Supplementary Material for “Human copy number variants are enriched in regions of low mappability”

Jean Monlong¹,², Patrick Cossette³, Caroline Meloche³, Guy Rouleau⁴, Simon L. Girard¹,³,⁵, and Guillaume Bourque¹,²,⁶,⁺

¹Department of Human Genetics, McGill University, Montréal, H3A 1B1, Canada
²Canadian Center for Computational Genomics, Montréal, H3A 1A4, Canada
³Centre de Recherche du Centre Hospitalier de l’Université de Montréal, Montréal, H2X 0A9, Canada.
⁴Montréal Neurological Institute, McGill University, Montréal, H3A 2B4, Canada.
⁵Département des sciences fondamentales, Université du Québec à Chicoutimi, Chicoutimi, G7H 2B1, Canada
⁶McGill University and Génome Québec Innovation Center, Montréal, H3A 1A4, Canada
⁺Correspondence: guil.bourque@mcgill.ca

May 9, 2018
## Contents

**1 Supplementary Tables**

**2 Supplementary Figures**

**3 Supplementary Information**

| Section                                                                 | Page |
|------------------------------------------------------------------------|------|
| 3.1 Data                                                               | 22   |
| 3.2 Read count across the genome                                       | 23   |
| 3.3 RD and mappability estimates                                       | 24   |
| 3.4 CNV detection with PopSV                                            | 25   |
| 3.5 Validation and benchmark                                            | 26   |
| 3.6 Experimental validation                                            | 28   |
| 3.7 Analysis of CEPH12878                                              | 29   |
| 3.8 Genomic patterns of CNVs                                            | 31   |

**Bibliography**

35
List of Tables

| Table | Description                                                                 | Page |
|-------|-----------------------------------------------------------------------------|------|
| S1    | Experimental validation results.                                            | 5    |
| S2    | Experimental validation in low-coverage regions.                            | 6    |
| S3    | Investigating low-mappability deletion calls with two CEPH12878 assemblies.| 7    |
| S4    | Properties of events in public CNV catalogs                                 | 8    |
| S5    | OMIM genes overlapping novel CNV regions of low-mappability                 | 9    |
List of Figures

S1 Coverage, mappability and population-based measures .................................. 11
S2 Average coverage in reference samples in the CageKid and GoNL datasets. .......... 12
S3 Rand index between pedigree and CNV-based dendogram in low-coverage regions... 13
S4 PopSV’s performance in low-mappability regions in CageKid dataset. ................ 14
S5 Distance to assembly gaps and supporting evidence from long-read sequencing in
   CEPH12878. ........................................................................................................ 15
S6 Overlap between PopSV catalog and calls from Pendleton et al. .......................... 16
S7 Distance to a centromere, telomere or assembly gap. ..................................... 17
S8 CNVs enrichment after controlling for segmental duplication overlap and distance to
   CTG. .................................................................................................................... 18
S9 Overlap between CNVs and repeats. ............................................................... 19
S10 Polymorphism likely caused by non-homologous allelic recombination between L1PA
    repeats. ............................................................................................................ 20
S11 Novel CNV regions and CNVs in other public catalogs ................................. 21
### 1 Supplementary Tables

| Validated | Chr. | Start       | End         | Class   | Left PCR primer                                      | Right PCR primer                                      |
|-----------|------|-------------|-------------|---------|------------------------------------------------------|-------------------------------------------------------|
| V         | 3    | 6649794     | 6654897     | large CN 0 | CCTTAGTATTTCACTCGTACATTTGCTAGGTTAT              | ATAAATATCGTGCCTAACTTTGAGCTCTT                       |
| V         | 5    | 127407030   | 127411341   | large CN 0 | TATTCATATTAACCTACTCTCCAGAAGA                      | TTTTTAGAGATTGGACTAAATAATTCCAC                       |
| V         | 3    | 5535139     | 5539535     | large CN 0 | TACTTTTTGAGATTTGAATTTTCTTTGCTAG               | GAATACTGAAATCAAAGATCATATCTGAG                       |
| V         | 1    | 116229111   | 11623162    | large CN 0 | GTGTACGAAATTTTATCTGAGGTGTC                      | ATCTATAGAACATTTTTTCAAAATACCA                       |
| V         | 1    | 158960188   | 15896058    | large CN 1 | GTGAGTACCAGCTGTGACGTGATAG                          | ATGACTTTCTATTGTTGAAATGTAGTAC                       |
| V         | 15   | 26748887    | 26752614    | large CN 1 | CAATTATCATCAACATTACCGCTACGAGTCTGAGTAG            | AGTGGAGATTCTTATTTAAGCTTTGCTCTG                      |
| V         | 6    | 33937344    | 33942846    | large CN 1 | ACATTGTAGCCTGATGACCTGTGTCAGG                      | TCACTTAAAACTATCAGTTCAGTTAAAAAGG                     |
| V         | 12   | 82095501    | 82099899    | large CN 1 | ACCTATAACTAAGTACTGTCGAGTACGAGG                   | TCAGTAATAAACTATCAGTTCAGTTAAAAAGG                     |
| V         | 5    | 8255604     | 8260914     | large CN 1 | TCAACATCATACATACATACGACTAACTAGAGTCTG             | TACACATCTGAAACACTTTTCAAAATACCA                       |
| V         | 20   | 7398397     | 7403743     | large CN 1 | AACCAACTGTCTGAAACCTAAACCATTTTTGAGAGT            | CTTTTGTCACATTTCTAATAACAGCTTTGCTCTG                 |
| V         | 18   | 40535859    | 40535873    | large CN 1 | TAATTTCTTTTTTCTAAAGCTTTTTTTGAGTAG              | GTGACTTTTAACTACTAGTTAAAAAGG                         |
| V         | 16   | 48904951    | 48906510    | small CN 0 | TCAATTTTTTGTGAGTCTGCTTGAAATACGAGG              | AGATAGATACATTTTTGAAAGCTTTTGCTGTCTT                  |
| V         | 2    | 24108664    | 24108801    | small CN 0 | TCAACATCATACATACATACGACTAACTAGAGTCTG             | TCAGTAATAAACTATCAGTTCAGTTAAAAAGG                     |
| V         | 13   | 110224121   | 11022431    | small CN 0 | ACCTACGGAGAACAATCTTTGACATTGAGG                  | GTAGAAACACACTTCACTGGAGATACCTTT                       |
| V         | 11   | 60571017    | 60572170    | small CN 0 | AATGTGGAAGTGCTGTTCTCCTGTAATAACTGCT             | GTGGTTTTGTTGCTATTCTTTGAGATTTA                       |
| V         | 5    | 166402295   | 16640419    | small CN 0 | TAATTTTTTATCATGACTCTACATCTTCTGACTGAGG           | GATGATAGCTATTAAAAAGCTTAATTGAG                       |
| N         | 3    | 16012828    | 16012828    | small CN 1 | TAGATAACAGAAAAATAGATAAAACACTGAGG                | TCAGAAACACTTTTTTTAAGAGAAGAGG                       |
| N         | 17   | 10612674    | 10612674    | small CN 1 | AATGTGGAAGTGCTGTTCTCCTGTAATAACTGCT             | GTGGTTTTGTTGCTATTCTTTGAGATTTA                       |
| V         | 10   | 70251515    | 70251515    | small CN 1 | AATATGAAATACCAGGGAAATGAAAGAAGAAGGAGG            | ATATATCGATTTTTCTATTTTGAGACAGAGAGG                   |
| V         | 8    | 53700365    | 53700365    | small CN 1 | TAAGGGAAATATCTATAGCTCTGAGGACCTGG                | ATGGAATATATATATCTGATAGGTGGAC                       |

