Human copy number variants are enriched in regions of low-mappability

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

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

December 10, 2015

Abstract

Germline 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 can help identify CNVs, existing analytical methods suffer from limited sensitivity and specificity. Here we show that this is in large part due to the non-uniformity of read coverage, even after intra-sample normalization, and that this is exacerbated in regions of low-mappability. To improve on this, we propose PopSV, an analytical method that uses multiple samples to control for technical variation and enables the robust detection of CNVs. We show that PopSV is able to detect up to 2.7 times more variants compared to previous methods, with an accuracy of about 90%. Applying PopSV to 640 normal and cancer whole-genome datasets, we demonstrate that CNVs affect on average 7.4 million DNA bases in each individual, a 23% increase versus previous estimates. Notably, we find that regions of low-mappability, which were often concealed in previous analyses, harbor approximately 10 times more CNVs than the rest of the genome and that this contrasts with somatic CNVs (sCNVs) that are nearly uniformly distributed. We also observe that CNVs are found more than expected near centromeres and telomeres, in segmental duplications, in specific types of satellite repeats and in some of the most recent families of transposable elements. Although CNVs are found to be depleted in protein-coding genes, we identify 7206 genes with at least one exonic CNV, 324 of which harbored CNVs that would have been missed if low-mappability regions had been excluded. Similarly, 2253 trait- and disease-associated loci are observed to overlap at least one CNV. Our results provide the most comprehensive map of CNVs across the human genome to date and demonstrate the broad functional impact of this type of genetic variation including in regions of low-mappability.

1 Introduction

Structural variants (SVs) are defined as genetic mutations affecting more than 100 base pairs and encompasses several types of rearrangements: deletion, duplication, novel insertion, inversion and translocation[1]. Deletions and duplications, which affect DNA copy number, are also collectively known as copy number variants (CNVs). SVs arise from a broad range of mechanisms and show a heterogeneous distribution of location and size across the genome[1,2,3]. In healthy individual, SVs are estimated to cumulatively affect a higher proportion of the genome as compared to single nucleotide polymorphisms (SNPs)[4]. Numerous diseases including Crohn's Disease[5], schizophrenia[6], obesity[7], epilepsy[8], autism[9], cancer[10] and other inherited diseases[11,12], harbor SVs with a demonstrated detrimental effect[13,14,15].

While cytogenetic approaches and array-based technologies have been used to identify large SVs, whole-genome sequencing (WGS) could in theory uncover the full range of SVs both in terms of size and type[16]. Numerous methods have been implemented to detect SVs from WGS data using either paired-end information[17,18], read-depth (RD) variation[19,20,21], breakpoints detection through split-read approach[22] or de novo assembly[23]. However, existing approaches suffer from limited sensitivity and specificity[3,16], especially in specific regions of the genome, including regions of low-complexity and low-mappability[24,25]. One strategy to improve the accuracy of SV detection has been to use an ensemble approach that combines information from different methods relying on different types of reads. Large re-sequencing projects such as the 1000 Genome Project[3,26] and the Genomes of Netherlands (GoNL) project[27,28] have adopted this strategy and have successfully identified many SVs using an extensive panel of detection methods combined with low-throughput validation. Such a strategy increases the specificity of the calls, although validation rates can still be very low (e.g. 41 of the 601 predicted de novo SVs in Kloosterman et al.[28]). Moreover, joining the different sets of calls can be somewhat arbitrary and leads to an approach that is less sensitive. Finally, in these studies, as in many others, repeat-rich regions and other problematic regions are usually removed from the analysis to improve the accuracy of the calls. This is unfortunate given that CNVs in a number of such regions have already been associated with various diseases[29,30,31,12].

Why is calling SVs from WGS data so challenging? A major limitation of many of the existing detection methods is the incorrect assumption that reads are uniformly distributed across the genome. Indeed, it has been shown that various features of sequencing experiments, such as mappability[24,25], GC content[32], replication timing[33], have a negative impact on the uniformity of the coverage[34]. 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 will particularly impair the detection of SV with weaker signal, which is inevitable in regions of low-mappability, for smaller SVs or in cancer samples with stromal contamination or cell heterogeneity.

In this work, we start by showing that technical variation challenges the uniformity of coverage assumption despite state-of-the-art intra-sample normalization. To correct for this, we propose a new method, PopSV, an approach that relies on RD but uses a set of reference samples to control for technical variation and detect abnormal read coverage. Our approach differs from previous RD methods, such as RDXplorer[35] or CNVnator[20], 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[19] or BIC-seq[36], we show that they do not sufficiently control for experiment-specific noise as compared to PopSV. PopSV is also different from approaches such as cn.MOPS[21] and Genome STRIP[37] 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 SVs present in several samples but may miss sample-specific events. Moreover, their basic normalization of coverage and fully parametric models forces them to conceal a sizable portion of the genome and variants with weaker signal.

To demonstrate the utility of PopSV in characterizing CNVs across the genome, we apply the method to 640 WGS individuals from three human cohorts: a twin study with 45 individuals[38], a renal cell carcinoma datasets with 95 tumor and control pairs[39] and 500 unrelated individuals from the GoNL dataset[27]. Using this data we compare the performance of PopSV with existing CNV detection methods and validate the quality of the predictions. We also characterize the patterns of CNVs across the human genome and show that CNVs are enriched in regions of low-mappability and in different classes of repeats. Finally, we look at the functional significance of these structural variants and show that thousands of genes and genome-wide associated studies (GWAS) loci overlap CNVs. A number of these potentially important CNVs would have been missed if regions of low-mappability had been excluded.

2 Results

Intra-sample normalization does not remove coverage biases It is usually assumed that after correction for known biases such as GC content[32] and mappability[24,25], sequencing reads in a WGS experiment are uniformly distributed across the genome. To test this hypothesis, we computed the RD in non-overlapping genomic windows (bins) of size 5 kilo bases (Kb) in the normal samples of the renal cancer dataset. Read counts in the bins were corrected for GC-bias and, to be conservative in this initial analysis, regions with extreme read coverage were removed (Methods). Bin scores were then quantile normalized to obtain the same distribution for all samples (Fig. S1). Unexpectedly, and in contrast to simulated datasets, the inter-sample mean coverage in each bin was observed to vary from one genomic region to the other, highlighting the presence of additional biases (Fig. 1a). Supporting this observation, the bin coverage variance across samples was lower than expected and also varied between genomic regions (Fig. 1b). Such region-specific bias is overlooked when global estimates and genome-scanning methods are used to detect coverage differences. To further investigate this bias, we computed the proportion of the genome where a given sample had either the highest or the lowest coverage of all samples. Some samples looked more affected by this bias than others, as they consistently showed the highest, or the lowest, coverage across large portions of the genome (Fig. 1c). Similar patterns were also observed in the other two cohorts (Methods and Fig. S2 and S3). In short, we observed significant coverage biases even after intra-sample normalization and when focusing on the least problematic regions of the genome. This effect was even stronger when the whole genome was being evaluated (Fig. S4). The implications of this inter-sample variation, is that CNV detection approaches that assume a uniform coverage distribution[19,20,36] will have higher rates of false positives as the coverage will artificially fluctuate. Moreover, this experimental noise will confuse the detection of weaker signal, e.g. in low-mappability regions, for smaller CNVs or in cancer samples with stromal contamination or cell heterogeneity.

A population-based normalization and CNV detection method The main idea behind PopSV is to assess whether the coverage observed in a given location of the genome diverges significantly from the coverage observed in a set of reference samples. In PopSV, the genome is first segmented into bins and RD is computed for each sample as the number of reads with proper

mapping in each bin. In a typical design, the genome is segmented in non-overlapping consecutive windows of equal size, but custom designs could also be used. After normalization, the value observed in each bin is compared to the values observed in the reference samples and a Z-score is calculated (Fig. 2a, 2b and Methods). False Discovery Rate (FDR) is estimated based on these Z-score distributions and a bin is marked as abnormal based on a user-defined FDR threshold. The normalization step is critical here since we have shown that simple approaches will fail to give acceptable normalized RD scores (Fig. 1c). Moreover, with global median/variance adjustment or quantile normalization, the remaining subtle experimental variation impairs the abnormal RD test (Fig. S5a). With PopSV, we propose a new normalization procedure, which we call targeted normalization, that retrieves, for each bin, other genomic regions with similar profile across the reference samples and uses these regions to normalize RD (Methods). In contrast to other methods, targeted normalization shows better distribution features (Fig. S5b). It is important to note that it is critical for the success of this targeted normalization that the set of reference samples used is comparable to the tested samples. We have included in PopSV a set of exploratory tools to help assess this (Methods).

To demonstrate the effectiveness of PopSV, we first applied it to the twin dataset (Methods). Using 5 Kb bins, we observe smooth normal-like Z-score distributions and overall consistency of the bin values in the twin pairs (Fig. 2c). Applying the same methodology to the normal/tumor cancer cohort lead to similar results and highlighted, as expected, a large number of duplications and deletions in the tumors (Fig. S6). Encouragingly, in regions of low-mappability, the Z-score distribution was found to be identical to the one in regions of normal mappability (Fig. S7). Next, we estimated the copy number of each bin by dividing the RD in a given sample by the average RD across the reference samples multiplied by two, to reflect the fact that reference set is assumed to be diploid in each bin. We anticipate the copy number estimate to be reliable if the detected event spans the entire bin but less accurate for smaller event or partial signal (e.g. contamination or cell heterogeneity in cancer). The distribution of these copy-number estimates further supported the quality of the PopSV calls, with clear peaks around integer values especially for longer events (Fig. 2d). It's important to note that, in contrast to some of the other methods[21,37], this aggregation around integer values is completely independent of the calling process which only marked bins with abnormal RD.

Sensitivity and specificity of PopSV To evaluate the performance of PopSV, we compared it to FREEC[19] and cn.MOPS[21], two popular RD methods that can be applied to WGS datasets to identify CNVs. FREEC segments the RD values of a sample using a LASSO-based algorithm. It can use GC content and mappability information or RD from a control sample to normalize the signal before segmentation. In contrast, cn.MOPS considers simultaneously several samples and detects copy number variation using a Poisson model and a Bayesian approach. Here, RD is simply normalized to adjust for the total coverage in each sample. We used cn.MOPS for this comparison instead of GenomeSTRiP[37] because it was developed for a design similar to ours, i.e. tens of samples sequenced at high coverage.

First, in the twin study, we measured the number of CNVs identified in each twin that were also found in the matching twin (Methods). In this analysis, we focused on CNV calls found in less than 50% of the reference samples, as calls at very high frequency could be systematic errors. Using 5 Kb bins, PopSV recovered on average more concordant CNV events per sample, 324 versus 102 and 92 for FREEC and cn.MOPS respectively, while maintaining comparable specificity (Fig. 3a

and 3b). Notably, focusing on the regions of low coverage that account for 13% of the genome (Fig. S8 and Methods), we found that PopSV also outperforms the other approaches with 174 replicated events per sample on average, while cn.MOPS and FREEC only found 70 and 19 respectively. In those regions, PopSV had a slightly higher specificity with 96% of the calls being concordant (Fig. 3b). To explore the quality of the CNV calls further, we clustered individuals according to the CNV calls and compared the result to the known pedigree for these samples (Methods). We found that PopSV shows better concordance as assessed by the Rand index (Fig. S9). Even using only the regions of extremely low coverage resulted in a clustering dendogram mimicking almost perfectly the family relationships (Fig. 3c). Additionally, the distribution of CNV recurrence shows a clearer peak at 3-sample for PopSV (Fig. S10), which is expected due to the aggregation of CNVs present in both twins and one parent.

To further assess the performance of PopSV, we also tested the approach on the cancer dataset by comparing the agreement between germline events in tumor/normal pairs in a similar way as was done for the twin pairs. We observed comparable results with PopSV reporting on average 293 consistent CNV calls per sample while cn.MOPS and FREEC only detected 75 and 48 such events respectively (Table S1). Once again, the specificity of the different methods was comparable at around 88%. This was true overall as well as in low coverage regions where PopSV found twice as many replicated calls.

Resolution and validation of the PopSV calls To evaluate the performance of PopSV at different resolutions, we repeated the analysis of the twin dataset using 500 bp bins. With smaller bins there is more noise and long stretches of bins of low significance might be missed. For this reason, the 500 bp PopSV calls were combined with the 5 Kb calls (Methods). At this resolution, we observed that PopSV still found on average 1.7 and 6.3 times more concordant calls per sample compared to cn.MOPS and FREEC while maintaining similar specificity (Table S1). PopSV also detected on average 1.3 and 23.2 times more replicated variants in regions of low coverage compared to cn.MOPS and FREEC respectively, and had the highest specificity of all tested methods. Similar results were also observed with 500 bp bins in the renal cancer data set (Table S1).

Next, we assessed the performance in each genomic bin individually. When more than 90% of the twin pairs are consistently called in a given bin, we classify it as a reliable bin. We could use this measure to show that bins with different levels of mappability and repeat content were as likely to be reliable (Fig. S11), supporting PopSV's robustness and superior performance even in these challenging regions (Fig. S12). Using this metric, we also observed that a higher fraction of the genome is reliably called with PopSV compared to cn.MOPS and FREEC (1.5 and 2.7 more, respectively, Table S2 and Methods). These observations were replicated in the renal cancer dataset, where reliability is defined based on the proportion of the normal samples with consistent calls in their paired tumor (Table S1 and S2).

We also wanted to assess the consistency between the 5 Kb and the 500 bp individual calls. In theory, calls from the 5 Kb analysis should be supported by many 500 bp calls. We also expect large stretches of 500 bp calls to be detected in the 5 Kb analysis. This comparison is informative as it explores the quality of the calls, the size of detectable events and the resolution for different bin size. Overall, we find that 5 Kb calls are well supported by 500 bp calls, with only 14% of the 5 Kb bins not supported by any 500 bp bin (Fig. S13a). Even for these unsupported 5 Kb calls, we find that the 500 bp bins RD was consistently enriched (or depleted) although not enough to be called with confidence (Fig. S13b and S13c). This is expected given the higher background noise

in the 500 bp analysis that will reduce the power to call these variants. Next, we looked at the proportion of 500 bp calls, grouped by size, that were found in the 5 Kb calls. We find that the concordance gradually increases until the 500 bp calls reach 5 Kb in size where the concordance rises to nearly 100% (Fig. 3d). This suggests that PopSV is able to detect approximately 75% of the events as large as half its bin size, and almost all events larger than its bin size. As expected, only a small proportion of the small 500 bp calls overlap 5 Kb calls and they likely corresponds to fragmented larger calls. Considering the trade-off between bin size and noise, this suggests to run PopSV with a few bin sizes to better capture variants of different sizes.

Finally, 23 variants were chosen for experimental validation. We randomly selected one-copy and two-copy deletions, among small ($\sim$ 700 bp) and large ($\sim$ 4 Kb) variants. In addition, although more challenging to validate using standard approaches, 3 deletions in low coverage regions were also randomly selected. We visually inspected the chosen deletions in order to map the breakpoints and PCR primers were designed to target the whole deletion region (Methods). In total, 19 were successfully validated (83%), close to our *in silico* estimates (Table S3). Of note, one of the three low coverage deletions was successfully validated despite being embedded in a region that was more repetitive, which made he design of primers more difficult.

Global patterns of CNVs across the human genome Having demonstrated the sensitivity, specificity and resolution of PopSV, we wanted to characterize the global patterns of CNVs across the human genome. We started with an analysis of the twins and of the normal samples in the renal cancer dataset, both of which have an average sequencing depth around 40X. PopSV was used to make calls using 500 bp and 5 Kb bins, which were then merged to create a final set of variants as before. On average, 7.4 Mb in each 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 4 Kb and 70% of the variants found were smaller than 3 Kb. We compared our numbers to equivalent CNVs detected in the recent human SV catalogue from the 1000 Genomes Project[26] (Methods). In that study, 6.0 Mb was found to be variable on average in each genome (Table S4). As expected, the set of variants identified by PopSV included more variants in low coverage regions, explaining in part the $\sim$ 23% increase. While the study from the 1000 Genomes Project[26] explored a wider range of SVs, our set of variant is likely more representative of the distribution of CNVs in a normal genome since a broader portion of the genome could be analyzed.

Next, we applied PopSV to the 500 unrelated samples from the GoNL cohort (Table 1). Due to a lower sequencing depth ($\sim$13X), we used bins of size 2 Kb and 5Kb that gave the best signal to noise ratio (Methods). Slightly fewer variants were found in these samples mainly because of the reduced sequencing depth, which limits the detection of smaller CNVs. 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, 326 Mb were found to be affected by a CNV with more duplications (325,602) detected than deletions (248,937). This contrasts with the CNVs reported by the 1000 Genomes Projects[26] that were heavily skewed towards deletions (Table 1 and Table S4), likely due to the usage of different methods to detect various types of CNVs. The frequency distribution of deletions and duplications found using PopSV was also much more balanced compared with the ones from Sudmant et al.[26] (Fig. S14). Of note, we observed the same when comparing PopSV with other methods: PopSV's frequencies are more similar between deletions and duplications compared to FREEC (Fig. S15). As expected, both deletions and duplications

detected by cn.MOPS tend to be skewed towards more common events.

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)[40]. 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 3 datasets to randomly distributed regions of similar sizes (Fig. S16 and Methods). In an average genome, we found that 34% of the CNVs calls were within 1 Mb of a CTG, while we would have expected 11% by chance. To verify that these observations were not simply a consequence of the methodology used, we also looked at the somatic CNVs that we could detected in the renal 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 (Methods). As expected, somatic CNVs (sCNVs) were found to be significantly larger and to affect a much larger fraction of the genome (Table S5). Reassuringly, and in contrast to germline CNVs, sCNVs were not preferentially found near CTGs (Fig. S16b).

Notably, when looking at the genomic distribution of CNVs, we also observed that 10.9 and 37.5 times more variants were found in region of low and high coverage respectively, compared to regions with expected coverage (Fig. 4a). This effect subsists even when controlling for proximity to CTGs, adjusting for difference in false discovery rate and cannot be explained by different detection power since variants in regions of low and high coverage are actually harder to detect (Methods). In contrast to germline CNVs, sCNVs were once again found to be more uniformly distributed and were actually depleted in low and high coverage regions (Fig. S17). These results suggest that the enrichments of germline CNVs near CTGs and in regions of low-mappability are trustworthy and are probably the consequence of reduced selection pressure on these variants rather than the detection methodology.

