100 reads on average in the reference samples) only. The Rand index estimated the concordance between the clustering and the known pedigree, grouping the samples per family (see Supplementary Information).

Replication in twins

For each twin and each method, a CNV call was defined as *replicated* if also found in the other monozygotic twin but in less than 50% of the population to remove systematic errors. The frequency was computed by counting samples with any overlapping CNVs. In order to avoid missing calls with borderline significance, we used slightly less confident calls for the second twin (see Supplementary Information). For each method, we computed the number and proportion of *replicated* calls per sample. We computed these metrics using all the calls, calls in low-mappability regions only, calls in segmental duplications, calls overlapping annotated repeats and calls overlapping annotated satellites, all using a minimum overlap of 90% of the call's sequence. Finally, we computed the replication estimates for calls located at 1 Mb or less from a centromere, telomere or assembly gap.

Replication between paired normal and tumor samples

The same approach was applied in the renal cancer dataset. Here, *replicated* calls were found in a normal sample and its paired tumor but in <50% of the normal samples.

Replication estimates and reliable regions

Using CNV calls found in <50% of the population, we defined as *reliable* a 10 kb region where more than 90% of the overlapping calls were *replicated* calls. We then compared the number and proportion of reliable regions for each method and in different types of region. As before, we compared regions overlapping low-mappability regions, segmental duplications, annotated repeats, satellites, or located at less than 1 Mb from a centromere, telomere or assembly gap.

Experimental validation

A subset of variants in the Twin study were experimentally validated. First, we randomly selected one-copy and two-copy deletions, among small (~700 bp) and large (~4 kb) variants among the calls produced with 500 bp and 5 kb bins. The calls were visually inspected to design PCR primers (see Supplementary Information). We randomly selected 20 regions from those with available PCR primers. Next, we randomly selected deletions overlapping low-mappability regions and called in 6 samples or fewer. Because RD could not be used efficiently to fine-tune the breakpoints' location, we retrieved the reads (and their pairs) mapping to the region and assembled them (see Supplementary Information). We randomly selected 17 regions from those with PCR primers. In addition to gel electrophoresis, the amplified DNA of some regions was sequenced by Sanger sequencing.

Analysis of CEPH12878

High coverage PCR-free Illumina WGS data for 30 samples, including CEPH12878, was downloaded from the 1000 Genomes Project (1000GP) (34) (see Supplementary Information). PopSV was run using 5 kb bins and all the samples as reference. Using the same coverage track as before we selected all deletions in CEPH12878 overlapping low-mappability regions (at least 90% of the call). We first looked for support in CEPH12878 assemblies that used Illumina short-read sequencing, BioNano Genomics genome maps and either single molecule sequencing from the Pacific Biosciences (PacBio) platform (54) or 10× Genomics linked-read sequencing (55). For each selected deletion from PopSV, we aligned the flanking reference sequences to the assemblies using BLAST (56) (see Supplementary Information). When both flanks could be mapped to a contig, we visually inspected MUMmer plots (57) which either supported the deletion, the reference genome sequence or were too noisy to assess. We further annotated the selected calls if they overlapped with the deletions identified in Pendleton *et al.* (54) over a minimum of 1 kb. Finally, we downloaded the corrected PacBio reads and built a local assembly and consensus around each selected PopSV deletion (see Supplementary Information). We visually inspected MUMmer plots of the assembled and consensus sequences to confirm the presence of the deletion.

CNV catalog

We called CNVs separately in each cohort with PopSV using as reference samples the 45 samples in the Twin study, the normal samples in the cancer dataset and 200 samples in the GoNL dataset. For the Twin study and the renal cancer dataset, PopSV was run using 500 bp bins and 5 kb bins. Because of the lower sequencing depth, PopSV was run using 2 kb bins and 5 kb bins for the GoNL dataset. For each sample, calls from the two different runs were merged when consistent (see Supplementary Information). To compute the total number of calls, we collapsed calls with a reciprocal overlap higher than 50%. The amount of sequence affected in a genome is computed by merging all the variants in the cohort and counting the affected bases in the reference genome.

Comparison with public CNV catalogs

We retrieved autosomal deletions, duplications and CNVs from four public CNV catalogs derived from large-scale WGS surveys: the 1000GP SV catalog (34), Genome STRiP's catalog from 847 individuals (33), Genome STRiP calls in 148 high-depth WGS genomes (58), and the GoNL SV catalog (35) (see Supplementary Information). To compare the amount of CNV with PopSV, we removed deletions smaller than 300 bp as well as variants with high frequency (>80%). We compared CNV frequency between the 620 unrelated samples and a down-sampled set of 620 randomly selected individuals from the 1000GP CNV catalog. The frequency was derived for all the nucleotide that overlaps at least one CNV as the proportion of individuals with a CNV in this locus. The frequency distribution was computed separately for the different CNV types.

Comparison with CNV catalogs from long-read studies

The SV catalog from Chaisson *et al.* (59) was downloaded and overlapped with the CNV catalogs from 1000GP and PopSV results on our 640 genomes. Here, the 1000GP catalog contained deletions, duplications and CNVs of any size and frequency. Using control regions and logistic regression we tested for an enrichment of variants in the SV catalog from Chaisson *et al.* (59) (see Supplementary Information). The analysis was performed separately on deletions, duplications, low-mappability regions and extremely low-mappability regions. The same analysis was performed using the SV catalog from Pendleton *et al.* (54).

Novel CNV regions

Using the 620 unrelated individuals across the three cohorts, we selected CNVs present in more than 1% of the population (seven individuals or more) and not overlapping any CNV in the 1000GP catalog (34). We used deletions, duplications and CNVs of any size and frequency from the 1000GP. Novel CNVs were collapsed into novel CNV regions, i.e. contiguous regions in which each base is overlapped by at least one novel CNV. The novel CNV regions were annotated using the low-mappability and extremely low-mappability tracks. We also compared CNVs from the three other public CNV catalogs to the novel CNV regions.

Distance to centromere, telomere and assembly gaps

The centromeres, telomeres and assembly gaps (CTGs) were retrieved from the gap track in UCSC (60). In chromosomes with missing telomere annotation, we defined the telomere as the 10 kb region at the ends of chromosome. The distance from each variant to the nearest CTG was computed and represented as a cumulative proportion. Because this distribution changes with the size of the variants, we sampled random regions in the genome with similar sizes and computed the same distance distribution (see Supplementary Information). Thanks to this null distribution we were able to see if variants were located closer/further to CTG than expected by chance.

Enrichment in genomic features

We tested for CNV enrichment in different genomic features: genes, exons, low-mappability regions, segmental duplications, satellites, simple repeats and transposable elements. The different satellite families, frequent simple repeat motives, transposable element families and sub-families were also tested. For each sample, we computed a fold-enrichment as the fold change in proportion of regions overlapping a feature between CNV and control regions (see Supplementary Information). The significance was assessed using logistic regression on the CNV and control regions. To control for the enrichment in segmental duplications we used control regions with similar overlap profile (see Supplementary Information). We also added a variable representing the overlap with segmental duplications as a cofactor in the logistic regression model. When numerous tests were performed, e.g. satellite families, simple repeat motives, transposable element families or sub-families, the *P*-values were corrected for multiple testing using Benjamini-Hochberg procedure. Finally, for each CNV and control region, we computed the proportion of the region overlapped by satellites, simple repeats and transposable elements.

Overlap with gene annotation

Exons of protein-coding genes and promoter regions (10 kb upstream of the transcription start site) were extracted from the Gencode annotation v19. We counted how many genes overlapped a CNV in the population when considering exons only, exons and promoter region, or gene body and promoter region. In addition, we computed these numbers using only genes associated with a disease or phenotype in the OMIM Morbid Map (Online Mendelian Inheritance in Man; http://omim.org/). These numbers were also computed for CNVs that overlapped >90% of various classes of repeats. For example, Satellite-CNVs are CNVs with >90% of their region annotated as satellites.