**Table S1: Experimental validation results.** Location of the validated (V) and non-validated (N) CNVs for different classes. The last two columns show the primer sequences used for PCR amplification.
## Table S2: Experimental validation in low-coverage regions.

The result of the PCR validation was either concordant with PopSV call (Yes), discordant (No) or inconclusive (NA). In some cases, Sanger sequencing was performed. The CN column is the estimated copy-number of the deleted allele.
| Deletion Support | Reference Support | Number of Calls |
|------------------|-------------------|----------------|
| 0                | 0                 | 11             |
| 0                | 1                 | 1              |
| 1                | 0                 | 1              |
| 2                | 0                 | 12             |

| Deletion Support | Reference Support | Number of Calls |
|------------------|-------------------|----------------|
| 0                | 0                 | 18             |
| 0                | 1                 | 10             |
| 0                | 2                 | 7              |
| 1                | 0                 | 6              |
| 1                | 1                 | 4              |
| 1                | 2                 | 4              |
| 2                | 0                 | 10             |
| 2                | 1                 | 3              |

Table S3: Investigating low-mappability deletion calls with two CEPH12878 assemblies.

The first two columns represent the number of assemblies (0, 1 or 2) supporting the deleted allele or the reference allele. The third column shows the number of PopSV calls in each category.
Table S4: **Properties of events in public CNV catalogs.** Deletions, duplications and CNVs from four public catalogs. Variants with high frequency (> 80%), variants on the chromosome X, and variants smaller than 300 bp were removed in order to compare with PopSV’s numbers (Table 1). 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.

| CNV catalog                  | Samples | Variants | Avg Size (Kbp) | Proportion | Affected genome (Mbp) |
|------------------------------|---------|----------|----------------|------------|-----------------------|
|                              |         | Total    | Per sample     | <3 Kbp     | Total                 | Per sample |
|                              |         | WG       | ELC            |            |                       |            |
| 1000GP                       | 2,504   | 41,979   | 1,024.44      | 2.22       | 6.00                  | 0.68       | 580.03 | 6.14 |
| deletion                     |         | 36,102   | 975.32        | 2.21       | 4.67                  | 0.72       | 342.97 | 4.56 |
| duplication                   |         | 8,503    | 48.26         | 0.00       | 32.54                 | 0.00       | 331.48 | 1.57 |
| GoNL                         | 750     | 9,592    | 1,048.14      | 0.63       | 2.93                  | 0.81       | 65.30  | 3.07 |
| deletion                     |         | 9,009    | 1,013.35      | 0.63       | 2.36                  | 0.82       | 34.79  | 2.39 |
| duplication                   |         | 528      | 21.11         | 0.00       | 29.19                 | 0.15       | 30.63  | 0.62 |
| Handsaker 2015 (Genome STRiP)| 847     | 8,657    | 212.03        | 1.88       | 27.80                 | 0.00       | 196.57 | 5.89 |
| deletion                     |         | 5,961    | 145.78        | 0.56       | 21.64                 | 0.00       | 108.03 | 3.15 |
| duplication                   |         | 3,469    | 66.26         | 1.32       | 41.35                 | 0.00       | 118.28 | 2.74 |
| Chiang 2017 (Genome STRiP)   | 148     | 7,932    | 828.49        | 9.23       | 6.35                  | 0.42       | 73.20  | 5.26 |
| Novel region          | OMIM gene |
|-----------------------|-----------|
| 1:25730001-25736000   | RHCE      |
| 1:16164001-161645000  | FCGR2B    |
| 1:207705001-207715000 | CR1       |
| 1:207725001-207745000 | CR1       |
| 5:68845001-68886000   | OCLN      |
| 5:69360001-69365000   | SMN2      |
| 5:69373001-69374000   | SMN2      |
| 5:70165001-70222000   | SMN1      |
| 5:70242001-70242500   | SMN1      |
| 5:70246501-70258000   | SMN1      |
| 6:29905001-29910000   | HLA-A     |
| 6:31960001-31975000   | C4A       |
| 6:31995001-31995500   | C4B       |
| 6:32522001-32560000   | HLA-DRB1  |
| 6:32590001-32602000   | HLA-DQA1  |
| 6:32628501-32629000   | HLA-DQB1  |
| 6:32630001-32634000   | HLA-DQB1  |
| 7:39052001-39055500   | POU6F2    |
| 7:74195001-74200000   | NCF1      |