Segmental duplications and various repeat families are more prone to harbor CNVs We wanted to characterize further the distribution of germline CNVs in relation to different genomic features, including looking at the contribution of segmental duplications, low-mappability regions and different repeat classes. To investigate these associations beyond the trends already observed with CTGs and low coverage regions, we simulated a set of matched control regions (Methods). Even after these corrections, we found that CNVs were enriched in simple repeats, segmental duplications and satellites repeats (Fig. 4b), with fold-enrichments of 1.2, 1.9 and 2.6 respectively. In contrast, and as expected, protein-coding genes and exons were found to be under-represented. The over-representation of CNVs in segmental duplications has been described before[2] and in a recent study[41], one-half of CNV base pairs were shown to overlap segmental duplications. Here, we found that on average $\sim 42\%$ of the CNVs overlap a segmental duplication in any given sample. The majority of these CNVs were fully contained in the overlapping segmental duplication (Fig. S18). In our results, segmental duplications with different levels of sequence similarity showed comparable levels of enrichment (Fig. S19).

Although it is known that satellites and simple repeats DNA are more unstable[42], 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 unit and we found that not all satellite repeats were equally likely to overlap a CNV (Fig. S20a). In particular, Alpha satellites, Satellite-like Repeat 1 and 2 (SATR1, SATR2), Centromeric Repeats (CER) and satellites with the specific motif $(GAATG)n$ and its reverse complement $(CATTC)n$

were found to have the highest fold enrichment. We also noted that the affected satellite repeats tend to span almost completely the variable regions (Fig. S18). Simple repeats are also grouped into families and we found the most significant enrichment for the $TA(/AT)$ tandem repeats (Fig. S20b). In contrast, the $GT(/TG)$ tandem repeats were significantly under-represented and the CAG repeat, known to cause Huntington disease[29], was not observed to be enriched nor depleted. Here the repeats tend to overlap just a fraction of the variant, but a clear subset of variant are fully covered by these tandem repeats (Fig. S18).

Finally, although transposable elements (TEs) as a whole did not show enrichment (Fig. 4b), the Other repeat class, which includes SVA repeats, was found to be significantly enriched in the three datasets (Fig. S20c). Moreover, looking at TEs at the level of individual repeat families, we found a number of them to be enriched including SVA D-F, L1Hs, and AluY (Fig. 4c). Surprisingly, a few older ERV families, including HERV-H that has been shown to be expressed and important in human embryonic stem cells[43,44], were also in the list of enriched TEs. Several families of older L1 repeats (e.g. L1PA2 to L1PA5) were also enriched and often implicated in what appears to be non-allelic homologous recombination (see examples in Fig. 4d and S21). Reassuringly, the somatic CNVs once again did not show any of these enrichments (Fig. S17).

Impact of CNVs on protein-coding genes and disease loci Although both small and large CNVs were depleted in genes (Fig. 5a and Methods), 7206 protein-coding genes were found to have an exon overlapping an event in at least one of the 640 normal genomes studied (Table S6). Moreover, if we included the promoter regions, at least 11341 protein-coding genes were potentially affected by at least one CNV (Methods). Next, to do a saturation analysis, we compared the number of genes found in our three cohorts at different sample sizes (Fig. 5b). None of the curves reached a plateau, suggesting that many more affected genes would have been identified with larger sample sizes.

At this point, we would also like to highlight that many genes contain or are located in close proximity to low-mappability regions (Fig. S22). Signal in these regions is frequently removed from CNV analyses to avoid misleading signal created by extreme read coverage. This is regrettable because these regions are known to harbour SVs and are likely to be unstable[45,46]. For instance, a recent analysis of long-reads sequenced from an haploid human cell line[46] found a large number of SVs around these regions. In our 640 normal samples, 324 genes were found to have an exon overlapping such a CNVs in a low coverage region and this number increased to 454 if we included the promoter regions (Table S6). Many of these genes belong to gene families known to be copy number variable, such as *ANKRD*, *NBPF*, *ZNF* or *CD* gene families. These CNVs are distinct from larger aberrations (example Fig. 5c) and could easily be missed by other approaches masking low coverage regions. Of note, we also found that 172 genes were affected by somatic exonic CNVs located within these low coverage regions (Table S7).

Finally, we wanted to see the number of Genome-Wide Association Studies (GWAS) loci that overlapped our comprehensive CNV catalogue. Similarly to protein-coding regions, GWAS hits were found to be significantly depleted for CNVs (Fig. 4b). Nevertheless, in our population of 640 genomes, more than two thousands different GWAS hits were covered by at least one CNV in one sample (Table 1). For example, a SNP associated with coronary artery disease locates within a low-coverage deletion affecting one of our samples (Fig. S23).

3 Discussion

Why are SVs so difficult to detect in WGS data? We have answered this question by showing that various experimental biases, which cannot be corrected for using basic intra-sample normalization, affect the uniformity of read coverage across the genome. These biases, if not considered, will impair the detection of CNVs. One thing that is important to note is that the amplitude of these biases varied from one cohort to the next and did not appear to be strictly linked to the sequencing platform used but also to the way the samples were prepared (Fig. 1, Fig. S2 and S3). With PopSV, samples that were sequenced with the same technology and protocols can now be analyzed jointly to control for these biases. When only a few samples are available this inter-sample normalization procedure might be less efficient but we estimate that with 20 reference samples or more PopSV will be preferable over methods working on single samples (or pairs of samples). We note that WGS is probably one of the most straightforward next-generation sequencing (NGS) protocol that only involves DNA extraction, shearing, sometimes amplification, and sequencing. It is likely that other NGS experiments, such as ChIP-Seq, are also similarly affected by sample preparation conditions and that these would also benefit from a similar inter-sample normalization procedure.

Comparing different calling methods is not straightforward, especially when different strategies are implemented. To begin, we compared PopSV cn.MOPS and FREEC using the same large bin size (5 Kb) in order to assess their ability to detect different types of signal: full versus partial signal, single versus multiple bin support, good versus low mappability. Next, we ran the methods with a smaller bin size (500 bp) to compare the methods in a situation with higher background noise. In each comparison we made sure that PopSV had similar specificity estimates compared to other methods, in order to reliably compare the sensitivity. We concluded that PopSV was more capable of detecting partial or single-bin signal (Fig. 3a and 3d), which is valuable to be able to observe smaller variants or variants in more challenging regions. Even when the background noise was significant, PopSV showed the best sensitivity and could reliably test a wider range of the genome (Table S1 and S2). In contrast to cn.MOPS, FREEC and ensemble methodologies[27,26], PopSV was also able to detect both deletions and duplications as efficiently (Fig. S14 and S15).

A notable strength of this new approach is that it enables the analysis of CNVs across the whole genome. Using PopSV on 140 normal genomes with high sequencing depth (∼40X) and 500 additional samples with medium coverage (∼13X), we found that regions of low coverage, which only represent 13% of the genome, overlap with 65% of the CNVs detected. 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. Having a more complete CNV catalogue also enabled an unbiased characterization of the CNV patterns across genome and potentially increases the power for trait-association studies. In particular, we were able for the first time to quantify the extent to which some regions in the genome are more prone to harbour such structural rearrangements. For example, we could describe genome-wide enrichment for different families of DNA satellites, simple repeats and several TE families, such as SVA, L1Hs and HERV-H.

Because PopSV looks for abnormal read coverage in each bin independently, it does not require the coverage to be uniform across the genome. For this reason, a natural extension of PopSV would be to apply it to targeted sequencing data, such as whole-exome sequencing data. In this context, the fragmented nature of the coverage and the differences in baseline from one region to another would seamlessly be integrated and corrected for by the set of samples used as a reference. Actually, several methods for CNV detection from whole-exome data that use information from other samples already exist[47,48], although they do not control for the biases described above the

way PopSV does. Similarly, another logical extension of PopSV would be to apply it not only to correctly mapped reads but also to discordant reads to detect abnormal discordant coverage. Here, any type of discordant mapping, such as read pairs with incorrect insert size, orientation or with only one pair mapped could be counted together or separately. Discordant reads are intrinsically difficult to work with because they are usually ambiguous and found in regions of low-mappability. Issues of ambiguous mapping are context-specific and are exceedingly difficult to model directly. The advantage of working with a set of reference samples, as in the PopSV framework, is that we would have a way to control for this variability empirically. A additional advantage of incorporating the discordant reads in PopSV is that it would also allow for defining more precise breakpoints for the SVs detected, including in regions of low-mappability.

In summary, we have presented a novel method that enables the systematic detection of CNVs across the genome. Applying this method to a set of 640 WGS datasets, we were able to produce the most comprehensive map of CNVs across the human genome to date and highlight the broad potential impact of this type of genetic variation including in regions of low coverage. In the future, 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.

4 Data and code availability

The PopSV R package and documentation are available at http://jmonlong.github.io/PopSV/. The scripts used to produce the graphs and numbers in this study have been deposited on https://figshare.com/s/ba79730bb87a1322480d. It also contains the necessary data to reproduce our results. The raw sequences of the different datasets have already beed deposited by their respective consortium (Methods).

5 Acknowledgments

This work was supported by a grant from the National Sciences and Engineering Research Council (NSERC-448167-2013) and a grant from the Canadian Institute for Health Research (CIHR-MOP-115090). SLG and GB are supported by the Fonds de Recherche Santé Québec (FRSQ-29493 and FRSQ-25348). Data analyses were enabled by compute and storage resources provided by Compute Canada and Calcul Québec. We are grateful to the team of the Québec Study of Newborn twins who provided the twin dataset and the Cagekid consortiums 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 July 9, 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.

6 Author Contributions

JM and GB conceived and designed the study. JM implemented the method and performed the analyses. JM, CM and SLG designed and performed experimental validation. GR, PC and SLG

contributed reagents/materials. Finally, JM and GB wrote the manuscript.

Abbreviation

Kb Kilo base.

SV Structural Variation or Structural Variant.

CNV Copy-Number Variation or Copy Number Variant.

WGS Whole-Genome Sequencing.

RD Read-Depth, also called read coverage or depth of coverage.

References

[1] I. M. Hall and A. R. Quinlan. Detection and interpretation of genomic structural variation in mammals. *Methods in molecular biology (Clifton, N.J.)*, 838:225–48, 2012.

[2] A. J. Sharp, Z. Cheng, and E. E. Eichler. Structural variation of the human genome. *Annual review of genomics and human genetics*, 7:407–442, 2006.

[3] R. E. Mills *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, 2011.

[4] A. W. Pang *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome biology*, 11(5):R52, 2010.

[5] S. a. McCarroll *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nature genetics*, 40(9):1107–1112, 2008.

[6] J. L. Stone *et al.* Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 455(7210):237–241, 2008.

[7] E. G. Bochukova *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, 463(7281):666–670, 2010.

[8] H. C. Mefford *et al.* Rare copy number variants are an important cause of epileptic encephalopathies. *Annals of Neurology*, 70:974–985, 2011.

[9] H. Stefansson *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature*, 505(7483):361–6, 2014.

[10] R. Beroukhim *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905, 2010.

[11] F. Balzola, C. Bernstein, G. T. Ho, and C. Lees. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls: Commentary. *Inflammatory Bowel Disease Monitor*, 11(1):26–27, 2010.

[12] S. Ayarpadikannan and H.-S. Kim. The Impact of Transposable Elements in Genome Evolution and Genetic Instability and Their Implications in Various Diseases. *Genomics & Informatics*, 12(3):98, 2014.

[13] H. V. Firth *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, 84(4):524–533, 2009.

[14] D. F. Conrad *et al.* Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, 2010.

[15] M. Spielmann and E. Klopocki. CNVs of noncoding cis-regulatory elements in human disease. *Current opinion in genetics {&} development*, 23(3):1–8, 2013.

[16] C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. *Nature reviews. Genetics*, 12(5):363–76, 2011.

[17] K. Chen *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9):677–81, 2009.

[18] M. R. Lindberg, I. M. Hall, and A. R. Quinlan. Population-based structural variation discovery with Hydra-Multi. *Bioinformatics (Oxford, England)*, pages 4–6, 2014.

[19] V. Boeva *et al.* Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics (Oxford, England)*, 27(2):268–269, 2011.

[20] 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):974–84, 2011.

[21] G. Klambauer *et al.* 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, 2012.

[22] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, 25(21):2865–71, 2009.

[23] A. Rimmer *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics*, 46(8):912–918, 2014.

[24] T. J. Treangen and S. L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics*, 13(1):36–46, 2012.

[25] S. M. Teo, Y. Pawitan, C. S. Ku, K. S. Chia, and A. Salim. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics (Oxford, England)*, 28(21):2711–8, 2012.

[26] P. H. Sudmant *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015.

[27] L. C. Francioli *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 46(8):818–825, 2014.

[28] W. P. Kloosterman *et al.* Characteristics of de novo structural changes in the human genome. *Genome Research*, 25(6):792–801, 2015.

[29] M. E. MacDonald *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6):971–983, 1993.

[30] S. M. Mirkin. Expandable DNA repeats and human disease. *Nature*, 447(7147):932–940, 2007.

[31] J. Rich, V. V. Ogryzko, and I. V. Pirozhkova. Satellite DNA and related diseases, 2014.

[32] Y. Benjamini and T. P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research*, 40(10):e72, 2012.

[33] A. Koren *et al.* Genetic Variation in Human DNA Replication Timing. *Cell*, 159(5):1015–1026, 2014.

[34] M.-S. Cheung, T. a. Down, I. Latorre, and J. Ahringer. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic acids research*, 39(15):e103, 2011.

[35] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome research*, 19(9):1586–92, 2009.

[36] R. Xi *et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proceedings of the National Academy of Sciences of the United States of America*, 108(46):E1128–36, 2011.

[37] R. E. Handsaker *et al.* Large multiallelic copy number variations in humans. *Nature Genetics*, 47(3):296–303, 2015.

[38] M. Boivin *et al.* The Quebec Newborn Twin Study into adolescence: 15 years later. *Twin research and human genetics : the official journal of the International Society for Twin Studies*, 16(1):64–9, 2013.

[39] G. Scelo *et al.* Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nature communications*, 5(May):5135, 2014.

[40] D.-Q. Nguyen, C. Webber, and C. P. Ponting. Bias of selection on human copy-number variants. *PLoS genetics*, 2(2):e20, 2006.

[41] P. H. Sudmant *et al.* Global diversity, population stratification, and selection of human copy number variation. *Science*, pages 1–16, 2015.

[42] K. A. Eckert and S. E. Hile. Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Molecular Carcinogenesis*, 48(4):379–388, 2009.

[43] D. Kelley and J. Rinn. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome biology*, 13(11):R107, 2012.

[44] X. Lu *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural & Molecular Biology*, 21(4):423–425, 2014.

[45] A. J. Sharp *et al.* Segmental duplications and copy-number variation in the human genome. *American journal of human genetics*, 77(1):78–88, 2005.

[46] M. J. P. Chaisson *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 2014.

[47] Y. Shi and J. Majewski. FishingCNV: a graphical software package for detecting rare copy number variations in exome-sequencing data. *Bioinformatics (Oxford, England)*, 29(11):1461–2, 2013.

[48] C. Wang *et al.* PatternCNV: a versatile tool for detecting copy number changes from exome sequencing data. *Bioinformatics*, pages 1–3, 2014.

[49] K. R. Rosenbloom *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Research*, 43(D1):D670–D681, 2015.