RESULTS

Modeling RD using population-based measures instead of mappability scores

When counting uniquely mapped reads, the mappability of a region is a major predictor of the observed RD. Theoretical mappability estimates (27) strongly correlated with the RD in a sample but many regions with intermediate mappability diverged from the predicted levels of RD (Supplementary Figure S1A). By computing the average RD across the 45 samples from the Twin study in each 5 kb bin we found that this divergence is consistent across samples and not simply due to a high RD variance (Figure 1A). These mappability estimates only approximate RD variation and cannot explain the RD profile in numerous regions. In contrast, population-based metrics more directly estimate the expected RD level (Supplementary Figure S1B). Similarly to what was done in Monlong *et al.* (29) in high-mappability regions, we hypothesized that population-based estimates of RD mean and standard deviation could be used directly and help analyze regions with reduced RD. To test this hypothesis, *Z*-scores corrected by the mappability-based estimates were compared to *Z*-scores derived from both the inter-sample mean and standard deviation. The population-based *Z*-scores better followed a Normal distribution with an excess kurtosis of 0.2 and skewness of 0.004 compared to 29.4 and –2.284 respectively for mappability-adjusted *Z*-scores (Figure 1B). The distribution of the population-based *Z*-scores was also more stable across the mappability spectrum (Figure 1C). When comparing samples from the three different datasets, we noticed cohort-specific profiles in term of RD level and variance even though RD had been quantile normalized (Figs S1C and D), suggesting that population-based estimates will be better at capturing subtle cohort-specific variation.

These results suggest that a population-based strategy such as PopSV (29) could be extended to investigate CNVs in regions of low-mappability. To define low-mappability regions in the population, we used the average RD in the reference samples track produced by PopSV. In the Twin study for example, 12.6% of the covered 5 kb bins were labeled as low-coverage (Figure 1D), more than half of which were regions with extremely low coverage (lower than 100 reads on average). Slightly fewer regions were labeled as low-coverage in the other cohorts (Supplementary Figure S2). As expected, low-coverage regions were depleted in gene content with only 15.3% of the 5 kb bins in these regions overlapping a protein-coding gene versus 48.8% for other regions. Nonetheless, 4044 protein-coding genes overlapped a low-coverage region. Finally, 23.2% of the low-mappability regions overlapped segmental duplications and 69.1% were located at less than 1 Mb from a centromere, telomere or assembly gap, versus respectively 2.9% and 8.8% for other regions.

Replication rates in regions of low-mappability

We previously demonstrated that CNV detection with PopSV was overall more sensitive than FREEC (16), CNVnator (17), cn.MOPS (18) and LUMPY (53) methods (29). In the following, we focused on the performance of PopSV in low-mappability regions. We first investigated the general concordance of the CNV calls with the pedigree in the Twin study. Using calls in extremely low-mappability regions (average RD below 100 reads) only, we clustered the individuals and compared the result to the known pedigree. We found that PopSV showed better concordance, as assessed by the Rand index (Supplementary Figure S3), compared to the other methods. Indeed, the clustering dendogram from PopSV calls, even in these challenging regions, captured almost perfectly the family relationships (Figure 2A). We then investigated if the call replication rate was stable across different mappability profiles. Using calls present in <50% of the population to avoid systematic bias, the overall replication rate in the other twin was found to be 89.7%. Focusing on calls in low-coverage regions, we found a comparable replication rate of 92.5%. The replication rate remained constant in regions with different repeat profiles (Figure 2B) such as regions overlapping segmental duplication, annotated repeats, or close to centromeres, telomeres and assembly gaps. In contrast, the other methods showed a reduced replication and higher variance in repeat-rich regions. The superior replication rate was complemented by a larger number of calls: PopSV called between 2.7 and 9.9 times more replicated CNVs per sample in low-coverage regions compared to the other methods. We observed the same results in the cancer dataset when comparing the agreement between germline events in normal/tumor pairs.

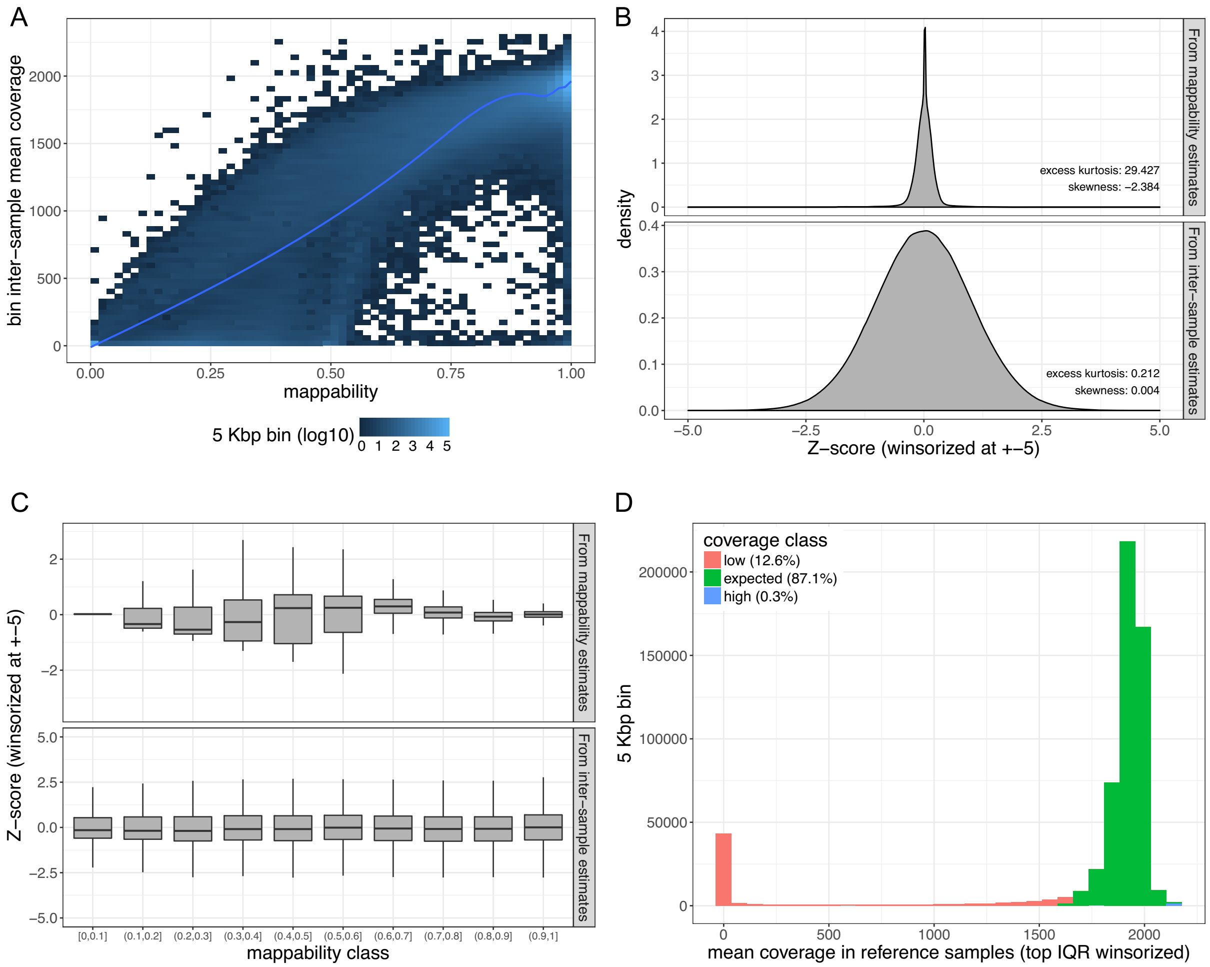

Figure 1. Mappability and population-based RD estimates. (**A**) Inter-sample mean RD and average mappability in 5 kb bins. Regions with the same mappability estimate can have different RD levels. (**B**) *Z*-score distribution. In *mappability*, *Z*-scores were computed from the mappability-predicted RD and global standard deviation; In *population estimates* from the inter-sample mean and standard deviation. (**C**) *Z*-score distribution across the mappability spectrum. (**D**) Average RD in the Twin study. The right-tail of the histogram was winsorized using the IQR and the different coverage classes are shown with colors.