| Novel region          | OMIM gene |
|-----------------------|-----------|
| 9:136418001-136420000 | ADAMTSL2  |
| 10:64130001-64135000  | ZNF365    |
| 10:101595001-101600000| ABCC2     |
| 10:135380001-135383500| SYCE1     |
| 11:320001-325000      | IFITM3    |
| 12:52683001-52685000  | KRT81     |
| 12:52696001-52696500  | KRT86     |
| 15:32454001-32460000  | CHRNA7    |
| 15:32464001-32464500  | CHRNA7    |
| 15:43902501-43903000  | STRC      |
| 15:43910001-43910500  | STRC      |
| 16:21760001-21765000  | OTOA      |
| 17:34504001-34545000  | CCL3L3    |
| 19:11535001-11540000  | CCDC151   |
| 19:41340001-41350000  | CYP2A6    |
| 22:18660001-18765000  | USP18     |
| 22:18904501-18905500  | PRODH     |
| 22:18909001-18909500  | PRODH     |

Table S5: **OMIM genes overlapping novel CNV regions of low-mappability** Novel CNV regions are polymorphic in more than 1% of the individuals across the three cohorts but absent from the 1000GP SV catalog\(^1\). OMIM genes are genes associated with a disease or phenotype in the OMIM Morbid Map (Online Mendelian Inheritance in Man; [http://omim.org/](http://omim.org/)).
2 Supplementary Figures
Figure S1: **Coverage, mappability and population-based measures.** a-b) Read coverage in a sample (y-axis) versus mappability (a) or the inter-sample average coverage (b). c-d) Inter-sample mean (c) and standard deviation (d) were fitted against the mappability in each cohort separately. The tiles represent all cohorts pooled together.
Figure S2: Average coverage in reference samples in the CageKid (a) and GoNL (b) datasets.
Figure S3: **Rand index between the pedigree information and the dendogram from CNV calls in low-coverage regions.** The dendogram for CNV-based clustering was cut at different levels (x-axis) and the groups compared to the pedigree (family-level) with the Rand index (y-axis). For each method, the line highlights the best performance across three linkage criteria.
Figure S4: PopSV’s performance in low-mappability regions in CageKid dataset. Proportion and number of calls replicated in the paired tumor. The point shows the median value per sample, the error bars the 95% confidence interval.
Figure S5: **Distance to assembly gaps and supporting evidence from long-read sequencing in CEPH12878.** Deletions in low-mappability regions were grouped by their supporting evidence (y-axis and colors). *assemblies:* deletion observed in at least one of the two public assemblies. *PB-SV:* overlap with a structural variant called from the PacBio reads\(^2\). *PB-reads:* deletion observed in the local assembly or consensus of the PacBio reads. Variants with no support are represented by the white boxplot.
Figure S6: **Overlap between PopSV catalog and calls from Pendleton et al.**. Recurrent calls were collapsed in each catalog (i.e. PopSV and the 1000 Genomes Project (1000GP)). The proportion of the collapsed calls overlapping calls from Pendleton et al.\(^2\) was computed. The fold-enrichment is produced by drawing control regions with similar size distribution as Pendleton’s calls. *low-map*: calls in low-mappability regions; *ext. low-map*: calls in extremely low-mappability regions.
Figure S7: **Distance to a centromere, telomere or assembly gap.** The y-axis represents the cumulative proportion of the affected genome. The *expected* curve is computed from uniformly distributed genomic regions with matched size.
Figure S8: CNVs enrichment after controlling for segmental duplication overlap and distance to CTG. Enrichment of CNVs in a) different genomic features, b) satellite families and c) simple repeats in the different cohorts (colors). Bars show the median fold enrichment across samples compared to control regions. The star represents significant enrichment from the logistic regression.
Figure S9: **Overlap between CNVs and repeats.** The histograms represent the proportion of the CNV region that overlaps a) a satellite, b) a simple repeat or c) a transposable element, when they do overlap. The *expected* distribution is computed from the control regions used for the enrichment analysis.
Figure S10: **Polymorphism likely caused by non-homologous allelic recombination between L1PA repeats.** Examples of CNV likely caused by non-allelic homologous recombination between two L1PA3 repeats (a) or L1PA6 (b). The line and points represent the coverage of one sample with a duplication (a) or a deletion (b), highlighted in yellow; the violin plots represent the distribution of the coverage in the reference samples.
Figure S11: **Novel CNV regions and CNVs in other public catalogs.** Cumulative proportion of the 3,455 novel regions (y-axis) that overlap CNVs in different public CNV catalogs (colour) depending on their frequency in the public catalog (x-axis). The labels highlight the proportion of novel regions that don’t overlap any CNV in the corresponding public catalog. Novel regions were defined as overlapping a CNV in more than 1% of the individuals but absent from the 1000GP catalog.
3 Supplementary Information

3.1 Data

Twin study All patients gave informed consent in written form to participate in the Quebec Study of Newborn Twins. Ethic boards from the Centre de Recherche du CHUM, from the Université Laval and from the Montreal Neurological Institute approved this study. Sequencing was done on an Illumina HiSeq 2500 (paired-end mode, fragment length 300 bp). The reads were aligned using a modified version of the Burrows-Wheeler Aligner (bwa version 0.6.2-r126-tpx with threading enabled). The options were ‘bwa aln -t 12 -q 5’ and ‘bwa sampe -t 12’. The aligned reads are available on the European Nucleotide Archive under ENA PRJEB8308. The 45 samples had an average sequencing depth of 40x (minimum 34x / maximum 57x).

Renal cell carcinoma WGS data from renal cell carcinoma is presented in details in the CageKid paper. In short, 95 pairs of normal/tumor tissues were sequenced using GAIIx and HiSeq2000 instruments. Paired-end reads of size 100 bp totaled an average sequencing depth of 54x (minimum 26x / maximum 164x). Reads were trimmed with FASTX-Toolkit and mapped per lane with BWA backtrack to the GRCh37 reference genome. Picard was used to adjust pairs coordinates, flag duplicates and merged lane. Finally, realignment was done with GATK. Raw sequence data have been deposited in the European Genome-phenome Archive, under the accession code EGAS00001000083.