7 Figures

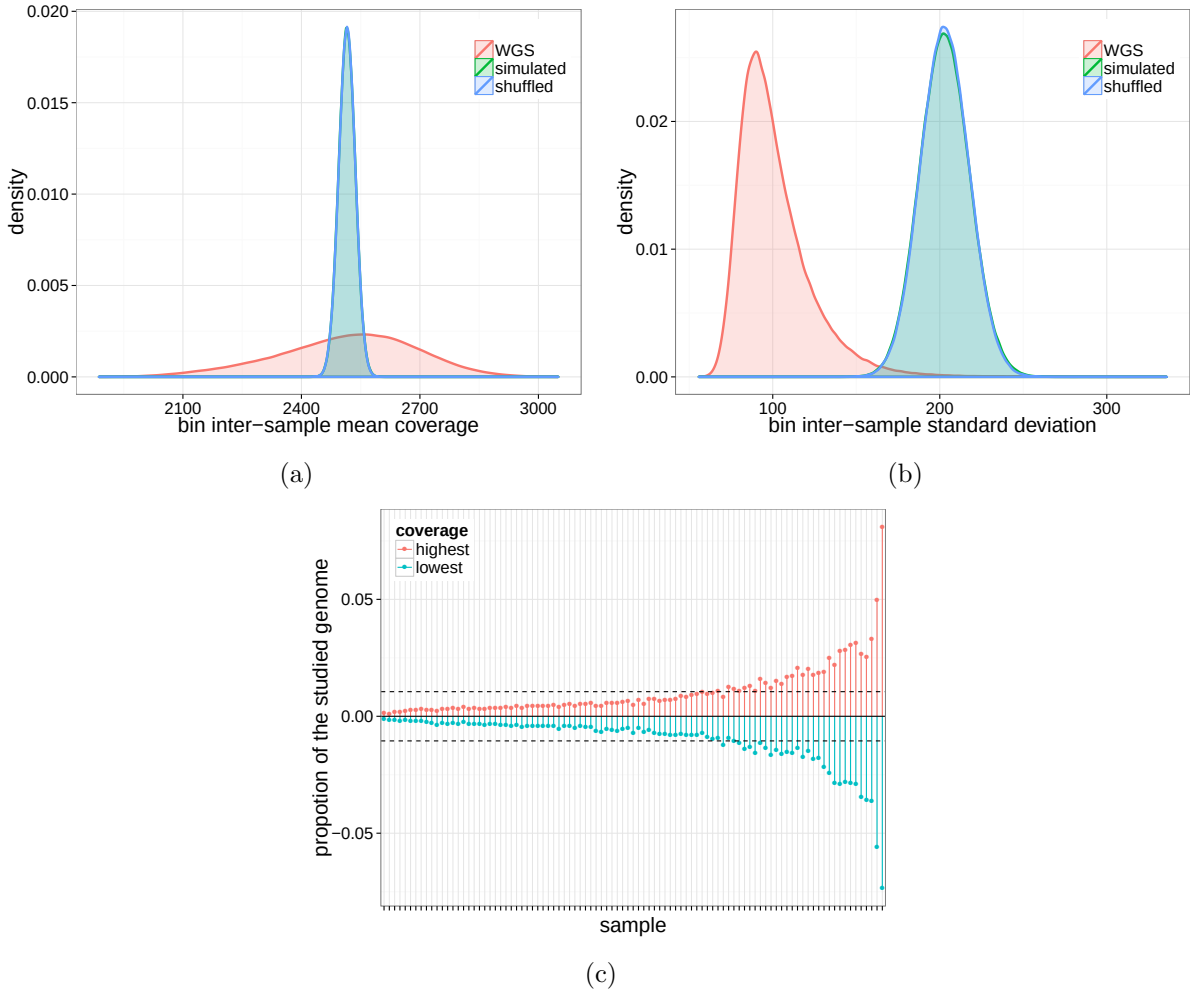
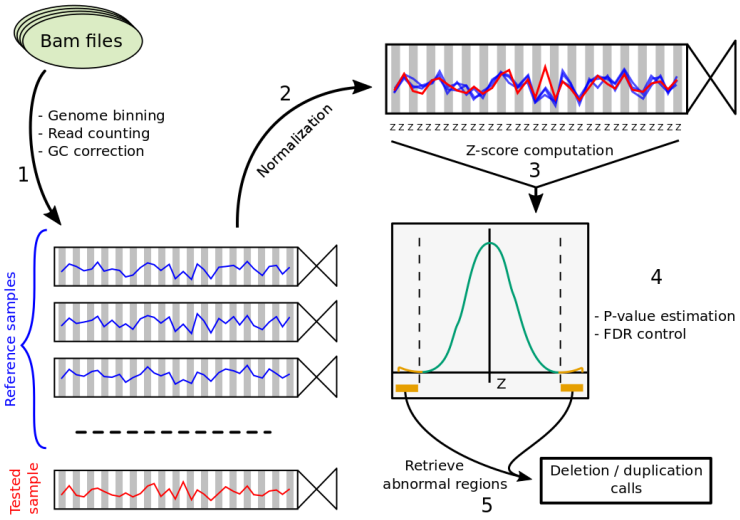
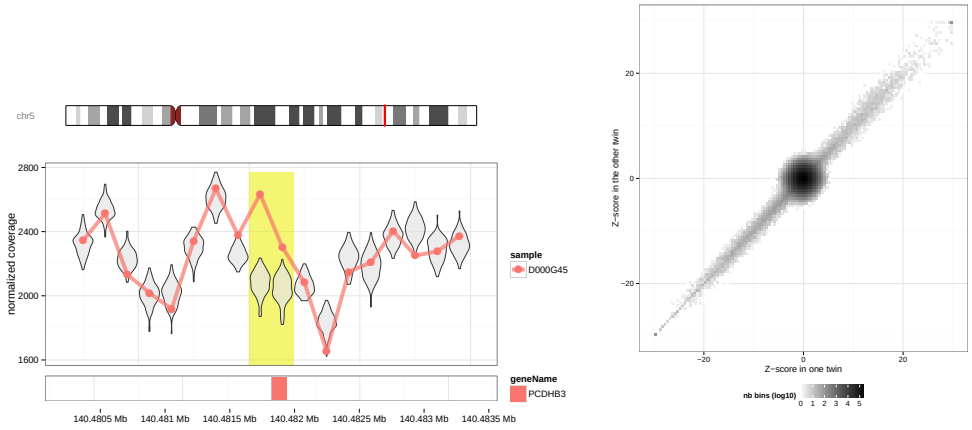


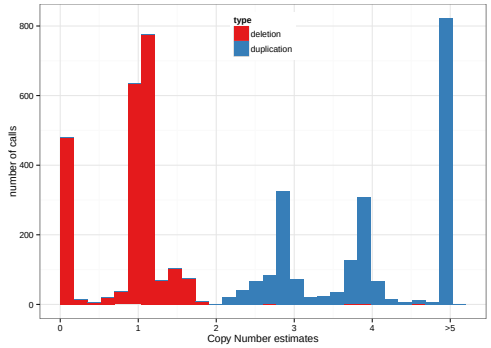
Figure 1: **Variation and bias in whole-genome sequencing experiments.** a) Distribution of the bin inter-sample mean coverage (red) and null distribution (blue: bins shuffled, green: simulated normal distribution). b) Distribution of the coverage standard deviation in each genomic bin. c) Proportion of the genome in which a given sample (x-axis) has the highest (red) or lowest (blue) RD. In the absence of bias all samples should be the most extreme at the same frequency (dotted horizontal line).



(a)



(b) (c)



(d)

Figure 2: **PopSV a Population-based approach for SV detection** a) First the genome is fragmented and reads mapping in each bin are counted for each sample and GC corrected (1). Next, coverage of the sample is normalized (2) and each bin is tested by computing a Z-score (3), estimating p-values (4) and identifying abnormal regions (5). b) The line and points represent the coverage of one sample with a duplication (highlighted in yellow); the violin plots represent the distribution of the coverage in the reference samples. c) Z-scores for all the genomic bins from two twins (x- y- axis). Non-zero positive or negative Z-score supports a duplication or a deletion, respectively. d) Focusing on large events, copy numbers can be estimated accurately and segregate close to integer values.

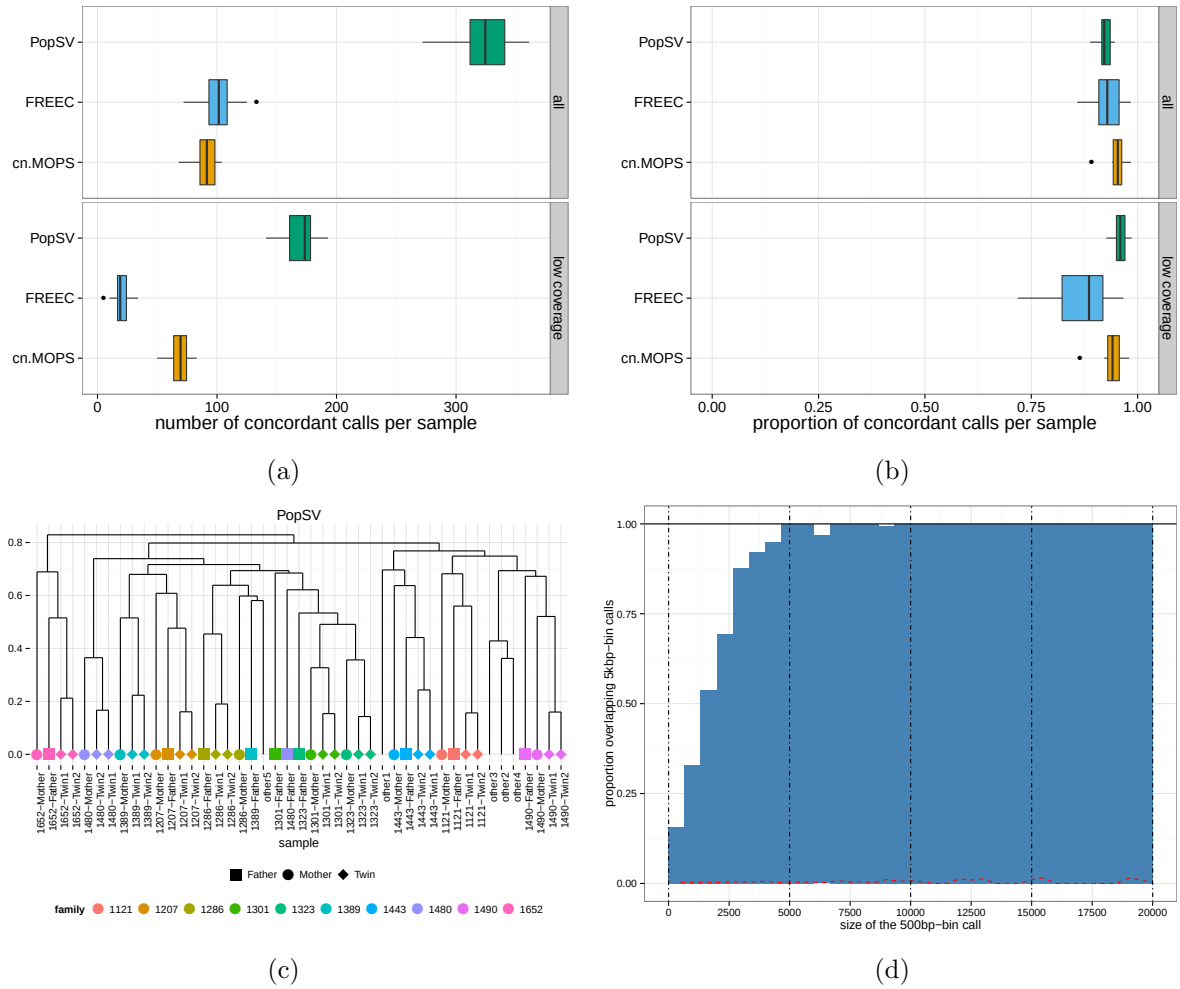


Figure 3: **Sensitivity, specificity and resolution of PopSV.** In the twin study and using 5 Kbp bins, number (a) and proportion (b) of variants from a twin also found in the other twin. c) Samples are clustered using PopSV calls in regions of extremely low coverage and recover almost perfectly the family structure. d) Proportion of 500 bp calls of different sizes (x-axis) overlapping a 5 Kbp call.

Set	Depth	Samples	Variants			Avg Size (Kbp)	Variants <3 Kbp		Affected genome (Mb)			
			Total	Per sample			Proportion	Per sample	Total	Per sample		
				WG	*LC*					*min*	*mean*	*max*
Twin study	42x	45	73677	1637.27	207.84	4.21	0.65	1056.84	62.22	5.30	6.89	9.03
deletion			32717	727.04	12.33	4.53	0.58	423.80	33.97	2.79	3.30	3.85
duplication			40960	910.22	195.51	3.94	0.70	633.04	34.20	2.50	3.59	5.29
Renal cancer germline	40x	95	202617	2132.81	251.31	3.58	0.71	1521.16	134.77	5.53	7.62	10.23
deletion			76483	805.08	6.48	4.30	0.63	508.56	70.65	2.65	3.46	7.26
duplication			126134	1327.73	244.96	3.14	0.76	1012.60	76.28	2.31	4.16	6.70
GoNL	13x	500	274759	549.52	84.59	8.71	0.46	250.24	226.50	3.05	4.78	8.12
deletion			131207	262.41	4.32	8.50	0.42	110.16	106.83	1.30	2.23	3.96
duplication			143552	287.10	80.32	8.91	0.49	140.08	139.21	1.45	2.55	5.67

Table 1: **CNVs in Twin, renal cancer germline and GoNL datasets.** WG: whole genome; LC: low-coverage regions. *Affected genome* represents the amount of the reference genome that affected by a CNV.

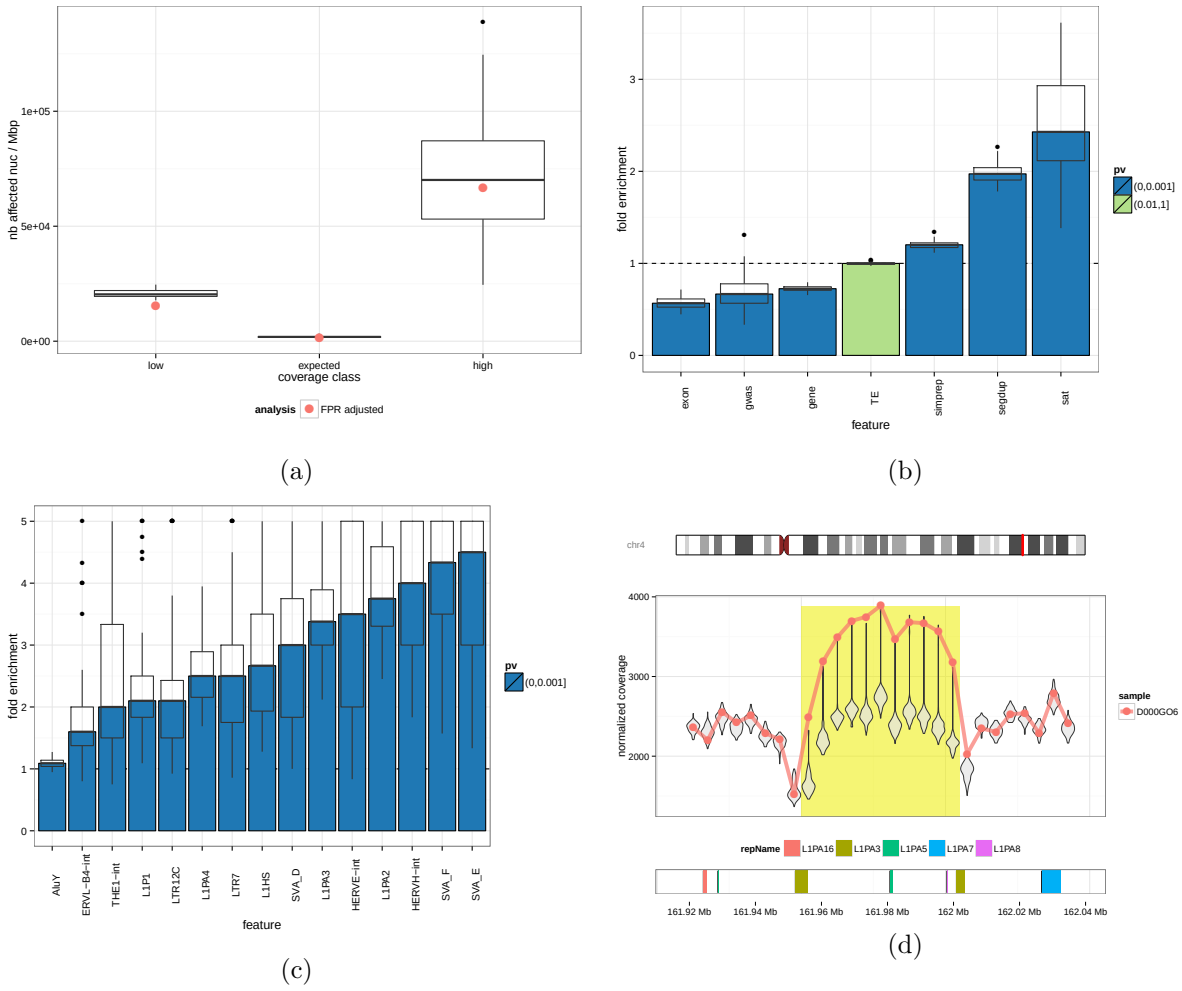


Figure 4: **CNVs are enriched in regions of low-mappability and near centromeres and telomeres.** a) The number of nucleotide affected by CNVs per Mbp in each sample varies across coverage classes. b) Enrichment of CNVs in different genomic classes. Bars show the inter-sample median, boxplot shows the variation across samples. c) Enrichment of CNVs in the top 15 most-enriched transposable element sub-families. Fold enrichment were winsorized at 5 for visibility. d) Example of CNV likely caused by non-allelic homologous recombination between two L1PA3 repeats. Same representation as Figure 2b.

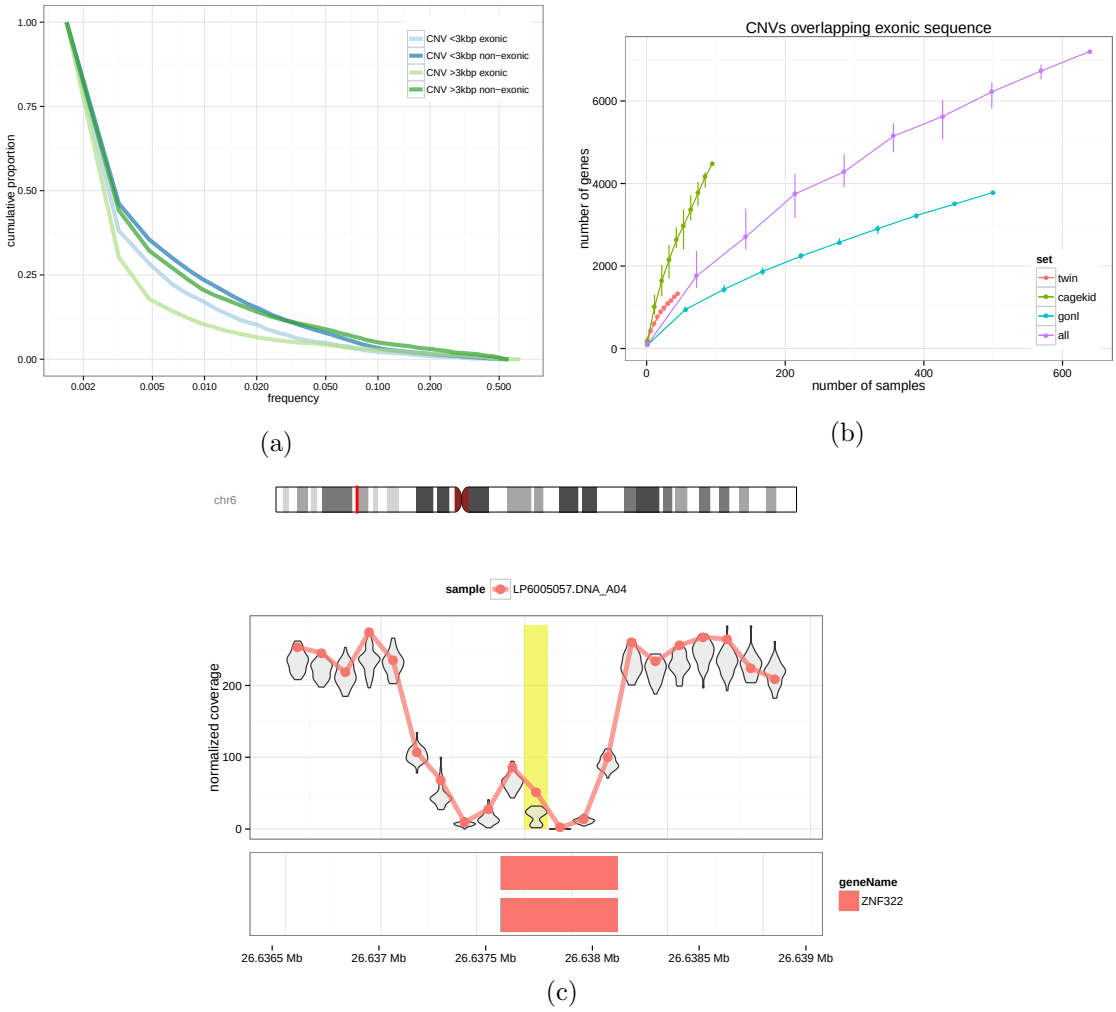


Figure 5: **Functional impact of CNVs.** a) Frequency of CNVs of different sizes and overlap with coding exons. b) The curve shows the number of protein-coding genes with exonic sequence affected by a CNV, at different sample size. c) Example of a duplication in *ZNF322* exon, located in a challenging region (low coverage). Same representation as Figure 2b.