PopSV had between 1.8 and 17.8 times more calls in low-mappability regions compared to the other methods and a stable replication rate across repeat profiles (Supplementary Figure S4). We next wanted to assess the performance in each region of the genome, rather than overall rates per sample, and used the replication in twins to identify regions with reliable calls. Again we observed that PopSV was as reliable overall as in regions with different repeat profiles (Figure 2C). This analysis also showed that PopSV provides reliable calls in a larger fraction of the genome compared to other methods. The strongest gain was observed for regions overlapping satellites or overlapping almost completely annotated repeats, with around twice as many regions reliably called by PopSV. cn.MOPS showed the second best performance, especially in regions overlapping segmental duplications or close to centromeres, telomeres and assembly gap.

Validation of CNVs in regions of low-mappability

Using Real-Time PCR validation across 151 regions, we previously demonstrated that the replication estimates from the Twin dataset are consistent with experimental validation (29). We had tested variants of different types, sizes and frequencies and validated 90.7% of the calls, similar to our twin-based replication estimates. Here we tested additional deletions in individuals from the Twin study using PCR validation. We first validated randomly selected deletions and found a validation rate close to the overall replication rate, with 18 out of 20 deletions (90%) successfully validated (Supplementary Table S1). In a second validation batch, we focused on rare deletions in low-mappability regions, of which 11 out of the 17 (65%) were successfully validated (Supplementary Table S2). We noticed that the majority of the non-validated deletions were predicted to be smaller than 100 bp and most likely due to a problem during the breakpoint fine-tuning. If we consider only deletions larger than 100 bp, the validation rate in regions of low-mappability increased to 83% (10/12) once again close to PopSV's replication rates in the Twin dataset.

Regions with extreme repeat content remained difficult to target and validate using PCR approaches. To further interrogate the performance of PopSV in those regions, we

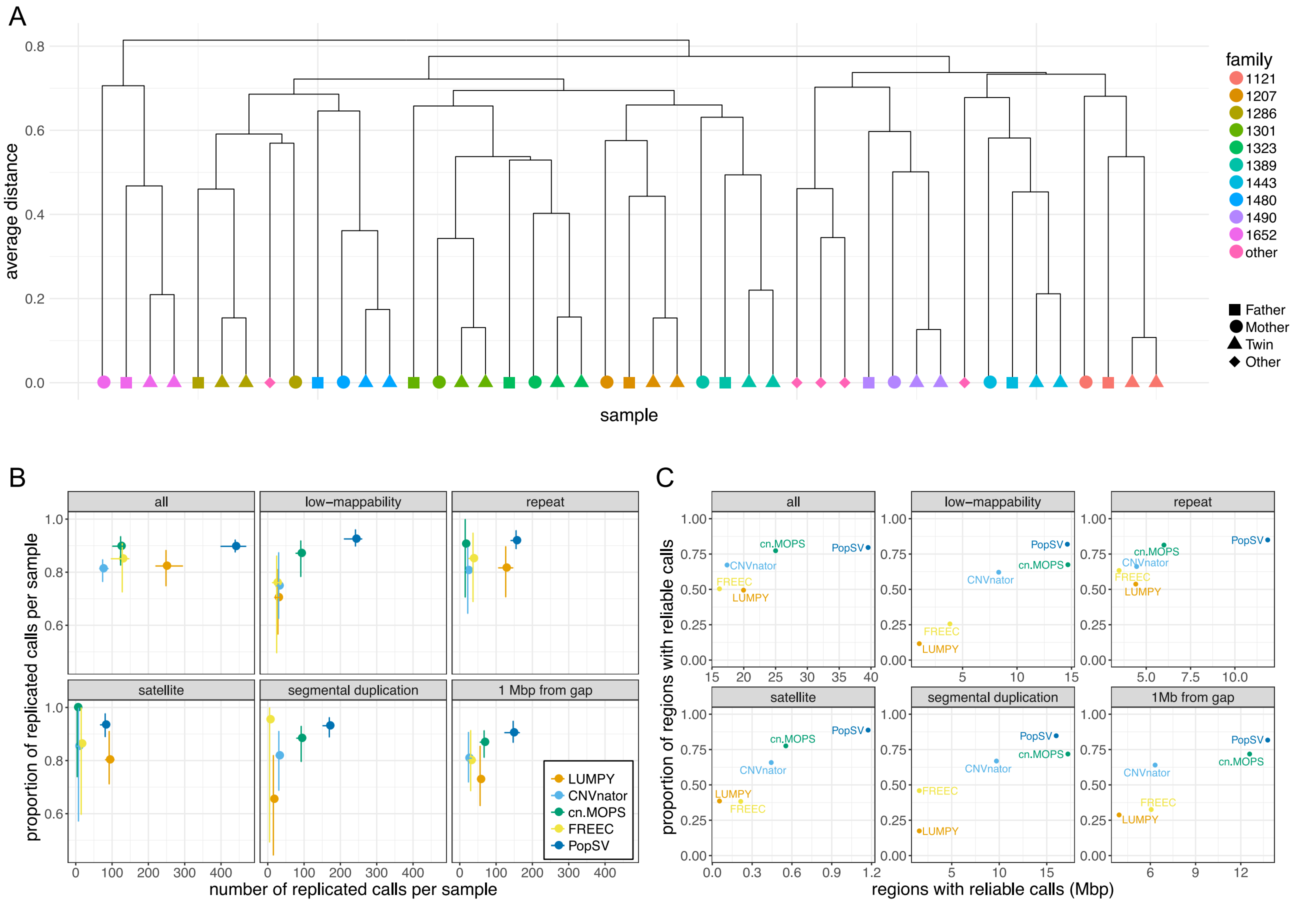

Figure 2. PopSV's performance in low-mappability regions. (**A**) Cluster using PopSV calls in extremely low coverage regions (below 100 reads). (**B**) Proportion and number of calls replicated in the monozygotic twin. The point shows the median value per sample, the error bars the 95% confidence interval. (**C**) Proportion and number of regions with reliable calls, computed from call replication in twins.

turned to whole-genome data from long-read sequencing technology. Publicly available assemblies for CEPH12878 samples confirmed several deletions called by PopSV in low-mappability regions. Out of the 14 homozygous deletions that could be assessed, 13 were confirmed in a contig, 12 of which were observed in both assemblies (54,55). Only one region seemed to be a false positive, an assembled contig supporting the reference sequence in one assembly. Eleven regions could not be assessed because the flanks in the reference genome didn't map to any assembled contigs or their MUMmer plots neither supported a deletion nor the reference sequence. In summary, we confirmed 92.8% of the homozygous deletions in low-mappability regions that could be compared with the assemblies. Deletions can be confirmed by direct comparison of the variant region and, if homozygous, should be present in the assembly. In contrast, heterozygous deletions could be missing from an assembly if only the reference allele was assembled. We confirmed 27 out of the 44 heterozygous deletions in low-mappability regions that could be assessed (Supplementary Table S3). As expected, only one allele was supported for many regions: 16 regions with only the deleted allele observed and 17 regions with only the reference allele observed. Both deleted and reference alleles were observed for 11 variants. Although only 61.3% of the heterozygous deletion were confirmed, many variants might have been missed because of assembly preference to one allele, as suggested by the similar number of regions with only one supported allele. Using variants identified by Pendleton *et al.* (54) and by assembling raw PacBio reads, we found support for three additional homozygous deletions and 15 heterozygous deletions that had remained inconclusive in the assembly comparison. Most of the regions that couldn't be confirmed were located close to assembly gaps in the reference genome (Supplementary Figure S5). This observation highlighted that even with long-read sequencing data, it is not straightforward to clearly assess some genomic regions close to assembly gaps.