Genome of the Netherlands WGS data from the GoNL project is described in details in Francioli et al. This data have been derived from different sample collections:

• The LifeLines Cohort Study (http://www.lifelines.nl/), supported by the Netherlands Organization of Scientific Research (NWO, grant 175.010.2007.006), the Dutch government’s Economic Structure Enhancing Fund (FES), the Ministry of Economic Affairs, the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the Northern Netherlands Collaboration of Provinces (SNN), the Province of Groningen, the University Medical Center Groningen, the University of Groningen, the Dutch Kidney Foundation and Dutch Diabetes Research Foundation.
• The EMC Ergo Study (http://www.ergo-onderzoek.nl/wp/).

• The LUMC Longevity Study, supported by the Innovation-Oriented Research Program on Genomics (SenterNovem IGE01014 and IGE05007), the Centre for Medical Systems Biology and the National Institute for Healthy Ageing (Grant 05040202 and 05060810).

• VU Netherlands Twin Register (http://www.tweelingenregister.org/).

In short, samples were sequenced on an Illumina HiSeq 2000 instrument (91-bp paired-end reads, 500-bp insert size). We downloaded the aligned read sequences (BAM) for the 500 parents in the data set. We further performed indel realignment using GATK 3.2.2, adjusted pairs coordinates with Samtools 0.1.19, marked duplicates with Picard 1.118, and performed base recalibration (GATK 3.2.2). The average sequencing depth was 14x (minimum 9x / maximum 59x).

Genomic annotations  Gencode annotation (V19) was directly downloaded from the consortium FTP server at ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz. Other genomic annotations were downloaded from the UCSC database server at http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database. The file names of the corresponding annotations are

| Annotation                          | File Name                                      |
|-------------------------------------|------------------------------------------------|
| Mappability                         | wgEncodeCrgMapabilityAlign100mer.bw            |
| Cytogenetic bands                   | cytoBandIdeo.txt.gz                            |
| Centromere, telomere, assembly gap  | gap.txt.gz                                     |
| Segmental duplication               | genomicSuperDups.txt.gz                        |
| Simple repeat / Short Tandem Repeats| simpleRepeat.txt.gz                            |
| RepeatMasker                        | rmsk.txt.gz                                    |

3.2 Read count across the genome

The genome was fragmented in non-overlapping bins of fixed size. The number of properly mapped reads was used as a coverage measure, 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 coverage and the bin’s GC content. For each bin, the
correction factor was computed as the mean coverage across all the bins divided by the predicted
coverage from the LOESS model and the GC content of the bin. We used a bin size of 5 Kbp for
most of the analysis. When specified, we used a smaller bin size of 500 bp.

### 3.3 RD and mappability estimates

To investigate the bias in RD we used the read counts in 5 Kbp bins. Bins with extremely high
coverage were identified and removed when deviating from the median coverage by more than 5
standard deviation. First the coverage of the 45 samples from the Twin study were combined and
quantile normalized. At that point the different samples had the same global coverage distribution
and no bins with extreme coverage or GC bias.

The mappability track was downloaded from UCSC\(^6\) \((\text{wgEncodeCrgMapabilityAlign100mer.bw})\)
and the average mappability was computed for each bin. One sample was randomly selected and we
compared its coverage with the mappability estimates. We then computed the mean and standard
deviation of the coverage in each bin across the other samples and compared it with the sample
coverage. We also compared the inter-sample average with the mappability estimates.

To compute Z-scores that integrates the observed coverage variation we used two approaches.
The first modeled the coverage metrics (average or standard deviation) using the mappability es-
timates and computed a Z-score from the predicted coverage and global standard deviation. A
generalized additive model was fitted using a cubic regression spline on the mappability estimates
\((\text{mgcv R package})\). In the second approach, Z-scores were 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. For the kurtosis and skewness computation, we removed outlier
Z-scores with an absolute value greater than 10. These bins could be regions of CNV and would bias
the estimates. The Z-score distributions were also compared in bins from 10 different mappability
intervals.

We repeated this analysis pooling 45 samples from each of the three datasets. After quantile
normalization, the inter-sample coverage mean and standard deviation were computed separately
in each cohort and compared with the mappability estimates.
3.4 CNV detection with PopSV

Binning the genome  We ran two separate analysis on the three datasets. Bin sizes of 5 Kbp and 500 bp were used on the Twin study and renal cell carcinoma. Because of its lower sequencing depth, the 500 bp run on GoNL gave only partial results. More precisely, we observed a truncated distribution of the copy-number estimates, with most of the 1 and 3 copy number variants missing. It means that at this resolution many one-copy variation cannot be differentiated from background noise. For this reason we ran GoNL analysis using 2 Kbp and 5 Kbp bins.

Constructing the set of reference samples  In each dataset we choose the reference samples as follows: in the renal cancer dataset from the normal samples, in the Twin study from all the samples, in GoNL from a subset of 200 samples (see below). For each dataset, a Principal Component Analysis (PCA) was performed across samples on the counts normalized globally (median/variance adjusted). The resulting first two principal components are used to verify the homogeneity of the reference samples. Although our three datasets showed different levels of homogeneity, we didn’t need to exclude samples or split the analysis. The effect of weak outlier samples was either corrected by the normalization step or integrated in the population-view.

In GoNL, we decided to use only 200 of the 500 samples as reference. They were selected to span a maximum of the space defined by the principal components. In contrast to random selection, this ensures that weak outliers are included in the final set of reference samples, hence maximizing the technical variation integrated in the population-view.

Moreover, the principal components were used to select one control sample from the final set of reference samples. This sample is used in the normalization step as a baseline to normalize other samples against. We picked the sample closest to the centroid of the reference samples in the Principal Component space.