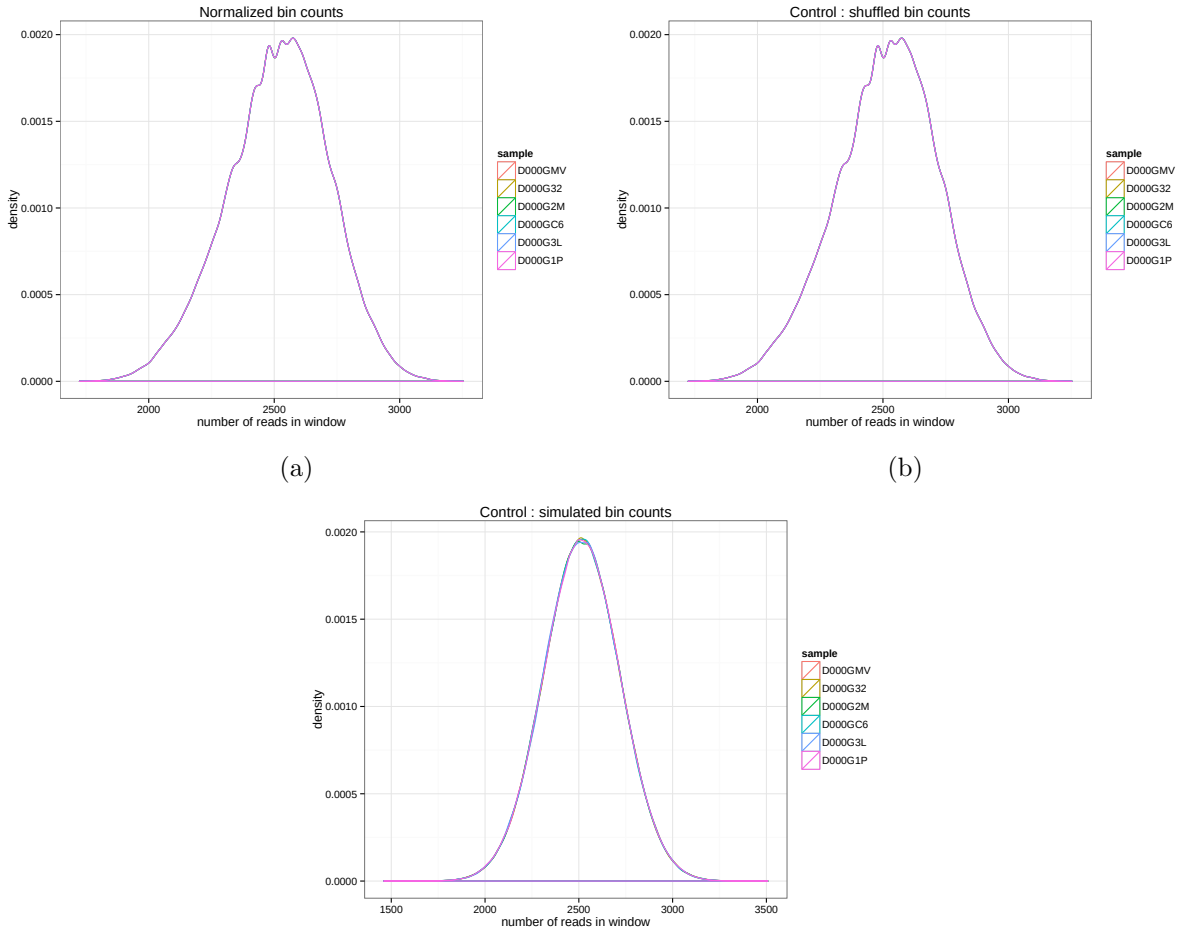
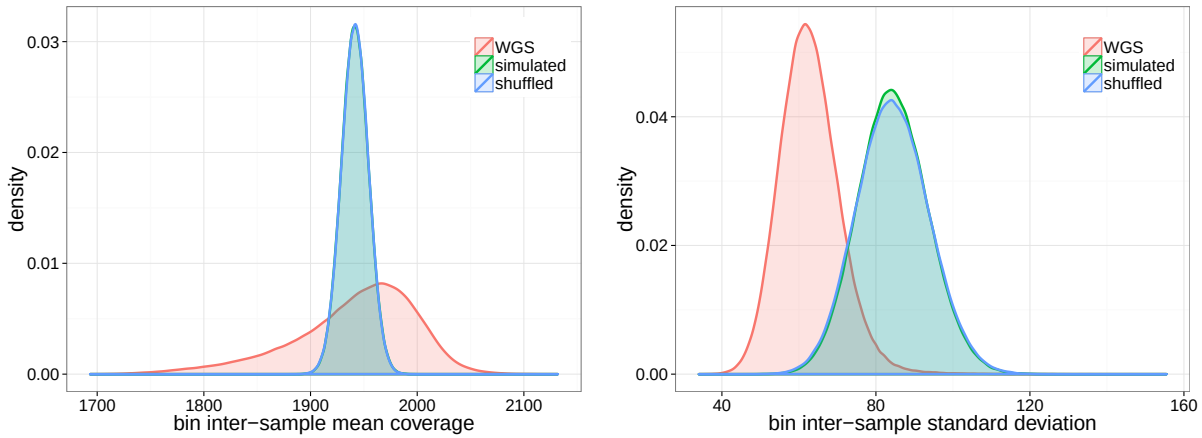


Figure S1: **Distribution of the bin counts after removal of regions of extreme coverage and normalization.** a) All samples have exactly the same RD distribution after quantile normalization. b) We build the distribution under the null hypothesis (i.e. uniform coverage) by shuffling the bins or c) simulating RD from a Normal distribution.



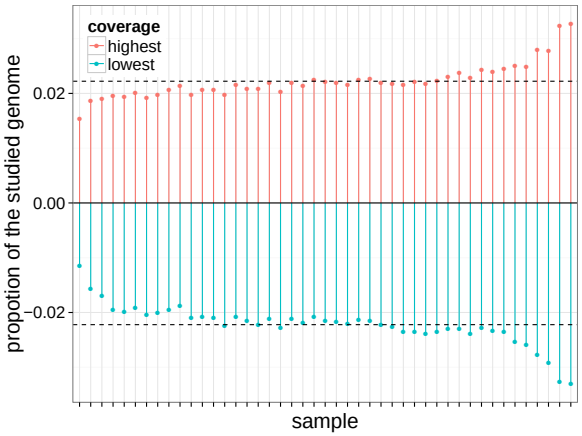
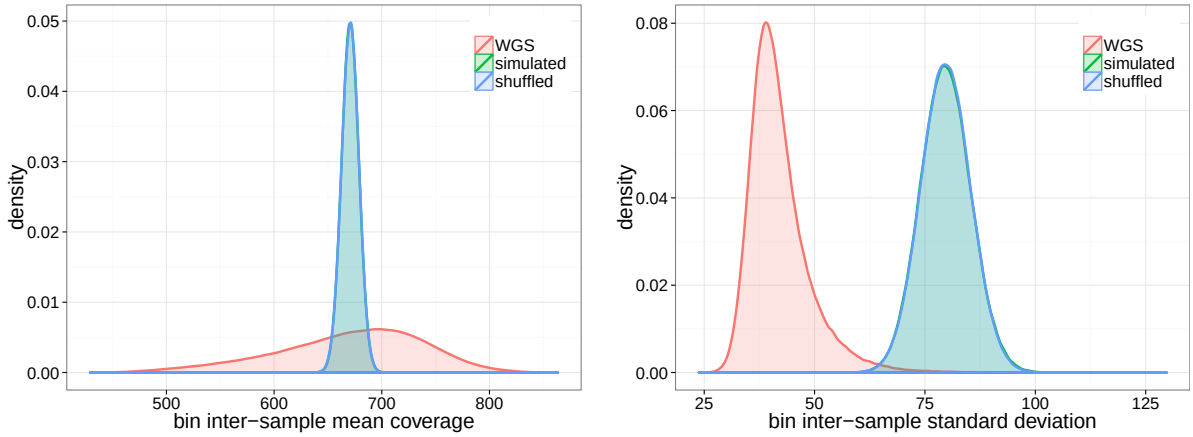


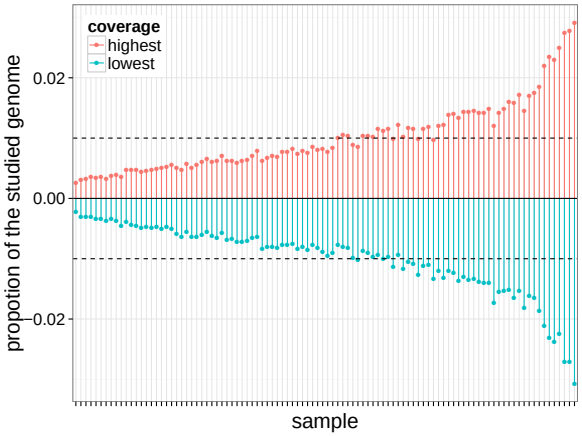
Figure S2: **Variation and bias in whole-genome sequencing in the *Twins* dataset.** a) Average bin RD across the samples (red) and null distribution (blue: bins shuffled, green: simulated normal distribution). b) Same with standard deviation. c) Proportion of the genome in which a sample (x-axis) has the highest(red) or lowest(blue) RD. In the absence of bias all samples should be the extreme one with the same frequency (dotted horizontal line).

Dataset	Region	Bin size	Number of concordant calls			Fold change PopSV vs		Proportion of concordant calls		
			PopSV	FREEC	cn.MOPS	FREEC	cn.MOPS	PopSV	FREEC	cn.MOPS
Twin study	whole genome	5kbp	324.5	101.5	91.5	3.20	3.55	0.92	0.93	0.95
		500bp-5kbp	883.0	140.0	506.5	6.31	1.74	0.89	0.92	0.88
	low coverage	5kbp	173.5	19.0	69.5	9.13	2.50	0.96	0.89	0.94
		500bp-5kbp	546.0	23.5	407.5	23.23	1.34	0.94	0.90	0.87
Renal cancer	whole genome	5kbp	293.0	48.0	75.0	6.10	3.91	0.88	0.80	0.88
		500bp-5kbp	949.0	80.0	564.0	11.86	1.68	0.79	0.72	0.75
	low coverage	5kbp	107.0	6.0	52.0	17.83	2.06	0.91	0.78	0.87
		500bp-5kbp	445.0	2.0	267.0	222.50	1.67	0.83	0.62	0.72

Table S1: **Concordance in different datasets, methods and bin size.**



(a) (b)

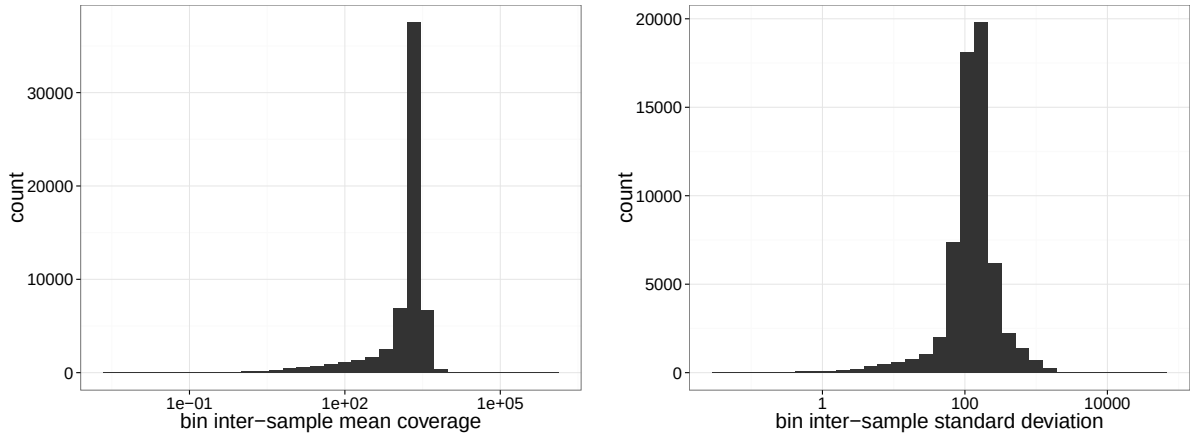


(c)

Figure S3: **Variation and bias in whole-genome sequencing in the GoNL dataset.** a) Average bin RD across the samples (red) and null distribution (blue: bins shuffled, green: simulated normal distribution). b) Same with standard deviation. c) Proportion of the genome in which a sample (x-axis) has the highest(red) or lowest(blue) RD. In the absence of bias all samples should be the extreme one with the same frequency (dotted horizontal line).

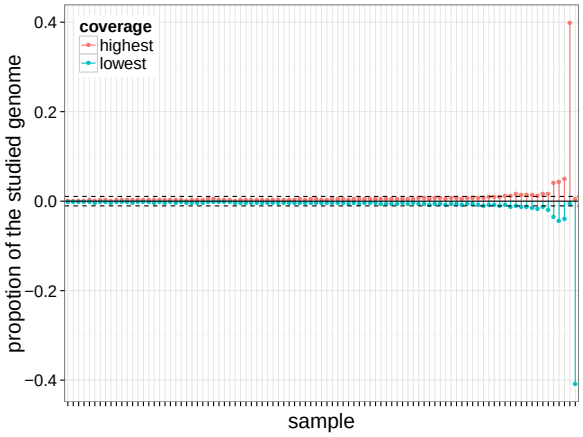
Dataset	Bin size	Number of reliable 1 Mb bins			Fold change PopSV vs	
		PopSV	FREEC	cn.MOPS	FREEC	cn.MOPS
Twin study	5kbp	1260	753	353	1.67	3.57
	500bp-5kbp	2034	762	1360	2.67	1.50
Renal cancer	5kbp	2107	808	484	2.61	4.35
	500bp-5kbp	2699	1149	2106	2.35	1.28

Table S2: **Amount of genome reliably tested in different datasets, methods and bin size.**



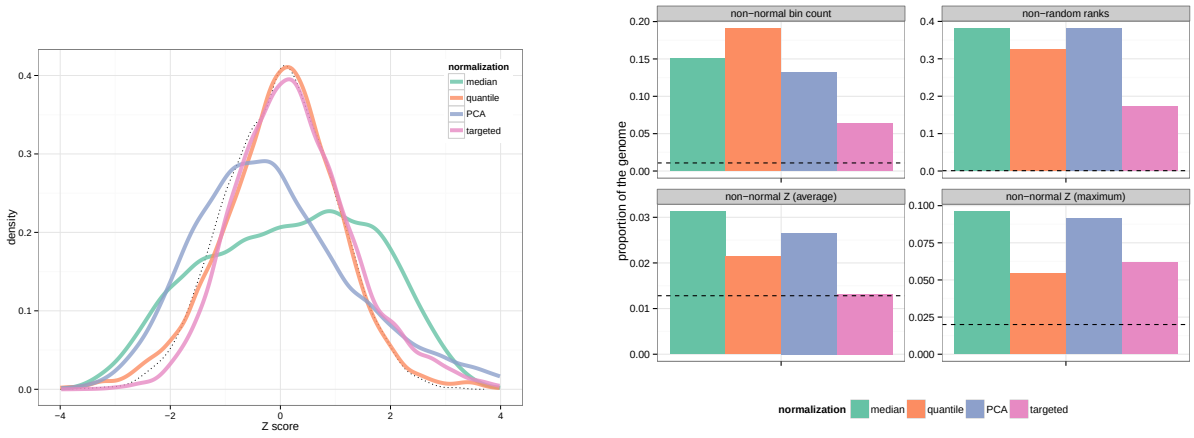
(a)

(b)



(c)

Figure S4: **RD bias is stronger when including all genomic regions.** In renal cancer normals, the same analysis as summarized in Fig. 1 is performed using all genomic regions, i.e. without filtering for extreme coverage. Quantile normalization is used again to force the same RD distribution in all samples. Of note, in a) and b) the distribution of the mean and variance across samples is shown on a log-scale as it spans several orders of magnitude.



(a) (b)

Figure S5: **Comparison of different normalization approaches.** a) For each normalization approach, the sample with the least normal Z-score distribution is shown. b) After targeted normalization, a lower proportion of the genome looks problematic for the analysis. Less bins have non-normal bin counts (top-left), the sample ranks are more random pointing at less sample-specific bias (top-right), and Z-scores fit better a Normal distribution on average (bottom-left) and in the worst sample (bottom-right). The dotted line is computed from simulated bin counts.