Global patterns of CNVs across the human genome

Having demonstrated the robustness of PopSV in low-mappability regions, we wanted to characterize the global patterns of CNVs across the human genome. We were especially interested in looking at calls in regions of low-mappability which represents between 9% and 12% of the human genome (Figure 1D and Supplementary Figure S2). We started with an analysis of the twins and the normal samples in the renal cancer dataset, both of which have an average sequencing depth ∼40×. PopSV was used to call

CNV using 500 bp and 5 kb bins, which were then merged to create a final set of variants. On average per genome, 7.4 Mb of the reference genome had abnormal read coverage, 4 Mb showing an excess of reads indicating duplications and 3.4 Mb showing a lack of reads indicating deletions (Table 1). In both datasets, the average variant size was around 3.7 Kbp and 70% of the variants found were smaller than 3 Kbp. We compared our numbers to equivalent CNVs detected in the most recent human SV catalog from the 1000 Genomes Project (1000GP), where 6.1 Mb was found to be copy-number variable on average in each genome (Supplementary Table S4). In those calls, we notice that no variants except for a few deletions were identified in regions of extremely low-mappability regions. Similarly, small duplications (<3 kb) were absent from that catalog. In contrast, the set of variants identified by PopSV included variants in extremely low-mappability regions as well as small deletions and duplications (Table 1), explaining in part the ∼20% increase in affected genome. While the study from the 1000GP (34) explored a wider range of SVs, our catalog is likely more representative of the distribution of CNVs in a normal genome since a larger portion of the genome could be analyzed. Small duplications and events in low-mappability regions were also under-represented in more recent CNV surveys that used higher sequencing depth or joint-calling of CNVs (33,35,58) (Supplementary Table S4), confirming the uniqueness of the PopSV catalog.

Next, we applied PopSV to the 500 unrelated samples from the GoNL cohort (Table 1). Due to a lower sequencing depth (∼13×), we used bins of size 2 and 5 kb, explaining the lower number of variants found in these samples. Nevertheless, a large sample size helps better characterize the frequency patterns and provides a more comprehensive map of rare CNVs. In total, across these three cohorts, 325.6 Mb were found to be affected by a CNV with more duplications (50 856) detected than deletions (44 110). This contrasts with the CNVs reported by the 1000GP (34) that were heavily skewed towards deletions (Supplementary Table S4), likely due to the conservative ensemble approached used to detect CNVs. The frequency distribution of deletions and duplications found using PopSV were also much more balanced compared with the ones from the 1000GP (34) (Figure 3A).

We also compared our CNV catalog with an orthogonal set of calls from Chaisson *et al.* (59) that were obtained using long-read sequencing. Although these calls came from a different genome, we expect both catalogs to share a number of common variants. We found a significant overlap between the two catalogs, overall and separately for deletions, duplications, low-mappability regions and extremely low-mappability regions (Figure 3B). In all categories, the overlap was stronger for PopSV's catalog compared to the 1000GP CNV catalog. We noted that the enrichment for the 1000GP catalog disappeared for duplications and low-mappability regions but was even stronger for PopSV's catalog. Like PopSV, the long-read sequencing study (59) also found a better balance between deletions and duplications. Similar observations were made using another set of calls from long-read sequencing of the CEPH12878 sample (54) (Supplementary Figure S6).

CNVs are enriched near centromeres and telomeres and in regions of low-mappability

Large CNVs have been shown to be enriched near centromeres, telomeres and assembly gaps (CTGs) (61). We were interested in exploring this observation further using the set of high resolution calls from PopSV. We compared the distribution of CNVs calls made across the three datasets to randomly distributed regions of similar sizes (Supplementary Figure S7). In an average genome, we found that 33.5% of the CNVs calls were within 1 Mbp of a CTG, while we would have expected only 11.2% by chance. To verify that these observations were not simply a consequence of the methodology used, we also looked at the somatic CNVs (sCNVs) that we could detect in the renal cell carcinoma dataset. For this purpose, we extracted the variants found by PopSV in the tumor sample of an individual but missing from its paired normal sample. Reassuringly, and in contrast to germline CNVs, sCNVs were not preferentially found near CTGs (Supplementary Figure S7), with 11.1% of the sCNVs within 1 Mb of a CTG.

After correcting for the distance to CTGs, we also observed a 4.7-fold-enrichment of variants in regions of low mappability (Figure 4A). Segmental duplications (SD), DNA satellites and Short Tandem Repeats (STR) were also significantly enriched with fold-enrichment of 3.6, 2.6 and 1.2, respectively. The over-representation of CNVs in SDs has been described before (2) and in a recent study (62), half of the CNV base pairs were shown to overlap a SD. To investigate the contribution of low-mappability regions beyond SDs, we used matched control regions and included segmental duplication overlap in the logistic regression model. Even after controlling for this known enrichment, we found that CNVs overlapped low-coverage regions more than twice as much as expected (Supplementary Figure S8A). This two-fold enrichment is independent of the SD association and consistently observed in the three cohorts of normal genomes. In contrast to germline CNVs, sCNVs were once again found to be more uniformly distributed (Figure 4A and Supplementary Figure S8A). These results suggest that the enrichments of germline CNVs near CTGs and in regions of low-mappability are unlikely to be the result of a methodological artifact.

Various repeat families are more prone to harbor CNVs

We wanted to further characterize the distribution of germline CNVs in relation to different repeat classes and families. By comparing CNVs to the same control regions with matched overlap with SD and distance to CTGs we can look for patterns that are specific to repeat sub-families without the risk of being biased by the global enrichments (Figure 4B). Using this approach, we found that CNVs were still significantly enriched in satellites repeats and in short tandem repeats (STRs) (P-value $< 10^{-4}$, Supplementary Figure S8A), with fold-enrichments of 2.3 and 1.2 respectively.

Although it is known that DNA satellites and simple repeats are more unstable (63), the extent to which CNVs are found in these regions in humans had, to our knowledge, not been systematically explored. Satellite repeats are grouped into distinct families depending on their repeated

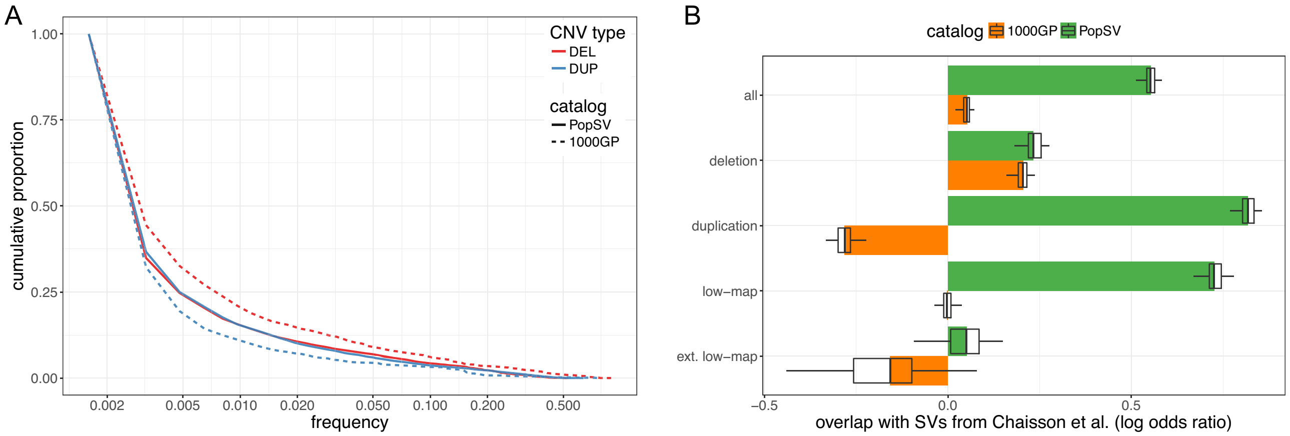

Figure 3. Comparison with CNV catalogs from the 1000 Genomes Project (34) (1000GP) and a long-read sequencing study (59). (**A**) The x-axis represents the proportion of individuals with a CNV overlapping a region. The y-axis represents the cumulative proportion of the affected genome. (**B**) Overlap with the SV catalog from Chaisson *et al.* (59). In each cohort (color), the proportion of collapsed calls overlapping calls from Chaisson *et al.* (59) or control regions with similar size distribution was modeled using a logistic regression. Boxplots show variation across 50 sampling of control regions. *low-map*: calls in low-mappability regions; *ext. low-map*: calls in extremely low-mappability regions.