CNV calling  After targeted normalization the coverage in each sample is compared to the coverage in the reference samples. A Z-score is computed and translated into a P-value that is then corrected for multiple testing. Consecutive bins with significant excess or lack of reads are merged and returned as potential duplication or deletion. Copy number estimates are derived from the cov-
verage across the bin and the average coverage across the reference samples. However, it is important to note that the definition of a variant is different from other methods. Here a variant is defined by the major allele in the population rather than the reference genome state. Most of the genome is in a diploid state compared to the reference genome and sufficiently covered by sequencing reads that the copy number state can be correctly estimated by PopSV’s population-based approach. However, highly polymorphic variants are called relative to the major allele in the population and additional efforts are required to assess the copy number state. Variants in extremely low-mappability regions are also difficult to fully characterize and might be caused by rare insertion in the reference genome or complex alleles. Nonetheless, PopSV can efficiently detect the presence of CNV in any situation. 

More details are available in the method paper\textsuperscript{8}.

**Coverage tracks** For each run, we constructed coverage tracks based on the average coverage in the reference samples. Bins where the reference samples had, on average, the expected coverage were classified as *expected coverage*. Bins with a coverage lower than 4 standard deviation from the median were classified as *low-mappability* (or *low coverage*). To ensure robustness, the standard deviation was derived from the Median Absolute Deviation. We use regions with low coverage to define *low-mappability regions*, as the low coverage is a result of the lower mappability of a region. Because the standard deviation is used, the number of regions classified as *low-mappability* is lower in datasets with more RD variance.

Eventually, we also defined *extremely low coverage* region which have an average coverage below 100. This sub-class of *low coverage* region was used in a few analyses to highlight the most challenging regions.

Regions were annotated with the overlap with protein-coding genes and segmental duplications (see Genomic annotations), and the distance to the nearest centromere, telomere or assembly gap. Finally, we computed the number of protein-coding genes overlapping at least one low-coverage region.

### 3.5 Validation and benchmark

**Running** FREEC, CNVnator, cn.MOPS and LUMPY  

FREEC\textsuperscript{9} segments the RD values of a sample
using a LASSO-based algorithm. It was run on each sample separately, starting from the BAM file, using the same bin sizes as for PopSV. FREEC internally corrects the RD for GC and mappability bias. In order to compare its performance in low-mappability region, the minimum “telocentromeric” distance was set to 0. The remaining parameters were set to default. Of note an additional run with slightly looser parameter (breakPointThreshold=0.6) was performed to get a larger set of calls used in some parts of the in silico validation analysis to deal with borderline significant calls.

CNVnator\textsuperscript{10} uses a mean-shift technique inspired from image processing. It was run on each sample separately, starting from the BAM file, using the same bin sizes as for PopSV. CNVnator also corrects internally for GC bias and we used default parameters. For the analysis using higher confidence calls, we used calls with either ‘eval1’ or ‘eval2’ lower than $10^{-5}$ (instead of the default 0.05).

cn.MOPS\textsuperscript{11} considers simultaneously several samples and detects copy number variation using a Poisson model and a Bayesian approach. It was run on the same GC-corrected bin counts used for PopSV. All the samples are analyzed jointly. Of note an additional run with slightly looser parameter (upperThreshold=0.32 and lowerThreshold=-0.42) was performed to get a larger set of calls used in some parts of the in silico validation analysis to deal with borderline significant calls.

LUMPY\textsuperscript{12} which uses an orthogonal mapping signal: the insert size, orientation and split mapping of paired reads. The discordant reads were extracted from the BAMs using the recommended commands. Split-reads were obtained by running YAHA\textsuperscript{13} with default parameters. All the CNVs (deletions and duplications) larger than 300 bp were kept for the upcoming analysis. 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. Calls with 5 or more supporting reads were considered high-confidence.

**Clustering samples from the Twin study** A distance between two samples A and B was defined as: \[ 1 - 2 \frac{|R_A \cap R_B|}{|R_A| + |R_B|} \] where $R_A$ represents the regions called in sample A, $R_A \cap R_B$ the regions called in both A and B, and $|R|$ the cumulative size of the regions. Hence, the similarity between two samples is represented by the amount of sequence found in both divided by the average amount of sequence called. This distance is used for hierarchical clustering of the samples in the
Twin dataset. The clustering was performed using only calls in regions with extremely low coverage (reference average ≤100 reads). Different linkage criteria (average, complete and Ward) were used for the exploration. In our dendograms we used the average linkage criterion. The concordance between the clustering and the pedigree was estimated by the Rand index, grouping the samples per family. For each method and linkage criteria, the Rand index was computed for every possible dendogram cut (x-axis in Figure S3).

3.6 Experimental validation

Experimental validation was performed on samples from the Twin study. In a first validation batch, variants were randomly selected among both one-copy and two-copy deletions. We selected both small (∼700 bp) and large (∼4 Kbp) variants in each class. The coverage at base pair resolution was visually inspected for each deletion and, when possible, the breakpoints were fine-tuned. PCR primers were designed to target the whole deleted region. We randomly selected 20 variants out of the variants for which we managed to design PCR primers. We then performed long-range PCR followed by gel electrophoresis. PCR was performed using 50 ng of DNA and the Phusion High-Fidelity DNA Polymerase from Thermo Fisher Scientific: 95 °C 5 minutes followed by 35 cycles (95 °C 30 seconds, 64 °C 30 seconds, 72 °C 45 seconds) and 72 °C 10 minutes. Either a 1% or 1.8% agarose gel was used, depending on the expected size of the amplified fragments. We used a 1 Kb Plus DNA Ladder from Thermo Fisher Scientific.

The presence of a deletion was tested by comparing the size of the amplified fragment in affected and control samples. If the affected sample showed a lower band than a control with a predicted 2 copies, the deletion was considered validated. On the other hand if affected sample and controls had one similar band, the deletion was considered non-validated. Of note, the validation rate might be under-estimated because visual prediction of the breakpoint is not always accurate.