Validated	Chr.	Start	End	Class	Left PCR primer	Right PCR primer
V	3	6649794	6654897	large CN 0	CCTTAGTATTTCAGTGGTTTCTGTAGGTAT	ATAAATATCAGTGCTCAACTTGGACTT
V	5	127407030	127411341	large CN 0	TATTCATATTAACCTATCCTCACAGAAAGA	TTTTTAAGAGATTTGAACTAAAATTCCAC
V	3	5535139	5539535	large CN 0	TACTTTTTGAATTTGTAAATTTCCTTTGTA	GAAATCAGAAAATCAAGATCATACTGAAG
V	1	116229111	116233162	large CN 0	GTGTTACAGAATTAGTTTTACTGAGTGGTC	ATCTATAAAGAACTTTTTCCAAATAAACCA
V	1	158961082	158966958	large CN 1	GTAGAATGAGCTGTGTTATGAGATGGT	ATGACTTTCTATTGTTTGAAATGTAGTGAC
V	15	26748887	26752614	large CN 1	CAATTTATCTATCAAGTTATTTCACGGTAG	AGTGAGATTTCATTTTAAGCTTGTCTTC
V	6	33937344	33942846	large CN 1	ACATTGTAGCCTGATGACCTTGTTC	TGTGTTCTGAGGTTTACTTTATAATCTAGG
V	12	82095501	82099389	large CN 1	ACCTATAACTAAGTGTAGCTGCTGTAACTG	TCAGTAAAAATGATTACTACAGTGGAAAAT
V	5	8255604	8260914	large CN 1	TGAACATACATTCATACACACATAATACAA	TACATCACTGAACAAACCTCTATAGTCATA
V	20	7398397	7403743	large CN 1	AATAAACATTCTCTATAAACCCTAAAATGG	CTTTGTACCATATTTCATAAACGTAGAGTC
V	18	40053822	40057873	large CN 1	TAACTTTCTTTTCTAAAGCTTTTGGAGTAT	GTGAATTAAGATTCAATGTCTCTGCTAATA
V	16	48904951	48906510	small CN 0	TCTTATTTATTTTGACAGTCCTTTACTCTG	AGATAATCAACTCTTTGTTTATTCTTTCAG
V	2	241086647	241087801	small CN 0	ATCAACATTTAGCCAGTGTTGTCTTAG	GTCTCTTGTGCTCTATCTTTGGCTT
V	13	110221621	110222631	small CN 0	ACCTCAGGAGAACTACTTCATACATTTCTA	GTATGAAAAACACTCATGGATATCATTTCT
V	11	60571017	60572170	small CN 0	AATGTTGAAGTGTGTCTTTCTGTAATATCT	GTGTTTTGTGTCGCTATTTGTTTAGTA
V	5	166402295	166404219	small CN 0	TCACTTTATTCATAACATTTCAGTGTAGAG	GATCATATGCTTAAAATGCTAATGAGG
N	3	160126422	160127288	small CN 1	TAAGATACAAGAAATAGAGATAACACTGGG	TCTGAACACTTATTTTAAGAAAATGAAAAA
N	17	10612674	10613775	small CN 1	AATTTAGCAGTCTCTTACATTTCTTCTACC	TCTCTTCTATAAAAATAAATGGCTAAAAGC
V	10	70253713	70255155	small CN 1	AATAAAATCAAAGGTGATATTACTGACAGA	ATATACTCTTTTAACTTTTGACCATTTTGG
V	8	53700635	53702050	small CN 1	TAAGGAAAATTTAGTATAGTCTGGACCTGT	ATGGAAATATATCTCTGATGGGTGAC
N	6	26636844	26637539	low coverage	GTACATAGATTCTCACCCACAATTAAATC	CTTCTTCAACATCAGACAGTACACATT
N	13	78236245	78238105	low coverage	GTCAGTCTGGTTCTTTTCTGTCAAG	ACTTTAGTAAAATTGTTATTTAGTCCCAGG
V	1	248546279	248548008	low coverage	CTATCTTTCTTACCATTTAATATCTGCCTT	AGACTTCATTTAGGAAAGTGAGAAATACAC

Table S3: **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.

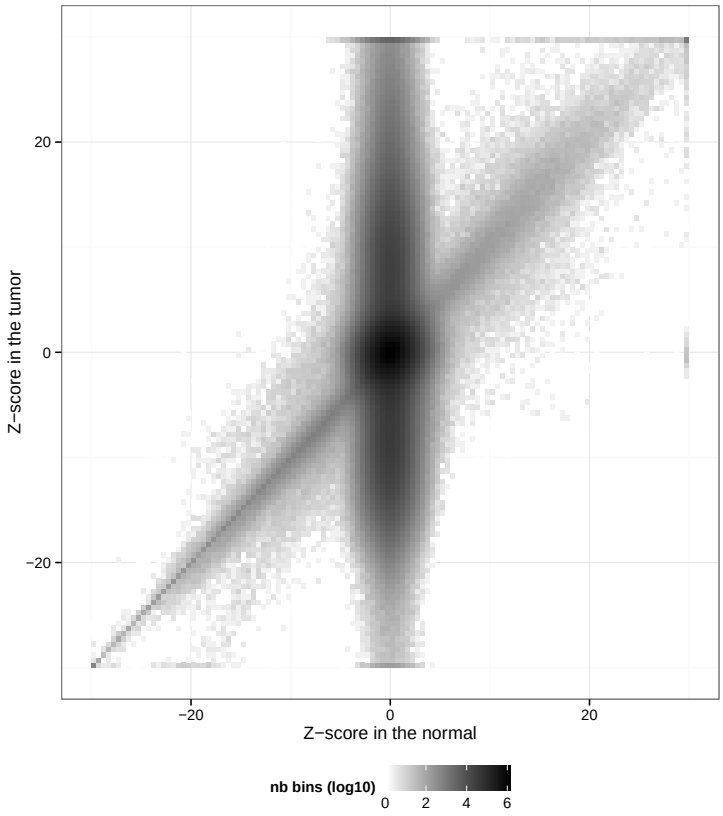


Figure S6: **ZZ plots between normal and tumor pairs.** In renal cancer, Z-scores from each normal samples is plotted against Z-scores from its tumor samples. This graph is an aggregation of all normal/tumor pairs. Z-scores are winsorized at -30 and 30 for visibility purpose.

Sample	Type	Variant		Variant per sample		Variant < 3 Kbp		Affected genome	
		All	Low-coverage	All	Low coverage	All	Per sample	All	Per sample
2504	All	2382489	3628	924	2	1420566	551	581.08	6.04
	CNV	312401	0	124	0	0	0	85.05	2.74
	DEL	2041543	3628	787	2	1420566	551	298.70	3.13
	DUP	28545	0	11	0	0	0	264.09	0.32

Table S4: **1000 Genomes deletions, duplications and CNVs.** We removed variants with high frequency (> 80%), variants in the chromosome X, and variants smaller than 300 bp in order to compare with PopSV's numbers (Table 1).

Set	Sample	Variant			Size (Kbp)	Variant <3 Kbp		Affected genome (Mb)			
		All	Per sample			Proportion	Per sample	All	Per sample		
			all	*low-coverage*					*min*	*mean*	*max*
Renal cancer somatic	95	391860	4124.84	44.40	58.54	0.48	1966.36	2455.18	4.16	232.83	664.86
deletion		194181	2044.01	2.72	70.81	0.42	865.64	1695.56	0.01	136.35	413.66
duplication		197679	2080.83	43.68	46.50	0.53	1100.72	1464.00	0.12	96.48	367.53

Table S5: **Somatic CNVs in renal cancer dataset.**

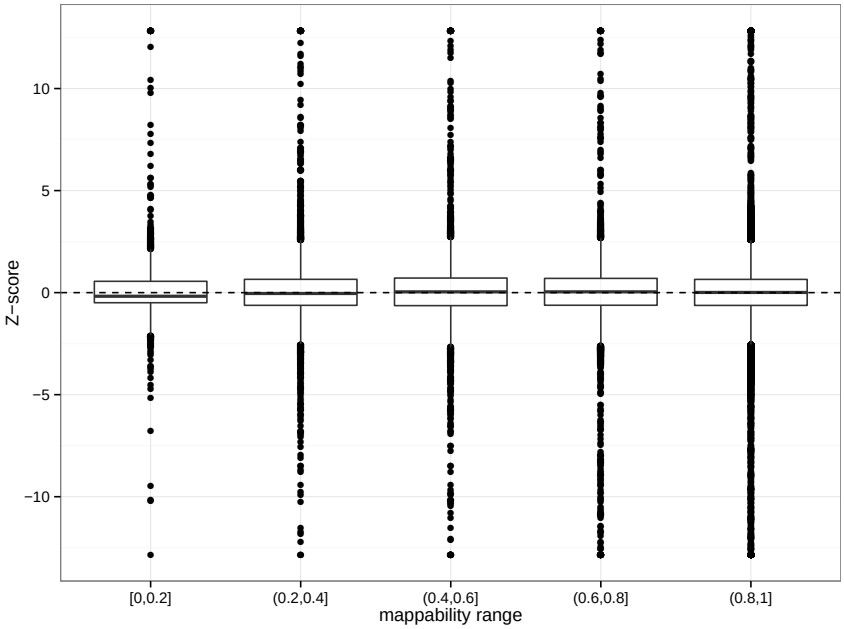


Figure S7: **Z-score distribution versus the mappability of the bin.** One randomly selected sample from the Twin dataset. Mappability was extracted from the UCSC track (Methods).

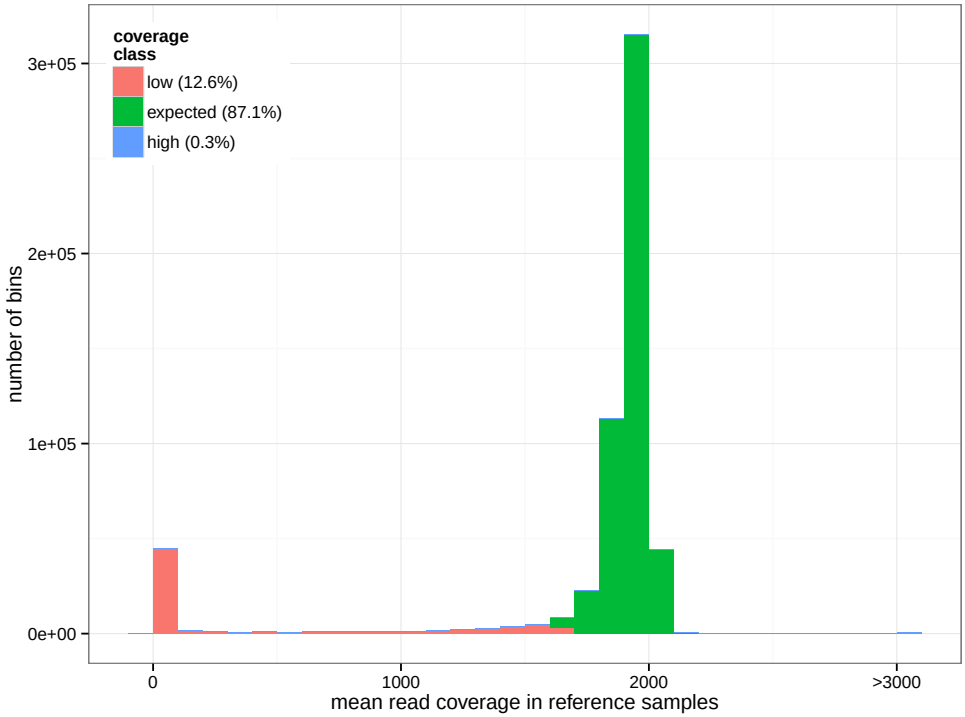


Figure S8: **Average coverage in 5 Kbp bins across reference samples in the** *Twins* **dataset.**

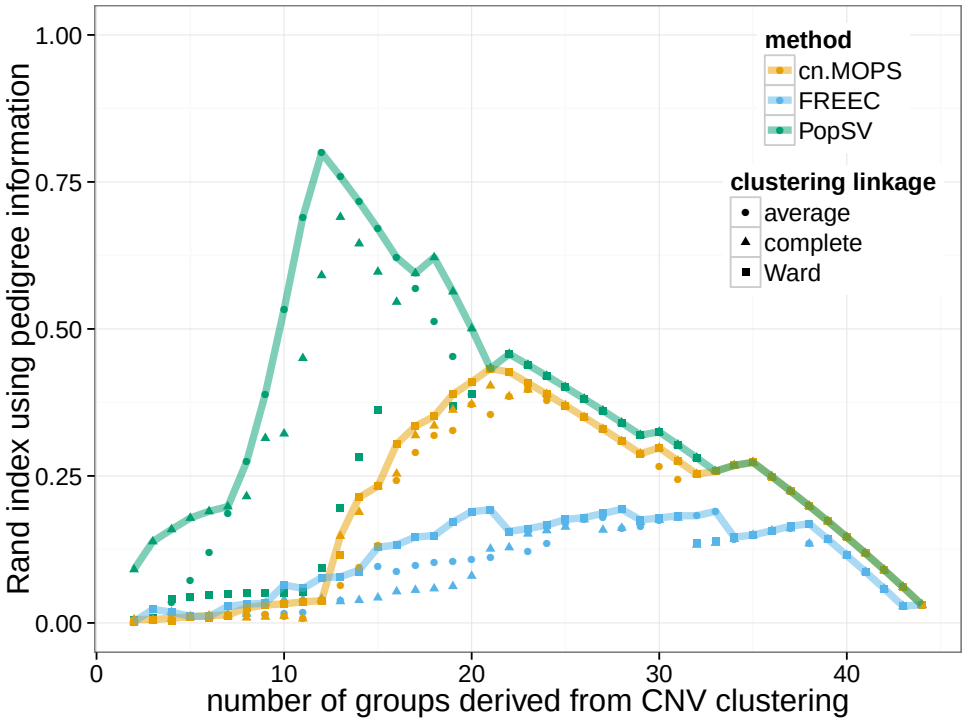


Figure S9: **Twin dataset: sample clustering and pedigree.** Samples are clustered using the CNV calls from the different methods (*colors*). The amount of genomic sequence called in only one of two samples defines the distance used for clustering. After cutting the hierarchical cluster at different levels (*x-axis*), cluster groups are compared to the known pedigree using the Rand index (*y-axis*). Different clustering linkage criterion (*point style*) are used and the one showing the best Rand index is highlighted by the line.

Set	Sample	Affected genome (Mbp)	Genes with CNVs			Genes with CNVs in low coverage			GWAS
			Exon	+ Promoter	+ Intron	Exon	+ Promoter	+ Intron	
Twin study	45	62.22	1337	2761	4617	147	283	329	351
deletion		33.97	824	1805	3300	12	28	35	245
duplication		34.20	664	1330	2263	145	277	322	179
Renal cancer germline	95	134.77	4487	8126	10638	224	339	381	992
deletion		70.65	2439	4814	7136	8	17	29	458
duplication		76.28	2567	4822	7042	223	338	380	575
GoNL	500	226.50	3785	5586	7173	226	357	406	1432
deletion		106.83	1728	2790	4130	10	21	24	652
duplication		139.21	2538	3638	4762	224	355	403	883
Total	640	325.66	7206	11341	13259	324	454	514	2253
deletion		165.26	4018	7157	9466	28	55	72	1108
duplication		194.42	4514	7304	9439	322	452	510	1338

Table S6: **Impact of CNVs.** Genes are protein-coding genes and promoter region is defined as the 10 Kbp region upstream of the transcription start site.

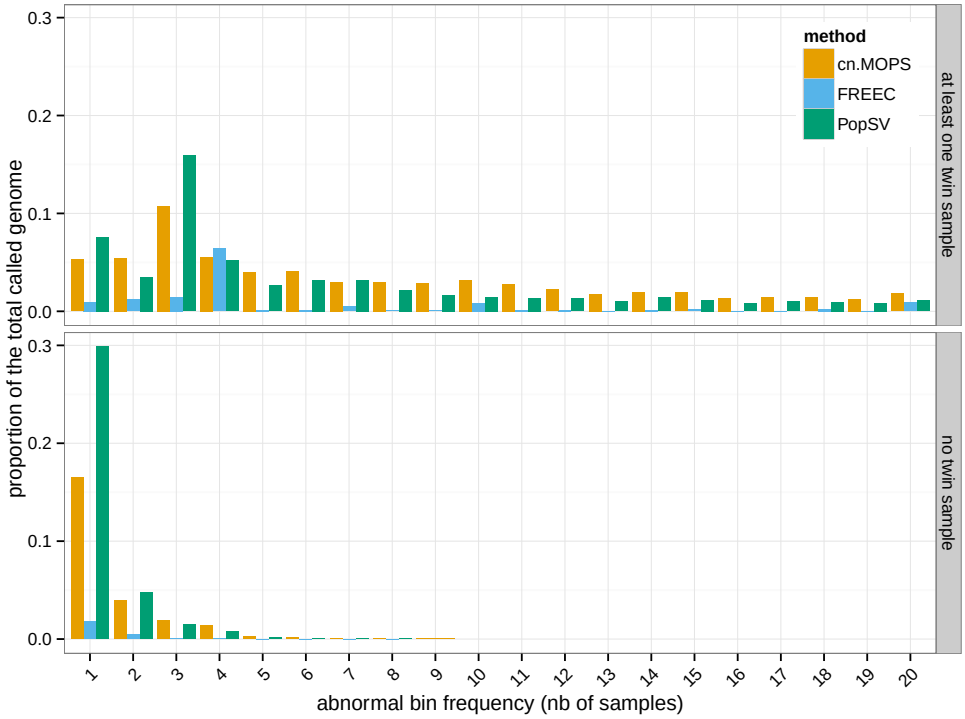


Figure S10: **Twin dataset: recurrence distribution.** The distribution of the event frequency shows a nice peak at 3-samples frequency when focusing on regions involving at least one twin (top). Using regions with no twin involved (bottom), the 3-samples peak should disappear.

Set	Sample	Affected genome (Mbp)	Genes with CNVs			Genes with CNVs in low coverage			GWAS
			Exon	+ Promoter	+ Intron	Exon	+ Promoter	+ Intron	
Renal cancer somatic	95	2455.18	18121	18909	18969	172	295	328	19886
deletion		1695.56	12931	14185	14677	6	16	24	13813
duplication		1464.00	13535	15535	16316	171	291	323	11570

Table S7: **Impact of somatic CNVs.** Genes are protein-coding genes and promoter region is defined as the 10 Kbp region upstream of the transcription start site.