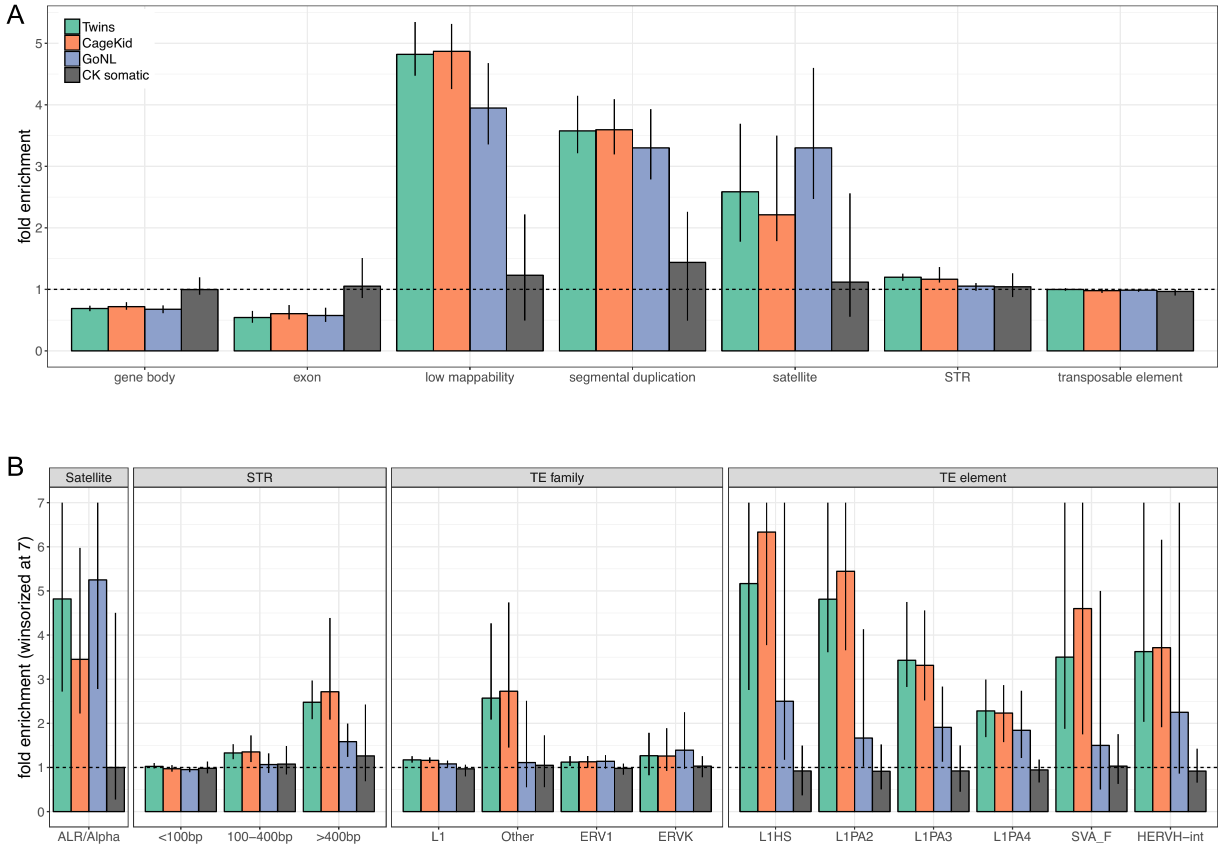

Figure 4. CNVs in normal genomes. (**A**) Enrichment of CNVs in different genomic classes (x-axis) across different cohorts (colors) and controlling for the distance to centromere/telomere/gap. Bars show the median fold enrichment compared to control regions. The error bar represents 90% of the samples in the cohort. (**B**) Enrichment of CNVs in repeat families (x-axis) controlling for the overlap with segmental duplication and distance to centromere/telomere/gap. The error bars were winsorized at 7 for clarity. *STR: Short Tandem Repeat*; *TE: Transposable Element*.

Table 1. CNVs in the Twins, CageKid normals and GoNL datasets. WG: whole genome; ELC: extremely low-coverage regions. The *Total* number of variants is the total number after collapsing recurrent variants. *Affected genome* represents the amount of the reference genome that overlaps at least one CNV

			Variants				Variants <3 Kbp		Affected genome (Mbp)			
Set	Depth	Samples	Total	Per sample		Avg Size (Kbp)	Proportion	Per sample	Total	Per sample		
				WG	*ELC*					*min*	*mean*	*max*
Twin study	42x	45	20 222	1 637.27	243.24	4.21	0.65	1056.84	62.22	5.30	6.89	9.03
deletion			10 661	727.04	13.20	4.53	0.58	423.80	33.97	2.79	3.30	3.85
duplication			10 396	910.22	230.04	3.94	0.70	633.04	34.20	2.50	3.59	5.29
CageKid normals	40x	95	56 256	2132.81	336.46	3.58	0.71	1 521.16	134.77	5.53	7.63	10.24
deletion			25 367	805.08	12.74	4.30	0.63	508.56	70.65	2.65	3.46	7.26
duplication			32 356	1327.73	323.73	3.14	0.76	112.60	76.28	2.31	4.17	6.70
GoNL	13x	500	27 945	549.52	81.97	8.71	0.46	250.24	226.50	3.05	4.79	8.16
deletion			13 818	262.41	1.45	8.50	0.42	110.16	106.83	1.30	2.23	3.96
duplication			15 291	287.10	80.52	8.91	0.49	140.08	139.21	1.45	2.56	5.72

unit and we found that not all satellite repeats were equally likely to overlap a CNV (Supplementary Figure S8B). In particular, Alpha satellites have the highest and most significant enrichment (P-value $< 10^{-5}$), with more than three times more CNVs than in the control regions (Figure 4B). We noted that satellites tend to span completely CNVs (Supplementary Figure S9), suggesting that satellites are likely directly involved in the CNV formation. Short and long tandem repeats can be highly polymorphic (43,44). Constrained by read length, recent studies (64,65) focused on variation of STRs smaller than 100 bp. In our analysis we found that CNVs were significantly enriched in the largest annotated STRs (>100 bp or >400 bp, Figure 4B). STR can be grouped by motif and we further tested the largest and most frequent families (Supplementary Figure S8C). Except for the weak enrichment in *AT* (*TA*) repeats, the STR enrichment appeared mostly independent of the repeat motif. Here the repeats tend to overlap just a fraction of the variant, but a clear subset of the variants are fully covered by these tandem repeats (Supplementary Figure S9). Finally, although transposable elements (TEs) as a whole did not show enrichment (Figure 4A), the 'Other' repeat class, which contains SVA repeats, was found to be significantly enriched in the two higher depth datasets (Figure 4B). Moreover, looking at TEs at the level of individual repeat families, we found a number of them to be significantly enriched including SVA F or L1Hs. Notably, HERV-H, an older ERV sub-family, was also in the list of enriched TEs. This sub-family has been shown to be expressed and important in human embryonic stem cells (66,67). Alu elements contributed to the formation of human segmental duplications (68) and are often found around SV breakpoints (69) but this TE family was not enriched in CNVs in our data. On the other hand, several families of L1 repeats older than the still active L1HS family were also enriched (e.g. L1PA2 to L1PA4) and often implicated in what appears to be non-allelic homologous recombination (see examples in Supplementary Figure S10). Reassuringly, the somatic CNVs once again did not show any of these enrichments (Figure 4B).