We then randomly selected deletions overlapping low-mappability regions and detected in 6 samples or fewer. We chose to test rare variants because they are likely enriched in false-positives. Hence, this batch of validation represents the most challenging regions to call and validate, and enriched in false-positives. Here we couldn’t use the base-pair coverage to fine-tune the breakpoints.
because the low-mappability blurs any clear signal. Instead, we retrieved the reads (and their pairs) mapping to the region and assembled them. With this approach we could sometimes get a better breakpoint resolution and design PCR primers that would amplify the deleted region. In addition to gel electrophoresis, the amplified DNA of some regions was sequenced using Sanger sequencing. We randomly selected 17 variants out of the variants for which we managed to design PCR primers.

### 3.7 Analysis of CEPH12878

**Whole-Genome Sequencing data** High coverage PCR-free Illumina WGS data for 30 samples, including CEPH12878, was downloaded from the 1000 Genomes Project. The ENA accession number is PRJNA260854. The files are also available on the FTP server at [http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/high_coverage_alignments/20141118_high_coverage.alignment.index](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/high_coverage_alignments/20141118_high_coverage.alignment.index). Although the sequencing depth is similar to the other datasets (average \( \sim 53X \)), the reads are 250 bp long so the average number of reads per region is lower. Because of the lower read coverage and sample size the CNV calls will be of slightly lower quality. Nonetheless, PopSV was run using 5 Kbp bins and all the samples as reference. Using the same coverage track as before we then selected all deletions in CEPH12878 and overlapping low-mappability regions (at least 90% of the call). We then looked for support in public assemblies, SV catalogs and reads from long-read sequencing technologies.

**Comparison with assemblies** We downloaded the genome assembly produced from short reads, Pacbio and BioNano reads from [ftp://ftp.ncbi.nlm.nih.gov genomes/all/GCA/001/013/985/GCA_001013985.1_ASM101398v1/GCA_001013985.1_ASM101398v1_genomic.fna.gz](ftp://ftp.ncbi.nlm.nih.gov genomes/all/GCA/001/013/985/GCA_001013985.1_ASM101398v1/GCA_001013985.1_ASM101398v1_genomic.fna.gz). We also downloaded a second assembly that was used 10X Genomics linked reads instead of the Pacbio reads. It is available at [http://kwoklab.ucsf.edu/resources/nmeth_201604_NA12878_hybrid_assembly.fasta.gz](http://kwoklab.ucsf.edu/resources/nmeth_201604_NA12878_hybrid_assembly.fasta.gz).

For each selected variant, we retrieved the two 50 Kbp flanking sequences in the reference genome and aligned them against the public assemblies with BLAST. The output was parsed to identify regions with two flanks aligning in at least 1 Kbp of a contig. MUMmer plots between the reference sequence and the contigs were visually inspected. The assembly supported PopSV calls...
when a deletion was visible in the expected region (between the flanks). The assembly supported the reference genome sequence when a contig crosses the variant without clear structural variant.

**SV calls from a long-read sequencing study** We downloaded the SV calls from the Pacbio reads and assembled contigs in Pendleton et al.\textsuperscript{2}. The VCF file is publicly available at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/NA12878.sorted.vcf.gz. We overlapped PopSV calls with deletions from this SV catalogs. Because we used 5 Kbp bins for PopSV, at least 1 Kbp of a PopSV calls needed to overlap a deletion from Pendleton et al.\textsuperscript{2} to be considered as sufficient support. Of note, the distribution of the overlap tended to be either null or higher than 1 Kbp supporting this choice.

**Local assembly of Pacbio reads** Corrected Pacbio reads from citetPendleton2015 were downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/corrected_reads_gt4kb.fasta. Each read was split in 200 bp fragments and mapped to the human reference genome (version hg19). From this mapping information we selected full Pacbio reads with at least one 200 bp mapping within a region of interest (with 30 Kbp flanks). For each region, the reads were mapped to the reference sequence with exonerate and we kept reads with partial mapping as they may support a SV. These reads were then assembled using Canu\textsuperscript{17}. A consensus sequence was also derived for reads clustered by alignment breakpoint and the clustalo\textsuperscript{18} software. The assembled contigs and consensus were mapped to the reference genome to identify a potential breakpoint. The two regions flanking the alignment breakpoint and the sequence spanning the breakpoint were mapped to the entire genome. We used the results of this genome-wide mapping to select the best candidates: assembled sequence whose flanks align uniquely to the region of interest and with reduced alignment quality for the “middle” sequence that spanned the breakpoint. Candidate contig/consensus were further visualized with MUMmer plots\textsuperscript{16}. The assembly supported PopSV calls when a deletion was visible in the expected region (between the flanks).
3.8 Genomic patterns of CNVs

Merging calls from two different bin sizes  Small bins gives better resolution for smaller variants. Large bins gives better sensitivity. For this reason we merged the calls from the 500 bp bin and 5 Kbp bin runs. Variant supported by both sets of calls were merged into one. To decide which set to use for the breakpoints and other information (e.g. copy number estimate), the proportion of overlap was used. If call(s) using small bins overlapped more than a third of a call from the large bin run, it was considered fully recovered by the small bin call which was then used to define breakpoints and other information. If not, the large bin run was considered more appropriate to define the final breakpoints and additional information. Calls unique to each run were simply added to the final set of calls. For the Twin dataset and the renal cancer dataset, calls from the 500 bp and 5Kbp runs were merged. For the GoNL dataset, calls from the 2 Kbp and 5Kbp runs were merged.

Computing global estimates of copy number variation  In Table 1, a call in extremely low coverage region is overlapped at more than 90% by the extremely low coverage track. To compute the total number of calls, we collapsed calls with an overlap higher than 50%. The amount of sequence affected in a genome was computed by merging all the variants in the cohort and counting the number of affected bases in this reference genome. After the merging step, each base of the genome either overlapped a merged variant or not. Each affected base was counted only once, even if it overlapped CNVs in several samples or with large copy number differences.