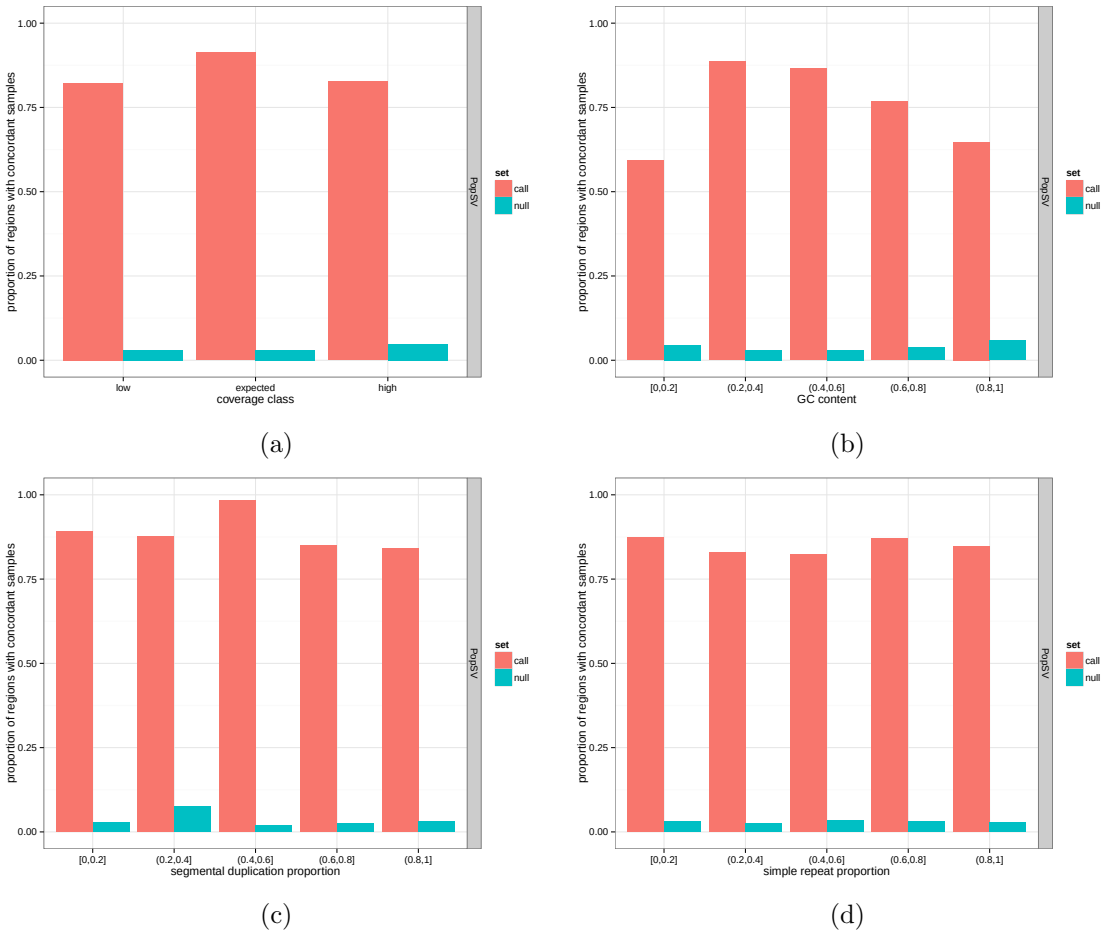


Figure S11: **Twins concordance per bin.** Bins are grouped by coverage class (top-left), GC content (top-right), segmental duplication content (bottom-left) and simple repeat content (bottom-right). Concordance is defined using the twin pairs. The null proportion (blue) represents the proportion expected by chance. It is computed as the concordance when randomly selecting samples.

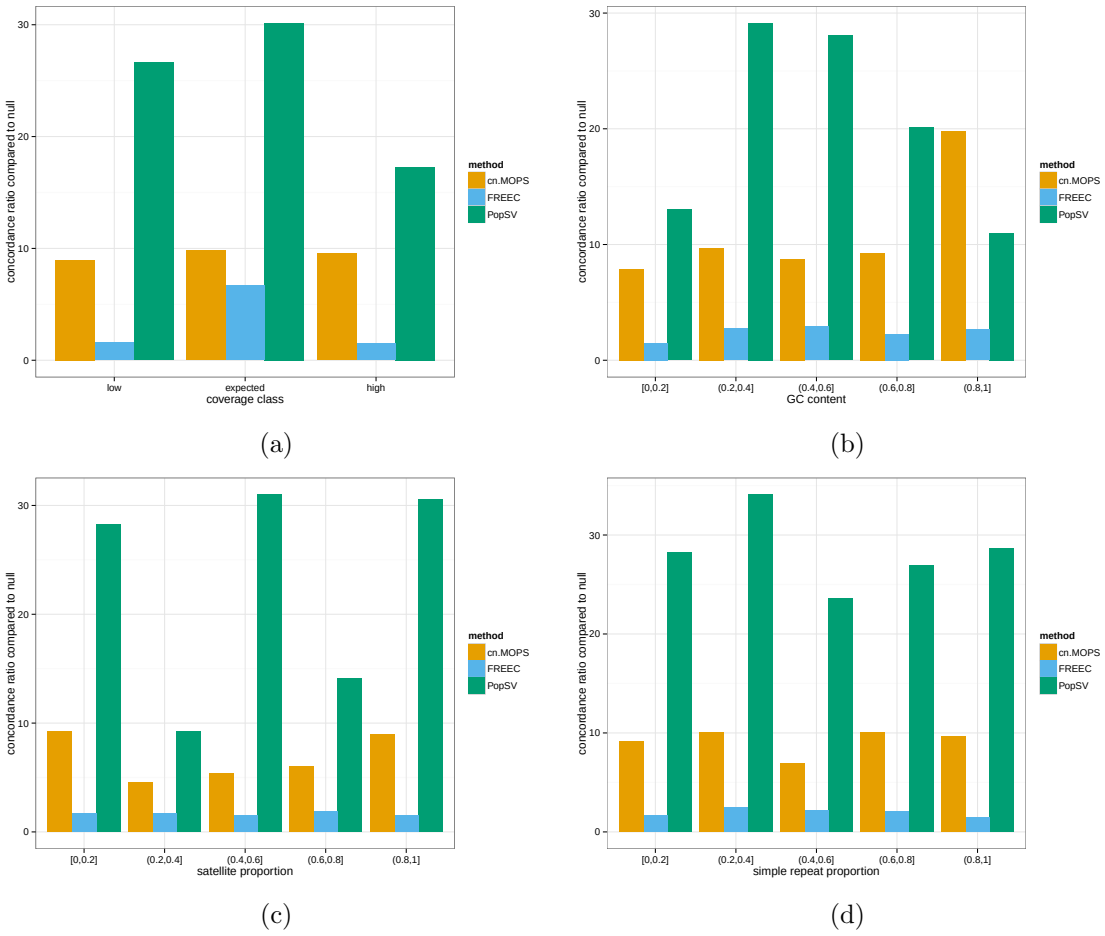
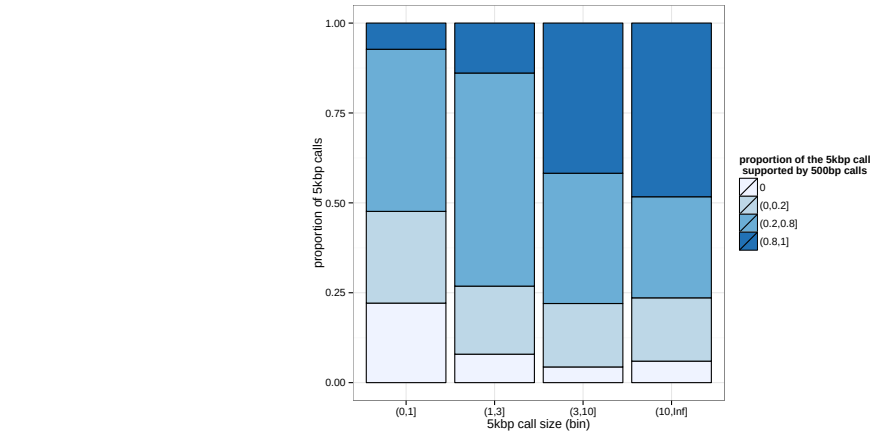
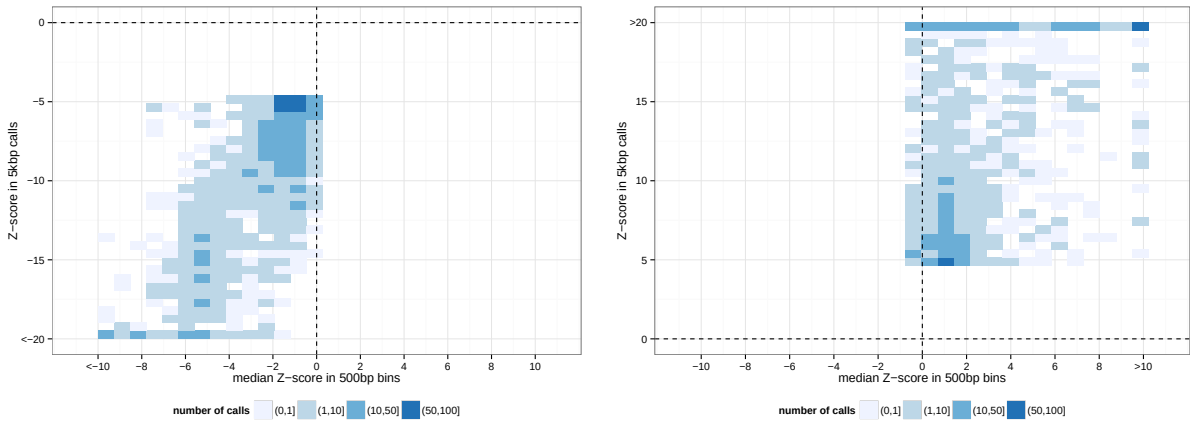


Figure S12: **Twins concordance per bin.** Bins are grouped by coverage class (top-left), GC content (top-right), segmental duplication content (bottom-left) and simple repeat content (bottom-right). The y-axis represents the fold change between the concordance ratio of the calls (see Figure S11) and a null concordance ratio (using random samples).



(a)



(b) (c)

Figure S13: **5Kbp calls supported by 500bp calls.** a) 5 Kbp calls of different sizes (x-axis) are split according to the proportion of the call supported by 500 bp calls. The Z-score of 500 bp bins in 5 Kbp calls is consistent with the call for deletion b) and duplication c) signal. 5 Kbp calls with lower significance (e.g. single-bin calls) are less supported by 500 bp calls (a) but their Z-scores are in the consistent direction (b,c) although not always significant enough to be called.

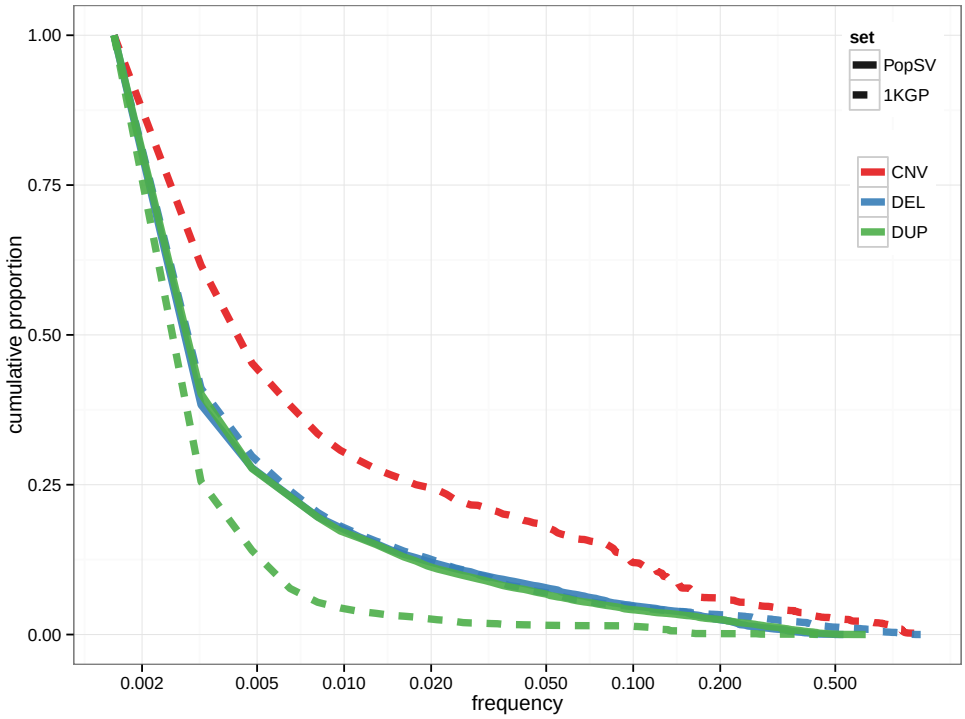


Figure S14: **Frequency of different CNVs.** The cumulative proportion (y-axis) shows the proportion of the affected bp with frequency greater or equal to a particular frequency (x-axis).

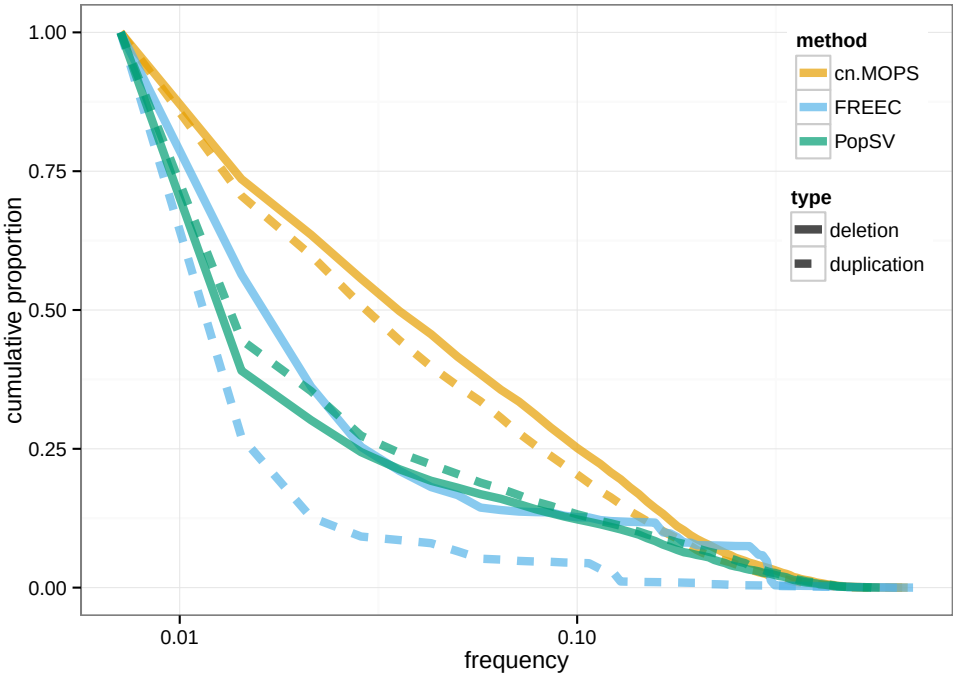
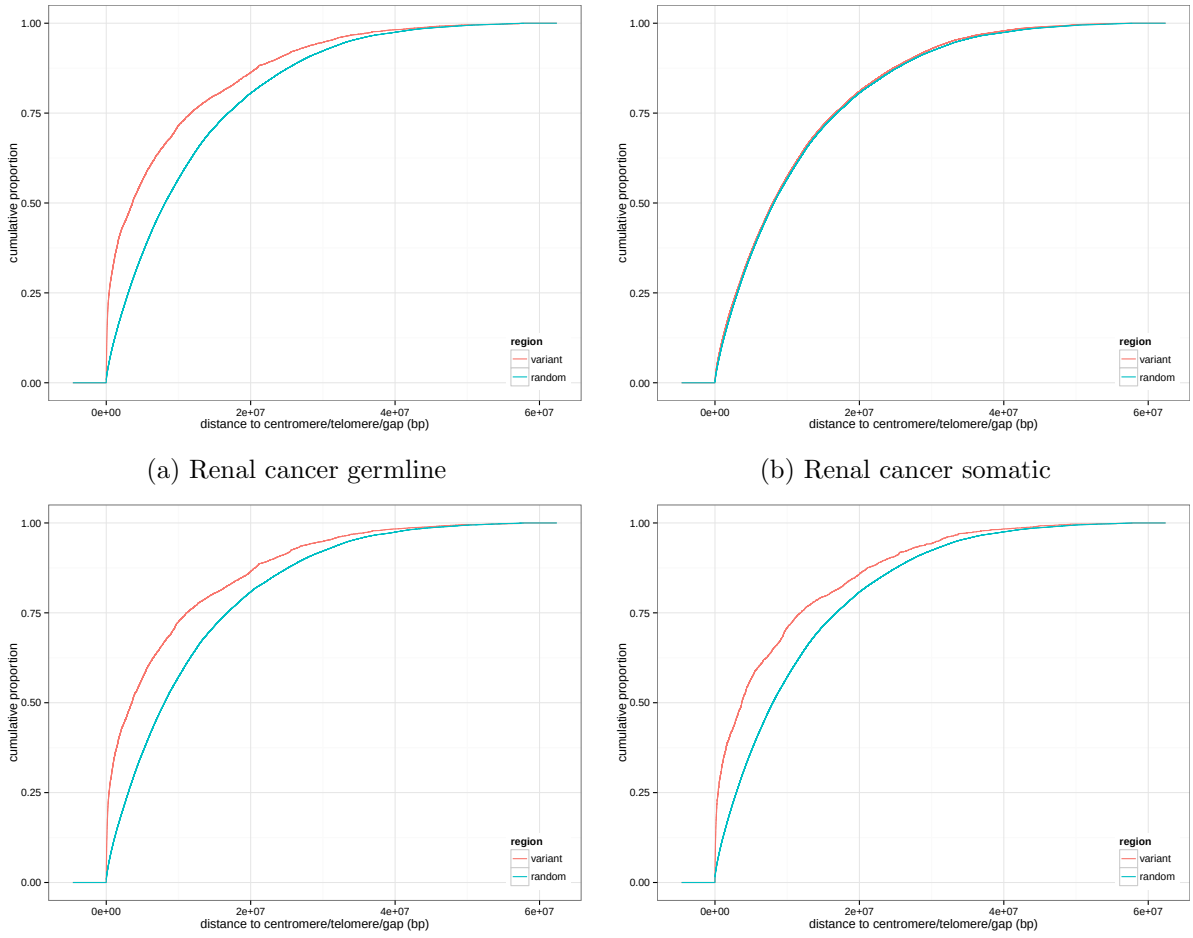


Figure S15: **Variant frequency with different methods.** The x-axis is log-scaled and represents the frequency at which a genomic region is affected by a CNV. The y-axis represent the cumulative proportion of the affected genome.



(a) Renal cancer germline

(b) Renal cancer somatic

(c) Twins

(d) GoNL

Figure S16: **Distance to a centromere, telomere or assembly gap (CTG).** The y-axis represents the cumulative proportion of the affected genome. The *random* curve is computed from uniformly distributed genomic regions with matched size.