Impact of CNVs in regions of low-mappability

Compared to the latest 1000GP catalog (34), we identified 3455 novel regions with CNVs in more than 1% of the population. 81.3% of these regions were located in low-mappability regions while 18.4% were located in extremely low-mappability regions. These novel CNV regions were missing from the 1000GP catalog and also mostly absent in other recent CNV surveys; only 7.9–15.1% of the novel regions overlapped with a CNV in three recent CNV catalogs (33,35,58) (Supplementary Figure S11). Among the regions with a CNV in the CEPH12878 sample, we identified a deletion in the second intron of the *TRIM16* gene that was found by both Pendleton *et al.* (54) and `PopSV`. Across the 640 individuals analyzed by `PopSV`, 12% carried the variant. Thanks to the long-read data, the exact breakpoints had been pinpointed in Pendleton *et al.* (54) and it was in fact a SVA-F transposable element located within the 6 kb intron in the reference genome but absent from the assembled sequence. SVA-F is one of the youngest repeat family in the human genome and their high similarity remains a challenge for CNV analysis. Furthermore, the variant is located within a segmental duplication with 98.5% similarity and absent from public catalogs such as the 1000GP or GoNL. Another deletion supported by both public assemblies and local reassembly of the PacBio read was located 12 kb downstream of *TMPRSS11E*. 6.6% of the individuals carried the variant in the `PopSV` catalog. The assembled sequence helped pinpoint the breakpoints to an annotated L1PA2 in the reference genome. The variant was also located in a segmental duplication and absent from public catalogs such as the 1000GP or GoNL. Finally, a deletion affecting 8 different exons from the *CR1* gene was found by both Pendleton *et al.* (54) and `PopSV` in CEPH12878. *CR1* has been associated with Alzheimer disease (70) and is located within embedded segmental duplications with high similarity. The deletion was present in 3% of the population analyzed with `PopSV` but is absent from public CNV catalogs.

Overall, 7206 protein-coding genes were found to have an exon overlapping a variant in at least one of the 640 normal genomes studied (Table 2). If we included the promoter regions (10 Kbp upstream of the transcription start site), at least 11 341 protein-coding genes were potentially

Table 2. Impact of CNVs on protein-coding genes. The *CNVs* number represents the number of different CNVs, after collapsing CNVs with more than 50% reciprocal overlap. Repeat CNV: more than 90% of the CNV is annotated as repeat. Genes are protein-coding genes and the promoter region is defined as the 10 kb region upstream of the transcription start site. *Novel CNVs* are located within regions annotated as novel compared to the 1000 Genome Project catalog

		Genes with CNVs			OMIM genes with CNVs		
Set	CNVs	Exon	+ Promoter	+ Intron	Exon	+ Promoter	+ Intron
All CNVs							
All	91 735	7206	11 341	13 259	1 241	1 857	2 196
Low coverage	32 707	848	1491	2 648	95	160	371
Extremely low coverage	9 348	304	401	442	11	14	25
TE	20 491	164	1747	3 998	29	233	664
STR	4 285	45	286	748	5	39	129
Satellite	1822	2	21	33	0	0	0
Novel CNVs							
All	17 046	418	680	1 102	38	59	135
Low coverage	15 263	347	560	894	29	47	111
Extremely low coverage	6591	189	263	285	5	6	8
TE	3 896	17	192	504	1	12	66
STR	1806	14	81	230	0	9	41
Satellite	890	1	4	5	0	0	0

affected by at least one CNV in the population. Focusing on regions of low-mappability, we found 4285 different CNVs that were completely included in regions annotated as STR. These STR-CNVs overlapped the coding sequence of 45 protein-coding genes, and 286 genes when including the promoter region (Table 2). In contrast, for CNVs included in satellite regions, only 21 genes had an exon or the promoter region overlapping one of the 1822 Satellite-CNVs. Finally, we focused on CNVs that were novel compared to the 1000GP (34) and in low-mappability regions. Even there, 347 genes were found to have an exon overlapping such CNVs and this number increased to 560 when including the promoter regions. Out of these 347 genes, 29 were previously associated to a mendelian disorder or phenotype in the OMIM database (Online Mendelian Inheritance in Man; http://omim.org/, Supplementary Table S5).

DISCUSSION

Despite the strong interest in CNVs because of their role in diseases, detecting them accurately has remained a challenge, especially in regions of low-mappability. This is mostly due to technical variation in RD that cannot be fully modeled by mappability estimates. Using a recently developed CNV-calling approach that relies on a set of reference samples to estimate the expected RD (29), we show that it is possible to accurately detect CNVs across the genome, even in repeat-rich regions. Indeed, using monozygotic twins and normal/tumor pairs, we were able to demonstrate that the performance of PopSV was stable and in most cases superior to other methods across different types of low-mappability regions. Although experimental validation can be challenging in these regions, we were able to confirm a number of deletions using PCR validation as well as variants in some of the most difficult regions by taking advantage of public datasets from long-read sequencing studies.

Notably, using PopSV on 140 normal genomes with high sequencing depth (∼40×) and 500 additional samples with medium coverage (∼13×), we found that regions of low mappability, which only represent ∼10% of the genome, were around 5 times more likely to harbor CNVs. The fact that this enrichment was observed for germline events and not somatic events was both reassuring and interesting because of the implications on the selection forces at play. In particular, we were able for the first time to quantify the extent to which some regions in the genome are more prone to harbor such structural rearrangements. For instance, beyond the known enrichment in segmental duplications, we found genome-wide enrichments for different families of DNA satellites, simple repeats and TE, such as SVA, L1Hs and HERV-H. Moreover, although PopSV doesn't fully characterize STR variation, it was able to detect CNVs in large annotated STRs. These CNVs could complement the output of STR detection methods that look for STR variation within sequencing reads and for this reason cannot test STRs longer than ∼100 bp. Here, we found a strong CNV enrichment in STRs larger than 400 bp suggesting that large STRs should be included in genome-wide STR variation screens. Overall, having a more complete CNV catalog enabled an unbiased characterization of the CNV patterns across the genome and could potentially increase the power for trait-association studies.

Fine-tuning the location of breakpoints is often possible by reanalyzing the local read coverage or using orthogonal methods such as split-read or local assembly. In repeat-rich regions however, these methods generally do not perform well. Long read sequencing is currently the only experimental method that actually results in unambiguous SV calls with nearly quasi-base-pair resolution in low-mappability regions. Indeed, recent studies using long-read sequencing (54,59) found many novel SVs and highlighted variation involving complex repetitive DNA. The increased resolution and ability to span repeated regions expanded existing SV catalogs but only a handful of genomes have been sequenced in this way so far due to the higher cost of this technology. Although breakpoint and allele characterization is limited with short reads, we were able to detect the presence of such CNVs across a large population of normal genomes. Compared to previous studies, our CNV catalog strongly overlaps with the variants found by long-read sequencing studies in low-mappability regions. With hundreds of genomes at our disposal we identified frequent CNVs in repeat-rich regions that had escaped previous population-scale surveys. In the CEPH12878 sample, we independently identified low-mappability variants and showed that some novel deletions were recurrent in our cohort. For example, an exonic deletion in the *CR1* gene absent from public CNV catalogs was identified by the long-read sequencing and found in ∼3% of the samples tested by PopSV. *CR1* has been associated with Alzheimer Disease (70) thus this exonic deletion in a low-mappability region might be relevant for association studies. Using our full CNV catalog, we identified 3455 novel regions that were not present in 1000G public SV database (34) but found in more than 1% of our 640 genomes. These regions overlapped exons of 418 protein-coding genes, 38 of which were associated with a disease phenotype in the OMIM database. The amount of genes hit by CNVs in novel or low-mappability regions and the enrichment of CNVs in repeat-rich regions suggest that they be included in genome-wide surveys. As other types of variant are likely enriched in repeat-rich regions, we anticipate that population-based methods, such as PopSV, will facilitate the identification not only of CNVs but also of other types of SVs in both normal and cancer genomes.