Comparison with public CNV catalogs  The SV catalog from Sudmant et al.\textsuperscript{1} was downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz. The CNV catalog from Handsaker et al.\textsuperscript{19} was downloaded from http://www.broadinstitute.org/%7Ehandsake/mcnv_data/bulk/1000G_phase1_cnv_genotypes_phased_25Jul2014.genotypes.vcf.gz. For the CNV catalog from Chiang et al.\textsuperscript{20}, we downloaded GTEx\_Analysis\_2016-10-24\_WholeGenomeSeq\_147Indiv\_SV\_sites.vcf.gz from the GTEx Data Portal at https://www.gtexportal.org/home/datasets (data release v6). The CNV catalog from Francioli et al.\textsuperscript{5} was downloaded from https://molgenis26.target.rug.nl/downloads/gonl_
public/variants/release6.1/20161013_GoNL_AF_genotyped_SVs.vcf.gz. We retrieved the set of autosomal deletion, duplication and CNVs. When comparing the global estimates of CNV with PopSV, we removed deletions smaller than 300 bp as well as variants with high frequency (> 80%). This remaining SVs represent CNVs that could in theory be detected by PopSV's approach. Using this sub-set, we derived the number of variants, number of variants smaller than 3 Kbp, number of variants in extremely low coverage regions, and amount of genome affected. These number are computed exactly as the one presented in Table 1 for PopSV's results.

**CNV frequency comparison** The frequency at which a region is affected by a CNV is computed using calls from the 620 unrelated samples. The copy-number change is not taken into account in the computation and the frequency is derived for all the nucleotide that overlaps at least one CNV. Using each catalog we computed, for each base in the genome, the proportion of individuals with a CNV. This frequency measure facilitates the comparison of catalogs with different methods and resolution. We represented the distribution as a cumulative proportion distribution in Figure 3a. The graphs read as “how much of the total affected genome is called in at more than X% of the population”. The frequency distribution was computed separately for deletions and duplications (and CNV in the 1000 Genomes Project catalog). Of note, the 1000 Genomes Project was downsampled to 640 random individuals in order to give comparable frequency curves.

**Comparison with CNV catalogs from long-read studies** First, the SV catalog from Chaisson et al.\textsuperscript{21} was downloaded from http://eichlerlab.gs.washington.edu/publications/chm1-structural-variation/. Recurrent calls were collapsed in both PopSV and the 1000 Genomes Project catalogs. PopSV's catalog corresponded to all germline calls in the Twin study, renal cancer dataset and GoNL. The 1000 Genomes Project catalog contained all the deletions, duplications and CNVs, no matter the size or frequency. The analysis was also performed separately on deletions, duplications, low-mappability regions and extremely low-mappability regions. For each comparison, we randomly selected control regions with sizes and overlap with assembly gaps similar to the SVs in Chaisson et al.\textsuperscript{21} (see Selecting control regions). A logistic regression tested the enrichment of CNVs in the Chaisson catalog.
versus the control regions. The regression was performed on 50 different sampling of the control regions for each comparison. The 50 samplings are represented by the boxplot in Figure 3b. We compared the estimates from the logistic regression. They represent the log odds ratio of a CNV overlapping the catalog from Chaisson. The same analysis was performed using the SV catalog from Pendleton et al.\textsuperscript{2} downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/NA12878.sorted.vcf.gz.

**Distance to centromere, telomere and assembly gaps** The centromeres, telomeres and assembly gaps (CTGs) are annotated in the gap track from UCSC\textsuperscript{6}. However, some chromosomes were missing telomere annotations. We defined them as the 10 Kbp region at the ends of chromosomes derived from the cytogenetic bands track.

The distance from each variant to the nearest CTG was computed and represented as a cumulative proportion, i.e. the proportion of variants located at a distance $d$ or closer to a CTG.

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 Selecting control regions). Thanks to this null distribution we were able to see if variants were located closer/further to CTG than expected by chance.

**Selecting control regions** In several analyses we compared the CNVs with control regions. The control regions have the same size distribution as the regions they are derived from (e.g. CNV, annotation). In some analysis we further controlled for the overlap with specific genomic features. For example, we controlled for the overlap with CTGs to avoid selecting control regions in assembly gaps where no CNV or annotation is available. Controlling for the overlap with regions flanking CTGs, we could simply control for the distance to CTGs. We also used this approach to control for the overlap with segmental duplications and investigate patterns independent from this repeat class.

To select control regions, thousands of bases were first randomly chosen in the genome. The distance between each base and the genomic features was then computed. At this point, simulating a region of a specific size and with specific overlap profile can be done by randomly choosing as
One of the bases that fit the profile:

\[
\left\{ b, \forall \text{ feature } f, O_f(d^b_f - \frac{S_r}{2}) < 0 \right\}
\]

with \( O_f \) equals 1 if the original region overlaps with feature \( f \), -1 if not; \( d^b_f \) is the distance between base \( b \) and feature \( f \); and \( S_r \) is the size of the original region.

For each input region, a control region was selected as described and had by construction the exact same size and overlap profile.

**Enrichment in genomic features**  We tested different genomic features, starting with: genes, exons, low-mappability regions, segmental duplications, satellites, simple repeats and transposable elements. The different satellite families, frequent simple repeat motives, transposable element families were also tested. We overlapped each genomic feature with CNVs and control regions. We then computed the fold change in proportion of regions overlapping a feature, in CNV versus control regions. A pseudo count was added when computing this ratio:

\[
\text{Fold enrichment} = \frac{|CNV \cap \text{Feature}| + 1}{N + 1} \frac{|Control \cap \text{Feature}| + 1}{|CNV \cap \text{Feature}| + 1} \frac{|Control \cap \text{Feature}| + 1}{|Control \cap \text{Feature}| + 1} \tag{2}
\]

where \( N \) is the number of CNVs (and control regions).