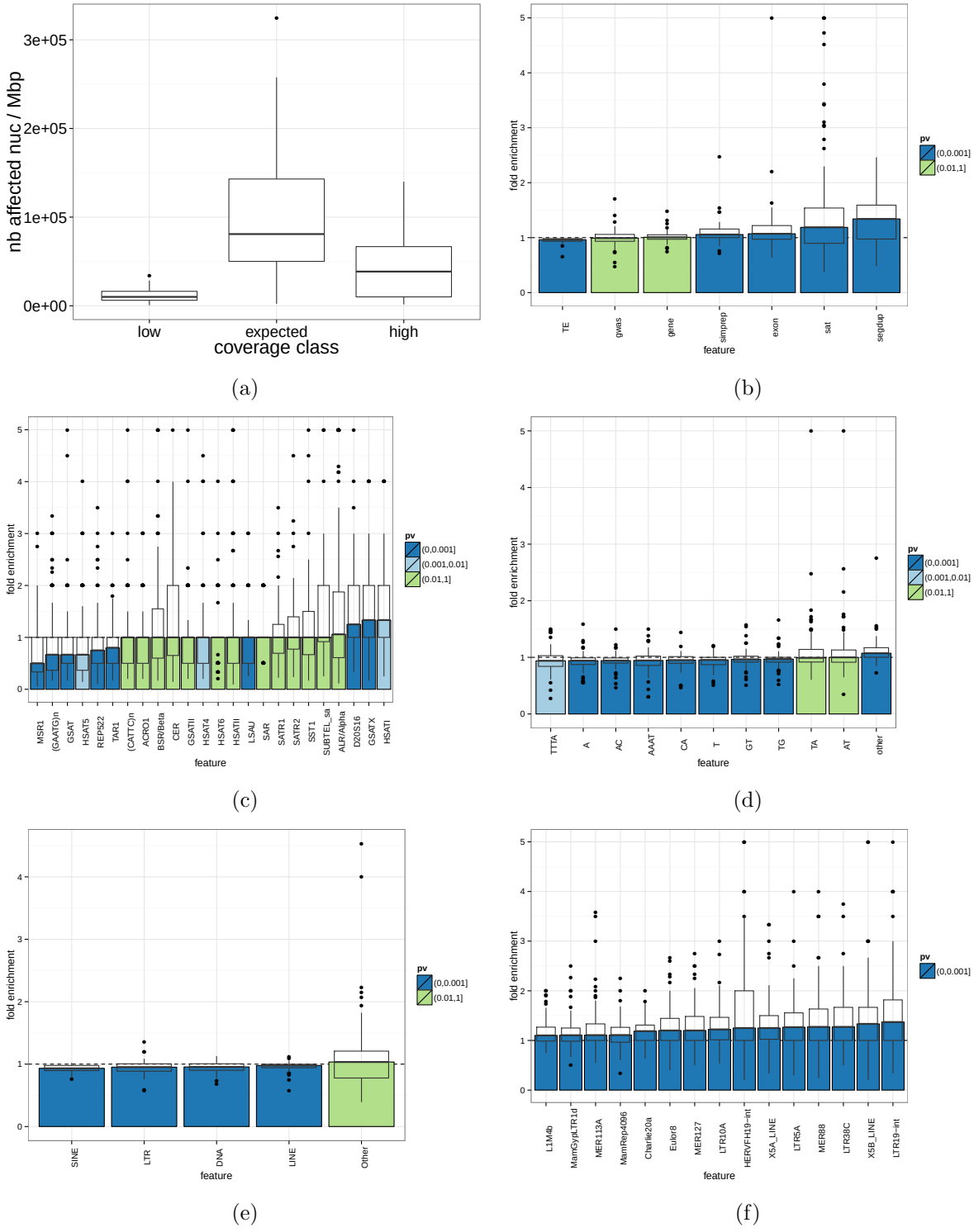


Figure S17: **Somatic CNVs in the renal cancer dataset.** a) The number of nucleotide affected by CNVs per Mbp in each sample, across coverage classes. Enrichment of CNVs in b) different genomic classes, c) satellite families, d) simple repeats, e) TE classes and f) TE sub-families. Bars show the inter-sample median, boxplot shows the variation across samples.

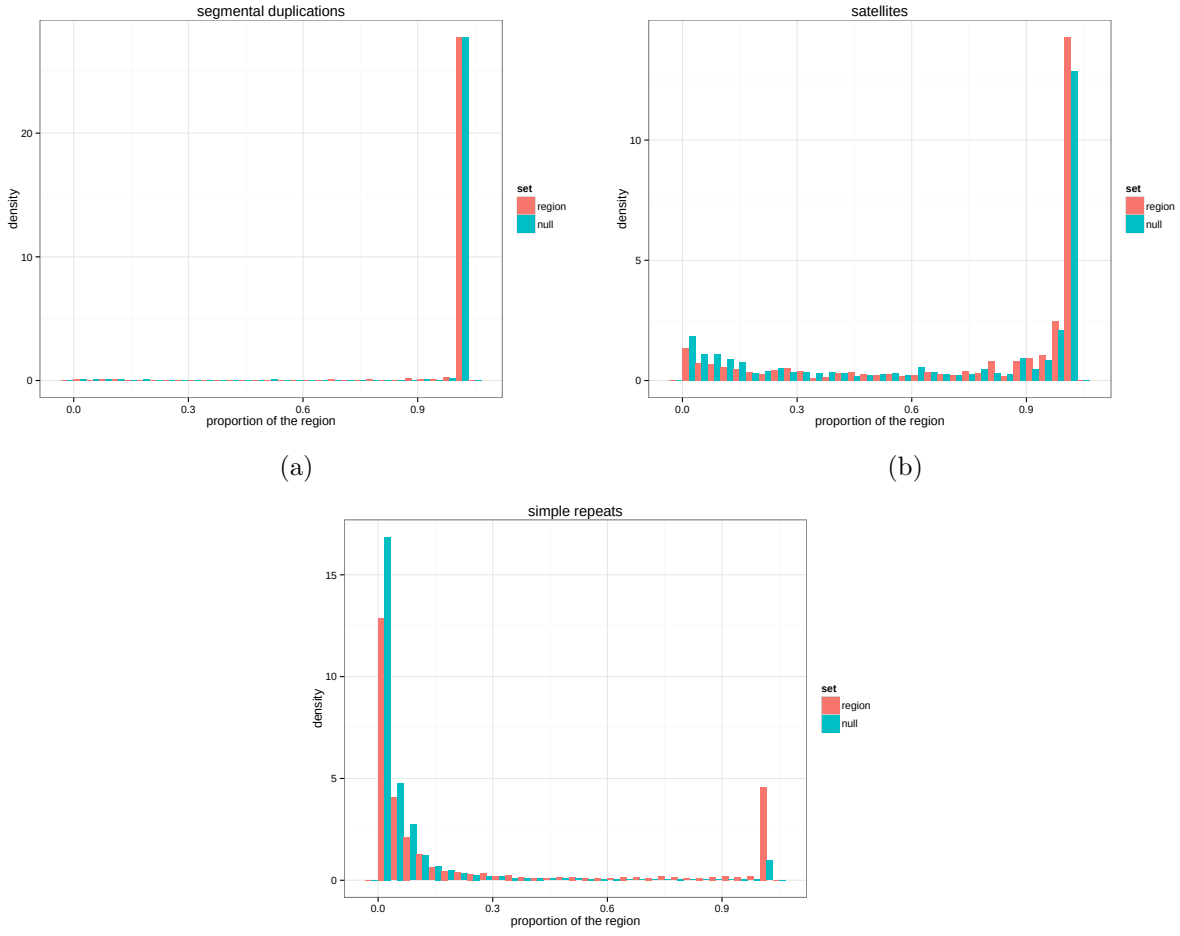


Figure S18: **Overlap between CNVs and repeats.** The histograms represent the proportion of the CNV region that overlaps a) a segmental duplication, b) satellite or c) a simple repeat, when they do overlap. The null distribution is computed from the same control regions used for the enrichment analysis (Methods).

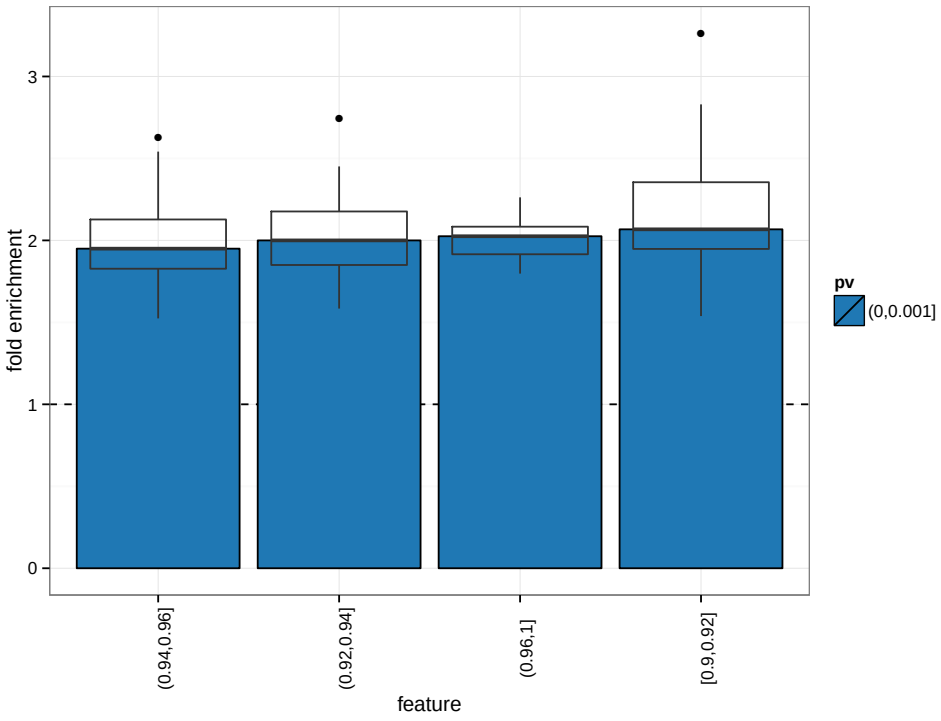
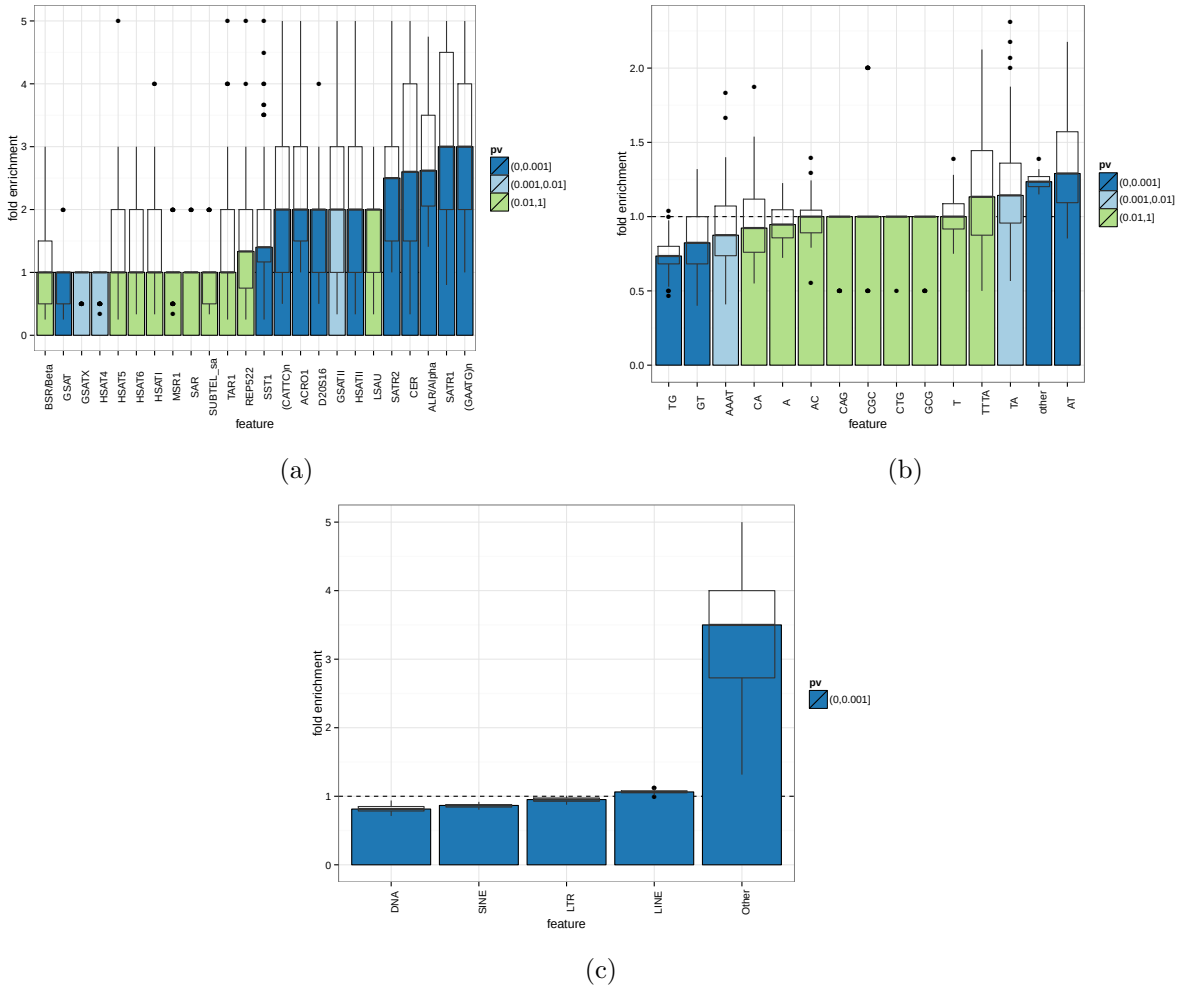


Figure S19: **Enrichment in segmental duplication grouped by age.** Segmental duplication are grouped by similarity to the duplicated region. The fold enrichment between variant and control regions is shown in the y-axis.



(a) (b)

(c)

Figure S20: **CNVs in normal genomes.** Enrichment of CNVs in a) different satellite families, b) simple repeats and c) transposable element classes. Bars show the inter-sample median, boxplot shows the variation across samples.

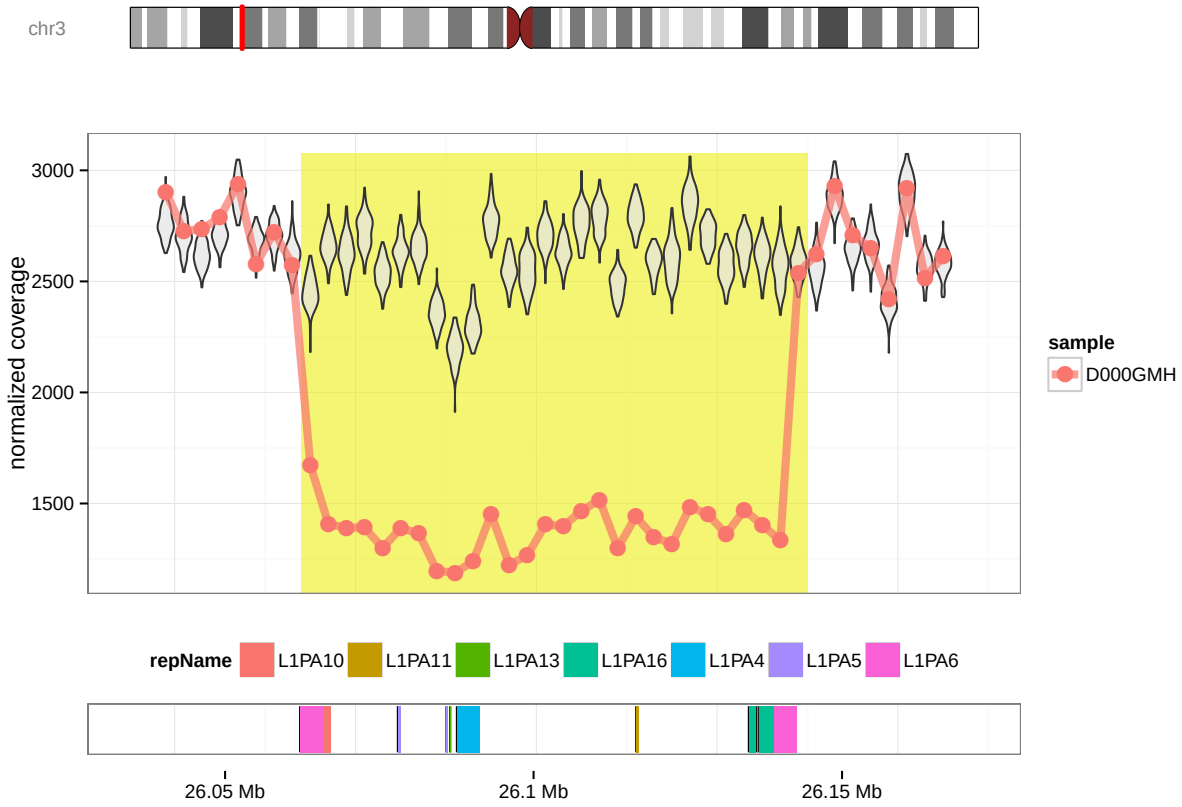
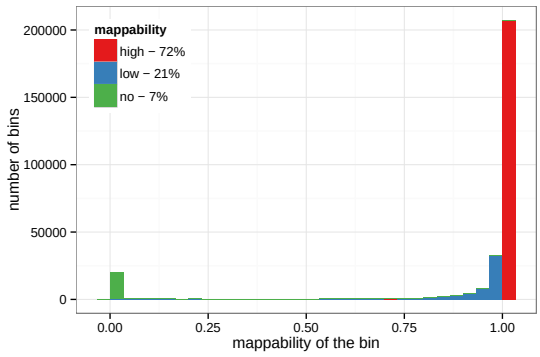
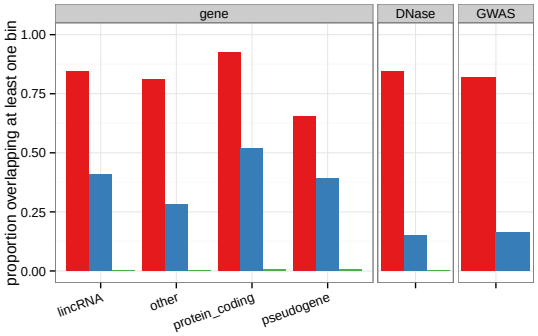


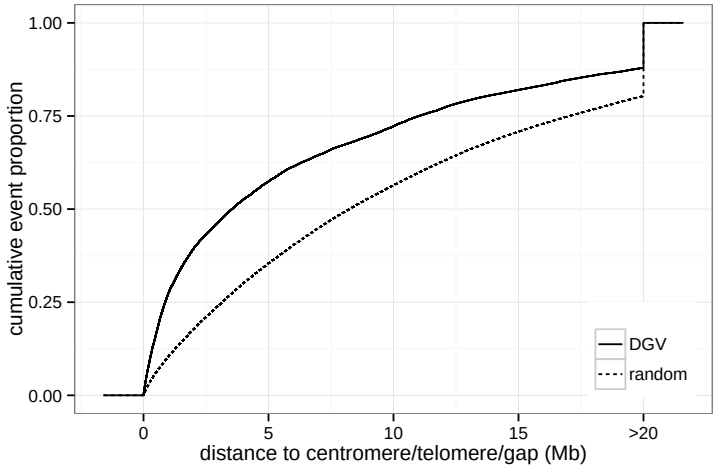
Figure S21: **Polymorphism likely caused by non-homologous allelic recombination between L1PA repeats.** Similar to Fig. 2b, violin plots represent the coverage in the reference samples, and the line and point the coverage in one sample.