One of the most promising future development of PopSV to further characterize low-mappability regions is its extension to detect balanced SV such as inversions or translocations. Indeed, instead of modeling the coverage of properly mapped reads, the same population-based strategy could test for an excess of discordant reads. By counting the number of reads in incorrect orientation or joining distant regions, one could recognize an excess of SV-supporting reads from discordant mapping caused by repeats. Such an approach could detect inversions and translocations that contains repeats around their breakpoints or complement SV calls from orthogonal approaches by providing a robust confidence score based on abnormal read coverage.

DATA AVAILABILITY

The PopSV R package and documentation are available at http://jmonlong.github.io/PopSV/. The scripts and instructions to reproduce the graphs and numbers in this study have been deposited at http://github.com/jmonlong/reppopsv/ and archived in https://doi.org/10.5281/zenodo.1241137.

The CNV catalog and annotations were deposited at https://figshare.com/s/8fd3007ebb0fbad09b6d. The raw sequences of the different datasets had already been deposited by their respective consortium (see Supplementary Information).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to the team of the Québec Study of Newborn twins who provided the twin dataset and the Cagekid consortium who provided the renal cancer dataset. This study also made use of data generated by the Genome of the Netherlands Project. A full list of the investigators is available from www.nlgenome.nl. Funding for the project was provided by the Netherlands Organization for Scientific Research under award number 184021007, dated 9 July 2009 and made available as a Rainbow Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL). The sequencing was carried out in collaboration with the Beijing Institute for Genomics (BGI). Finally, we would like to thank Simon Gravel, Mathieu Blanchette, Mathieu Bourgey and Toby Dylan Hocking for helpful discussions.

FUNDING

National Sciences and Engineering Research Council [NSERC-448167-2013]; Canadian Institute for Health Research [CIHR-MOP-115090]; Fonds de Recherche Santé Québec [FRSQ-29493, FRSQ-25348 to S.L.G., G.B.]. Data analyses were enabled by compute and storage resources provided by Compute Canada and Calcul Québec. Funding for open access charge: NSERC [NSERC-448167-2013].
Conflict of interest statement. None declared.

REFERENCES

1. Hall,I.M. and Quinlan,A.R. (2012) Detection and interpretation of genomic structural variation in mammals. In: *Methods in molecular biology*. Springer Science, Clifton, Vol. **838**, pp. 225–248.
2. Sharp,A.J., Cheng,Z. and Eichler,E.E. (2006) Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 407–442.
3. Mills,R.E., Walter,K., Stewart,C., Handsaker,R.E., Chen,K., Alkan,C., Abyzov,A., Yoon,S.C., Ye,K., Cheetham,R.K. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.
4. Pang,A.W., MacDonald,J.R., Pinto,D., Wei,J., Rafiq,M.A., Conrad,D.F., Park,H., Hurles,M.E., Lee,C., Venter,J.C. *et al.* (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.*, **11**, R52.
5. McCarroll,S.A., Huett,A., Kuballa,P., Chilewski,S.D., Landry,A., Goyette,P., Zody,M.C., Hall,J.L., Brant,S.R., Cho,J.H. *et al.* (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.*, **40**, 1107–1112.
6. Stone,J.L., O'Donovan,M.C., Gurling,H., Kirov,G.K., Blackwood,D.H.R., Corvin,A., Craddock,N.J., Gill,M., Hultman,C.M., Lichtenstein,P. *et al.* (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, **455**, 237–241.
7. Bochukova,E.G., Huang,N., Keogh,J., Henning,E., Purmann,C., Blaszczyk,K., Saeed,S., Hamilton-Shield,J., Clayton-Smith,J., O'Rahilly,S. *et al.* (2010) Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, **463**, 666–670.
8. Mefford,H.C., Yendle,S.C., Hsu,C., Cook,J., Geraghty,E., McMahon,J.M., Eeg-Olofsson,O., Sadleir,L.G., Gill,D., Ben-Zeev,B. *et al.* (2011) Rare copy number variants are an important cause of epileptic encephalopathies. *Ann. Neurol.*, **70**, 974–985.
9. Stefansson,H., Meyer-Lindenberg,A., Steinberg,S., Magnusdottir,B., Morgen,K., Arnarsdottir,S., Bjornsdottir,G., Walters,G.B., Jonsdottir,G.A., Doyle,O.M. *et al.* (2014) CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature*, **505**, 361–366.
10. Beroukhim,R., Mermel,C.H., Porter,D., Wei,G., Raychaudhuri,S., Donovan,J., Barretina,J., Boehm,J.S., Dobson,J., Urashima,M. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.
11. Balzola,F., Bernstein,C., Ho,G.T. and Lees,C. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases

and 3,000 shared controls: Commentary. *Inflamm. Bowel Dis. Monitor*, **11**, 26–27.