The fold enrichments were computed separately for each sample using control regions that fitted perfectly the profile of the variants in the sample. To assess the significance of the enrichment, a logistic regression was performed using CNV and control regions. The model to test one feature in one sample was:

\[
\log \left( \frac{P(\text{feature overlap})}{P(\text{no overlap})} \right) = \beta_0 + \beta_{CNV} \cdot CNV \tag{3}
\]

with \( CNV = \begin{cases} 
0 & \text{if control region} \\
1 & \text{if CNV} 
\end{cases} \)

To control for the enrichment in segmental duplication we used control regions with similar overlap profile (see Selecting control regions). We also added a variable representing the overlap with segmental duplication in the model:
\[
\log \left( \frac{P(\text{feature overlap})}{P(\text{no overlap})} \right) = \beta_0 + \beta_{CNV} \cdot CNV + \beta_{SD} \cdot SD
\] (4)

with \(SD = \begin{cases} 
0 & \text{if no SD overlap} \\
1 & \text{if SD overlap}
\end{cases}\)

For each feature and cohort we computed the median P-value. When numerous tests were performed (e.g. satellite families, simple repeat motives, transposable element families or subfamilies), the P-values were first corrected for multiple testing using Benjamini-Hochberg procedure.

Finally, we computed the proportion of the region overlapped by the different features (satellites, simple repeats and transposable elements). We compared CNV regions and control regions.

**Somatic variant definition**  Somatic variants were defined as variant in a tumor samples with low overlap with variant in the paired normal sample. In CageKid data, overlapping tumor variant with the ones from the paired normal showed almost only two peaks, at 0 and 100% overlap. A tumor variant was defined as somatic if it overlapped less than 10% of any variant in the paired normal.

**Bibliography**

[1] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Hsi-Yang Fritz et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, oct 2015. ISSN 0028-0836.

[2] M. Pendleton, R. Sebra, A. W. C. Pang, A. Ummat, O. Franzen, T. Rausch, A. M. Stütz, W. Stedman, T. Anantharaman, A. Hastie et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods*, 12(8):780–6, aug 2015. ISSN 1548-7105.

[3] M. Boivin, M. Brendgen, G. Dionne, L. Dubois, D. Pérousse, P. Robaey, R. E. Tremblay, and F. Vitaro. The Quebec Newborn Twin Study Into Adolescence: 15 Years Later. *Twin Research and Human Genetics*, 16(01):64–69, feb 2013. ISSN 1832-4274.
[4] G. Scelo, Y. Riazalhosseini, L. Greger, L. Letourneau, M. González-Porta, M. B. Wozniak, M. Bourgey, P. Harnden, L. Egevad, S. M. Jackson et al. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. Nature Communications, 5(May):5135, oct 2014. ISSN 2041-1723.

[5] L. C. Francioli, A. Menelaou, S. L. Pulit, F. van Dijk, P. F. Palamara, C. C. Elbers, P. B. T. Neerincx, K. Ye, V. Guryev, W. P. Kloosterman et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nature Genetics, 46(8):818–825, jun 2014. ISSN 1061-4036.

[6] K. R. Rosenbloom, J. Armstrong, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler et al. The UCSC Genome Browser database: 2015 update. Nucleic Acids Research, 43(D1):D670–D681, 2015. ISSN 0305-1048.

[7] T. Derrien, J. Estellé, S. Marco Sola, D. G. Knowles, E. Raineri, R. Guigó, and P. Ribeca. Fast computation and applications of genome mappability. PloS one, 7(1):e30377, jan 2012. ISSN 1932-6203.

[8] J. Monlong, S. L. Girard, C. Meloche, M. Cadieux-Dion, D. M. Andrade, R. G. Lafreniere, M. Gravel, D. Spiegelman, A. Dionne-Laporte, C. Boelman et al. Global characterization of copy number variants in epilepsy patients from whole genome sequencing. PLoS Genetics, 14(4):e1007285, 2018.

[9] V. Boeva, A. Zinovyev, K. Bleakley, J. P. Vert, I. Janoueix-Lerosey, O. Delattre, and E. Barillot. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics, 27(2):268–269, jan 2011. ISSN 13674803.

[10] A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research, 21(6):974–984, jun 2011. ISSN 10889051.

[11] G. Klambauer, K. Schwarzbauer, A. Mayr, D. A. Clevert, A. Mitterecker, U. Bodenhofer, and S. Hochreiter. Cn.MOPS: Mixture of Poissons for discovering copy number variations in next-
generation sequencing data with a low false discovery rate. *Nucleic Acids Research*, 40(9): e69–e69, may 2012. ISSN 03051048.

[12] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):R84, jun 2014. ISSN 1465-6906.

[13] G. G. Faust and I. M. Hall. YAHA: Fast and flexible long-read alignment with optimal breakpoint detection. *Bioinformatics*, 28(19):2417–2424, oct 2012. ISSN 13674803.

[14] Y. Mostovoy, M. Levy-Sakin, J. Lam, E. T. Lam, A. R. Hastie, P. Marks, J. Lee, C. Chu, C. Lin, Ž. Džakula *et al.* A hybrid approach for de novo human genome sequence assembly and phasing. *Nature Methods*, 13(7):587–590, may 2016. ISSN 1548-7091.

[15] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. BLAST+: architecture and applications. *BMC bioinformatics*, 10:421, dec 2009. ISSN 1471-2105.

[16] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12, 2004. ISSN 1474-760X.

[17] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27(5):722–736, may 2017. ISSN 1549-5469.

[18] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7:539, oct 2011. ISSN 1744-4292.

[19] R. E. Handsaker, V. Van Doren, J. R. Berman, G. Genovese, S. Kashin, L. M. Boettger, and S. A. McCarroll. Large multiallelic copy number variations in humans. *Nature Genetics*, 47(3):296–303, jan 2015. ISSN 1061-4036.
[20] C. Chiang, A. J. Scott, J. R. Davis, E. K. Tsang, X. Li, Y. Kim, T. Hadzic, F. N. Damani, L. Ganel, GTEx Consortium et al. The impact of structural variation on human gene expression. *Nature genetics*, 49(5):692–699, may 2017. ISSN 1546-1718.

[21] M. J. P. Chaisson, J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig, F. Hormozdiari, F. Antonacci, U. Surti, R. Sandstrom, M. Boitano et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608–11, jan 2015. ISSN 1476-4687.