12. Ayarpadikannan,S. and Kim,H.-S. (2014) The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases. *Genomics Informatics*, **12**, 98.
13. Alkan,C., Coe,B.P. and Eichler,E.E. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.
14. Chen,K., Wallis,J.W., McLellan,M.D., Larson,D.E., Kalicki,J.M., Pohl,C.S., McGrath,S.D., Wendl,M.C., Zhang,Q., Locke,D.P. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.
15. Lindberg,M.R., Hall,I.M. and Quinlan,A.R. (2014) Population-based structural variation discovery with Hydra-Multi. *Bioinformatics* **31**, 1286–1289.
16. Boeva,V., Zinovyev,A., Bleakley,K., Vert,J.P., Janoueix-Lerosey,I., Delattre,O. and Barillot,E. (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, **27**, 268–269.
17. Abyzov,A., Urban,A.E., Snyder,M. and 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.
18. Klambauer,G., Schwarzbauer,K., Mayr,A., Clevert,D.A., Mitterecker,A., Bodenhofer,U. and Hochreiter,S. (2012) Cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.*, **40**, e69.
19. Ye,K., Schulz,M.H., Long,Q., Apweiler,R. and Ning,Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.
20. Rimmer,A., Phan,H., Mathieson,I., Iqbal,Z., Twigg,S. R.F., Wilkie,A. O.M., McVean,G. and Lunter,G. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.*, **46**, 912–918.
21. Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, e72.
22. Treangen,T.J. and Salzberg,S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36–46.
23. Teo,S.M., Pawitan,Y., Ku,C.S., Chia,K.S. and Salim,A. (2012) Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, **28**, 2711–2718.
24. Koren,A., Handsaker,R.E., Kamitaki,N., Karlić,R., Ghosh,S., Polak,P., Eggan,K. and McCarroll,S.A. (2014) Genetic variation in human DNA replication timing. *Cell*, **159**, 1015–1026.
25. van Dijk,E.L., Jaszczyszyn,Y. and Thermes,C. (2014) Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.*, **322**, 12–20.
26. Cheung,M.S., Down,T.A., Latorre,I. and Ahringer,J. (2011) Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res.*, **39**, e103.
27. Derrien,T., Estellé,J., Marco Sola,S., Knowles,D.G., Raineri,E., Guigó,R. and Ribeca,P. (2012) Fast computation and applications of genome mappability. *PLoS One*, **7**, e30377.
28. Scheinin,I., Sie,D., Bengtsson,H., van de Wiel,M.A., Olshen,A.B., van Thuijl,H.F., van Essen,H.F., Eijk,P.P., Rustenburg,F. *et al.* (2014) DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res.*, **24**, 2022–2032.
29. Monlong,J., Girard,S.L., Meloche,C., Cadieux-Dion,M., Andrade,D.M., Lafreniere,R.G., Gravel,M., Spiegelman,D., Dionne-Laporte,A., Boelman,C. *et al.* (2018) Global characterization of copy number variants in epilepsy patients from whole genome sequencing. *PLoS Genet.*, **14**, e1007285.
30. Yoon,S., Xuan,Z., Makarov,V., Ye,K. and Sebat,J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.
31. Xi,R., Hadjipanayis,A.G., Luquette,L.J., Kim,T.-M., Lee,E., Zhang,J., Johnson,M.D., Muzny,D.M., Wheeler,D.A., Gibbs,R.A. *et al.* (2011) Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, E1128–E1136.
32. Glusman,G., Severson,A., Dhankani,V., Robinson,M., Farrah,T., Mauldin,D.E., Stittrich,A.B., Ament,S.A., Roach,J.C., Brunkow,M.E. *et al.* (2015) Identification of copy number variants in whole-genome data using reference coverage profiles. *Front. Genet.*, **5**, 1–13.
33. Handsaker,R.E., Van Doren,V., Berman,J.R., Genovese,G., Kashin,S., Boettger,L.M. and McCarroll,S.A. (2015) Large multiallelic copy number variations in humans. *Nat. Genet.*, **47**, 296–303.
34. Sudmant,P.H., Rausch,T., Gardner,E.J., Handsaker,R.E., Abyzov,A., Huddleston,J., Zhang,Y., Ye,K., Jun,G., Hsi-Yang Fritz,M. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
35. Francioli,L.C., Menelaou,A., Pulit,S.L., van Dijk,F., Palamara,P.F., Elbers,C.C., Neerincx,P.B.T., Ye,K., Guryev,V., Kloosterman,W.P. *et al.* (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.*, **46**, 818–825.
36. Kloosterman,W.P., Francioli,L.C., Hormozdiari,F., Marschall,T., Hehir-Kwa,J.Y., Abdellaoui,A., Lameijer,E.-w., Moed,M.H., Koval,V., Renkens,I. *et al.* (2015) Characteristics of de novo structural changes in the human genome. *Genome Res.*, **25**, 792–801.
37. Hormozdiari,F., Hajirasouliha,I., Dao,P., Hach,F., Yorukoglu,D., Alkan,C., Eichler,E.E. and Sahinalp,S.C. (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**, i350–i357.
38. He,D., Hormozdiari,F., Furlotte,N. and Eskin,E. (2011) Efficient algorithms for tandem copy number variation reconstruction in repeat-rich regions. *Bioinformatics*, **27**, 1513–1520.
39. MacDonald,M.E., Ambrose,C.M., Duyao,M.P., Myers,R.H., Lin,C., Srinidhi,L., Barnes,G., Taylor,S.A., James,M., Groot,N. *et al.* (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**, 971–983.
40. Mirkin,S.M. (2007) Expandable DNA repeats and human disease. *Nature*, **447**, 932–940.
41. Rich,J., Ogryzko,V.V. and Pirozhkova,I.V. (2014) Satellite DNA and related diseases. *Biopolymers Cell*, **30**, 249–259.
42. Carvalho,C. M.B. and Lupski,J.R. (2016) Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.*, **17**, 224–238.
43. Gymrek,M., Golan,D., Rosset,S. and Erlich,Y. (2012) lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.*, **22**, 1154–1162.
44. Warburton,P.E., Hasson,D., Guillem,F., Lescale,C., Jin,X. and Abrusan,G. (2008) Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics*, **9**, 533.
45. Korbel,J.O., Urban,A.E., Affourtit,J.P., Godwin,B., Grubert,F., Simons,J.F., Kim,P.M., Palejev,D., Carriero,N.J., Du,L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
46. Kazazian,H.H. and Moran,J.V. (2017) Mobile DNA in health and disease. *N. Engl. J. Med.*, **377**, 361–370.
47. Hannan,A.J. (2018) Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.*, **19**, 286–298.
48. Boivin,M., Brendgen,M., Dionne,G., Dubois,L., Pérusse,D., Robaey,P., Tremblay,R.E. and Vitaro,F. (2013) The Quebec newborn twin study into adolescence: 15 years later. *Twin Res. Hum. Genet.*, **16**, 64–69.
49. Scelo,G., Riazalhosseini,Y., Greger,L., Letourneau,L., Gonzàlez-Porta,M., Wozniak,M.B., Bourgey,M., Harnden,P., Egevad,L., Jackson,S.M. *et al.* (2014) Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat. Commun.*, **5**, 5135.
50. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
51. Seshan,V. and Olshen,A. (2017) DNAcopy: DNA copy number data analysis. *R package version 1.50.1.*
52. Faust,G.G. and Hall,I.M. (2012) YAHA: fast and flexible long-read alignment with optimal breakpoint detection. *Bioinformatics*, **28**, 2417–2424.

53. Layer,R.M., Chiang,C., Quinlan,A.R. and Hall,I.M. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.
54. Pendleton,M., Sebra,R., Pang,A. W.C., Ummat,A., Franzen,O., Rausch,T., Stütz,A.M., Stedman,W., Anantharaman,T., Hastie,A. *et al.* (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods*, **12**, 780–786.
55. Mostovoy,Y., Levy-Sakin,M., Lam,J., Lam,E.T., Hastie,A.R., Marks,P., Lee,J., Chu,C., Lin,C., Džakula,Ž. *et al.* (2016) A hybrid approach for de novo human genome sequence assembly and phasing. *Nat. Methods*, **13**, 587–590.
56. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
57. Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C. and Salzberg,S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
58. Chiang,C., Scott,A.J., Davis,J.R., Tsang,E.K., Li,X., Kim,Y., Hadzic,T., Damani,F.N., Ganel,L., GTEx,Consortium *et al.* (2017) The impact of structural variation on human gene expression. *Nat. Genet.*, **49**, 692–699.
59. Chaisson,M.J.P., Huddleston,J., Dennis,M.Y., Sudmant,P.H., Malig,M., Hormozdiari,F., Antonacci,F., Surti,U., Sandstrom,R., Boitano,M. *et al.* (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, **517**, 608–611.
60. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
61. Nguyen,D.-Q., Webber,C. and Ponting,C.P. (2006) Bias of selection on human copy-number variants. *PLoS Genet.*, **2**, e20.
62. Sudmant,P.H., Mallick,S., Nelson,B.J., Hormozdiari,F., Krumm,N., Huddleston,J., Coe,B.P., Baker,C., Nordenfelt,S., Bamshad,M. *et al.* (2015) Global diversity, population stratification, and selection of human copy-number variation. *Science*, **349**, aab3761.
63. Eckert,K.A. and Hile,S.E. (2009) Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Mol. Carcinogen.*, **48**, 379–388.
64. Willems,T.F., Gymrek,M., Highnam,G., Mittelman,D. and Erlich,Y. (2014) The landscape of human STR variation. *Genome Res.*, 1894–1904.
65. Fungtammasan,A., Ananda,G., Hile,S.E., Su,M. S.-w., Sun,C., Harris,R., Medvedev,P., Eckert,K. and Makova,K.D. (2015) Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Res.*, **25**, 736–749.
66. Kelley,D. and Rinn,J. (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.*, **13**, R107.
67. Lu,X., Sachs,F., Ramsay,L., Jacques,P.-É., Göke,J., Bourque,G. and Ng,H.-H. (2014) The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat. Struct. Mol. Biol.*, **21**, 423–425.
68. Bailey,J.A., Liu,G. and Eichler,E.E. (2003) An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.*, **73**, 823–834.
69. Kidd,J.M., Graves,T., Newman,T.L., Fulton,R., Hayden,H.S., Malig,M., Kallicki,J., Kaul,R., Wilson,R.K. and Eichler,E.E. (2010) A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, **143**, 837–847.
70. Lambert,J.-C., Heath,S., Even,G., Campion,D., Sleegers,K., Hiltunen,M., Combarros,O., Zelenika,D., Bullido,M.J., Tavernier,B. *et al.* (2009) Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.*, **41**, 1094–1099.