(a)



(b)



(c)

Figure S22: **Low mappability regions overlaps protein-coding genes, functional elements and are enriched in structural variants.** a) Mappability distribution of the genome fragmented in 10kb bins. UCSC[49] mappability track is used here (Methods). Three mappability classes are defined. b) Proportion of genes, DNase clusters and GWAS catalog overlapping regions of different mappability. c) Cumulative proportion of DGV annotated regions and distance to centromere/telomere/gap. Only regions annotated in at least 4 different studies were used. The random distribution was computed from regions of similar sizes uniformly distributed across the genome.

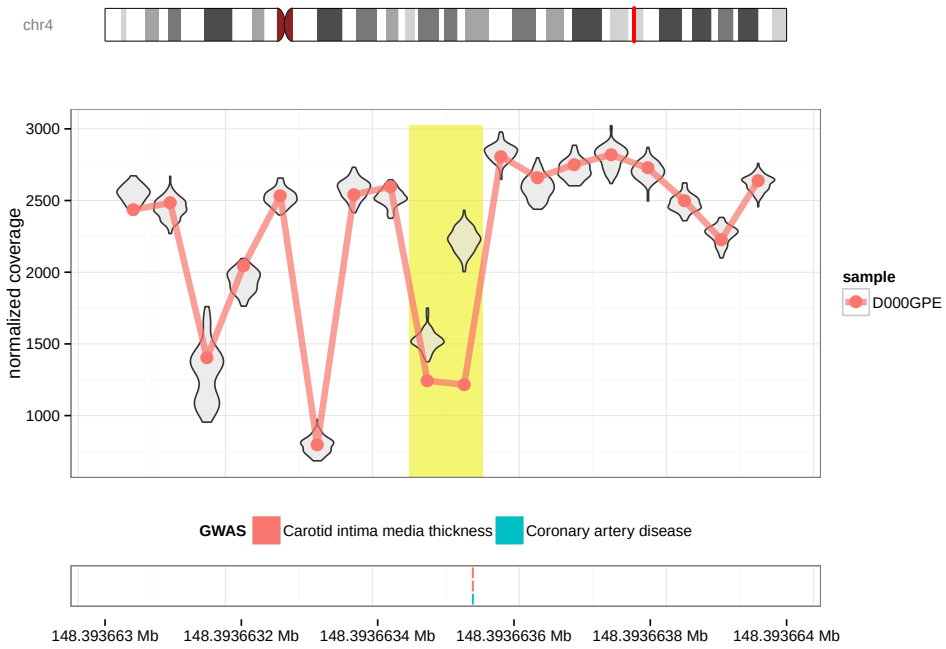


Figure S23: **Example of a CNVs on a GWAS locus.** Example of a germline deletion near a low-coverage region and overlapping a locus associated with coronary artery disease.

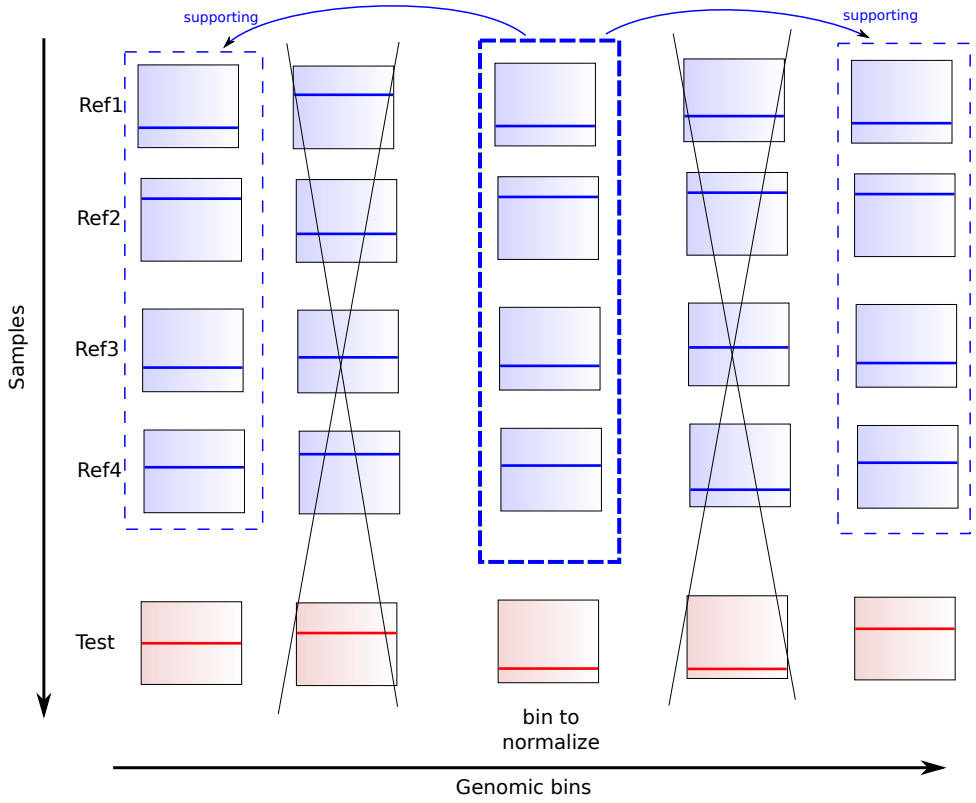


Figure S24: **Targeted normalization.** The coverage across the reference samples (blue) in the bin to normalize is used to find supporting bins across the genome. These supporting bins only are used to compute the normalization factor. The same supporting bins will be used to normalize the bin count in a test sample (red).

8 Methods

8.1 Data

Twin study All patients gave informed consent in written form to participate in the Quebec Study of Newborn Twins[38]. 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[39]. 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.[27]. This data have been derived from different sample collections:

- The LifeLines Cohort Study, 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.
- 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.

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 UCSC database[49] from http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database server. The file names of the corresponding annotations are

Mappability	`wgEncodeCrgMapabilityAlign100mer.bw`
Cytogenetic bands	`cytoBandIdeo.txt.gz`
Centromere, telomere, assembly gap	`gap.txt.gz`
Segmental duplication	`genomicSuperDups.txt.gz`
Simple repeat	`simpleRepeat.txt.gz`
RepeatMasker	`rmsk.txt.gz`
GWAS catalog	`gwasCatalog.txt.gz`

8.2 Technical variation in Read-Depth from Whole-Genome Sequencing

To investigate the bias in RD we first fragmented the genome in non-overlapping bins of 5 Kbp. The number of reads mapping in each bin with a mapping quality higher than 30 (Phred score) was used as RD measure. In each sample, GC bias was corrected by fitting a Loess model between the bin's RD and the bin's GC content. Using this model, the correction factor for each bin was estimated from its GC content. Bins with extreme coverage were identified when deviating from the median coverage by more than 3 standard deviation. After these conventional intra-sample corrections, RD across the different samples were combined and quantile normalized. At that point the different samples had the same global RD distribution and no bins with extreme coverage or GC bias.

Two control RD datasets were constructed to represent our expectation if there was no bias. One was derived from the original RD by shuffling the bins' RD in each sample. In the second, RD was simulated from a Normal distribution with mean and variance fitted to the real distribution. Simulation or shuffling ensures that no region-specific or sample-specific bias remains. To investigate region-specific bias, we computed the mean and standard deviation of the RD in each bin across the different samples. The same was performed in the control datasets. If there was no bias, the distribution of these estimators should be similar in the original, shuffled and simulated RD.

Next, to investigate experiment-specific bias, we retrieved which sample had the highest coverage in each bin. Then we computed, for each sample, the proportion of the genome where it had the highest coverage. If no bias was present, e.g. in the shuffled and simulated datasets, each sample should have the highest coverage in $\frac{100}{N}\%$ of the genome (with N the number of samples). If some experiment are more affected by technical bias it would be more often extreme. The same analysis was performed monitoring lowest coverage.

Finally, the same analyses were repeated with the challenging regions. Instead of excluding any bin with an extreme coverage in a sample, we kept any bin that was extreme in at least one sample. Hence it is the exact complement of the bins kept previously, i.e. all the bins previously removed by this filter.

8.3 PopSV a population-based approach

Binning and coverage measure The genome is fragmented in non-overlapping consecutive bins of fixed size. We ran three 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 finally ran GoNL analysis using 2 Kbp and 5 Kbp bins.

In each bin and each sample the number of reads that overlap the bin and are properly mapped are counted to get a measure of coverage. Here proper mapping means read pairs with correct orientation and insert size, and a mapping quality of 30 (Phred score) or more. The bin counts were then corrected for GC bias. In each sample, a LOESS model was fitted between the bin's count and bin's GC content. A normalization factor was then defined for each bin from its GC content.

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 Twins dataset 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 globally normalized. The resulting first two principal components are used to verify the homogeneity of the reference samples. In the presence of extreme outliers or clear sub-groups, a more cautious analysis is recommended. For example, outliers can stay included in the set of reference samples keeping in mind it might harbor more false calls later. Independent analysis in each sub-group identified is also a solution, especially when all the samples are to be used as reference. 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 the 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 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.

Normalization Although uniformity of the coverage across the genome is not required for our approach, RD values must be comparable across samples. When a particular region of the genome is tested, sample specific variation of technical origin must be minimized. This is done through a normalization step.

Naive global normalization approaches like the Trimmed-Mean M(TMM) or quantile normalization have been first implemented and tested. The TMM normalization robustly aligns the mean RD value in the samples. Quantile normalization forces the RD distribution to be similar in each samples. After witnessing the presence of un-characterized sample-specific variation, we implemented a more suited normalization.

Targeted normalization uses information across the large set of samples to identify similar bins across the genome and normalize their counts separately (see Fig. S24). For each bin, the top 1000 bins with similar coverage patterns across the reference samples are used to normalize the coverage

of the bin. TMM normalization is used on these top 1000 bins to derive the correct normalization factor for the bin to normalize. Similarity between two bins is measured using Pearson correlation between the counts across the reference samples. Hence the top 1000 bins are most similar in term of relative coverage across the samples to the coverage in the bin to normalize. If some bias is present in some samples, the top 1000 bins should also harbor this bias. This is a way to select other regions with similar bias patterns in order to correct it. In this targeted approach, each genomic region is normalized independently. The 1000 supporting bins are saved and used to normalized new samples (e.g. case sample). Although computationally expensive it ensures that all bins are normalized with the same effort. In contrast global normalization or even PCA-based approaches corrects for the most common or spread bias.

In order to compare the performance of the different normalization approaches we computed a set of quality metrics. The normalized RD will need to be suited for testing abnormal pattern across samples: under the null hypothesis, i.e. for normal bins, the RD should be relatively normally distributed and the samples rank should vary randomly from one bin to the other. The first metric is the *proportion of bins with non-normal RD* across the samples. Shapiro test was performed on each bin and a P-value lower than 0.01 defined non-normal RD. Then the randomness of the sample ranks was tested by comparing the RD of each sample a region with the median across all samples. In a regions of 100 consecutive bins, we counted how many times the RD in a sample was higher than the median across sample. If the rank are random this value should be around 0.5. The probability under the Binomial distribution is computed for each sample and corrected for multiple testing using Bonferroni correction. If any sample has an adjusted P-value lower than 0.05 we consider that the bin has non-random ranks. The resulting QC metric is simply the *proportion of bins with non-random sample ranks*. This QC is specifically testing how much sample-specific bias remains. The remaining QC metrics look at the Z-score distribution in each sample. The proportion of non-normal Z-scores is computed by comparing the density curves of the Z-scores and simulated Normal Z-scores. We compute the proportion of the area under the density curve that doesn't overlap the Normal density curve. This estimate of the *proportion of non-normal Z-scores* is computed in each sample. The final metrics are the *average and maximum* across the samples.

Abnormal RD test and Z-score computation The test is based on Z-scores computed for each bin, corrected afterward for multiple testing. The Z-scores represents how different the read count in the tested sample is from the reference samples. It is simply:

$$z = \frac{BC_t^b - \overline{BC_{ref}^b}}{sd(BC_{ref}^b)}$$

where BC_t^b is the bin count, i.e. the number of reads, in bin b and sample t.

Inevitably some samples are hosting common CNVs. We observed that just a couple of samples hosting a CNVs could be enough to bias the standard deviation used in the score computation and mask these CNVs in the coming tests. In many cases the RD signal was clearly showing several groups of samples with proportional read counts. To improve the Z-score computation in those regions, a simple approach was used: the samples were stringently clustered using their RD and the group with higher number of samples was chosen as reference and used to compute the mean and standard deviation for the Z-score computation. In practice, this clustering affects only bins with clear clusters but would remove just a few or no samples in most situations. Furthermore, a

median-based estimator was used for the standard deviation as it is less sensitive to outlier removal. A trimmed mean was also preferred over normal mean for its robustness to outliers.

Significance and multiple testing correction The Z-scores for all the bins of a sample are pooled and significance is estimated. Under the null hypothesis of normally distributed read counts, the Z-scores should also follow a normal distribution. For multiple testing correction, the Z-score empirical distribution is used to fit a normal and estimate the P-value and Q-value of each test. This step is performed using fdrtool R package.

By default the null distribution fitting for P-value computation assumes only a low proportion of bins violates the null hypothesis. In aberrant genomes, e.g. in tumor samples, it is often an unrealistic assumption. We devised a new strategy to set the proportion of the empirical distribution used to estimate the null distribution variance. Here the null Z-score distribution is assumed to be centered on 0 and only its variance is estimated by trimming the tails of the empirical distribution. To find a correct trimming factor, an iterative approach started from a low trimming factor and increased its value until reaching a plateau for the variance estimator. Once the plateau is reached, additional trimming doesn't change the estimated variance because there is no more abnormal Z-scores, only the central part of the null distribution. Samples with an important proportion of abnormal genome, e.g. tumor samples, showed more appropriate fit.

Of note the P-values for positive Z-scores (duplication) and negative Z-scores (deletion) are estimated separately. Thus imbalance in the deletion to duplication ratio, or large aberration that lead to asymmetrical Z-score distribution doesn't affect the P-value estimation. Multiple testing correction is still performed on all the P-values.

Copy number estimation and other metrics Following the significance estimation, consecutive bins with abnormal coverage are merged into a call. In addition to the Z-score, P-value, Q-value and number of bins of each call, PopSV retrieves the average coverage in the reference samples and the fold change in the sample tested.

Copy number is also estimated by dividing the coverage in a region by the average coverage across the reference samples, multiplied by 2 (as diploidy is expected). In our bin setting, the estimation is correct if the bin spans completely the variant. For this reason we trust the copy number estimate for calls spanning 3 or more consecutive bins, as it is computed using the middle bin(s) which completely span the variant. In other cases we expect the copy number estimate to under-estimated.

All this additional information can be used to order or retrieve high confidence calls. For examples, several consecutive bins or a copy number estimate around an integer value increases our confidence in a call. In our validation and analysis however, we used the entire set of calls.

The ZZ plots are computed directly from the Z-score of each bin in two different samples (e.g. paired normal/tumor samples, twins). The global distribution of the Z-score is also compared to the mappability estimate of the bins. At this point, we use the mappability track available from UCSC[49] (see Genomic annotations) and compute the mean level across the bin.

Coverage tracks For each run, we computed 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 higher or lower than 4 standard deviation

from the median were classified as *high coverage* or *low coverage* respectively. To ensure robustness, the standard deviation was derived from the Median Absolute Deviation.

Eventually, we also defined *extremely low coverage* region which have an average coverage close to 0. These region are defined by the peaks around 0 in the distribution of average coverage (see Figure S8). This sub-class of *low coverage* region is used in a few of the following analysis to highlight the most challenging regions.

8.4 Validation

Running FREEC and cn.MOPS FREEC was run on each sample separately, starting from the BAM file. 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.

cn.MOPS 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.

Clustering the Twins samples A distance between two sample A and B is defined as : $1 - 2\frac{|V_A \cap V_B|}{|V_A|+|V_B|}$ where V_A represents the variants found in sample A, $V_A \cap V_B$ the variants found in both A and B, and $|V|$ the cumulative size of the variants. 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. Different linkage criteria (*average*, *complete* and *Ward*) were used for the exploration. In our dendograms we used the *average* linkage criterion. The same clustering was performed using only calls in regions with extremely low coverage (reference average $\leq$10 reads).

Frequency peak in Twins The frequency at which a region is affected by a CNV was compared between the different methods. In the Twins dataset, we expect a peak around frequency of 3 samples : the two twins and one parent. To compare the different methods the height of the peak, in the frequency distribution, represent the proportion of the affected genome called at each frequency.

Concordance between two twins For each twin, a CNV call was defined as *concordant* if also found in the other twin. In order to avoid missing calls with borderline significance, we used slightly less confident calls for the second twin. We removed calls present in more than 50% of the samples as they could be systematic errors that would look concordant. Hence a *concordant* call is most likely true as it is present in a minority of samples but consistently in the twin pair. The proportion of *concordant* calls per sample gives an estimate of specificity. The level of sensitivity is represented by the number of *concordant* calls. Even if we removed systematic calls, the most frequent calls in the cohort are more likely to look *concordant* by chance. To normalize for this effect, we use the frequency distribution to compute the number of *concordant* calls expected by chance. In practice the null concordance for each call is simulated by a Bernoulli distribution of

parameter the frequency of the call. This number of *concordant* calls by chance is subtracted to the original number of *concordant* calls to give a adjusted measure of sensitivity. Although we don't know the true number of variant, this number of *concordant* calls is used to compare the different methods. The same analysis is also performed using only calls located in *low coverage* (as defined by the coverage track) regions in order to get an estimate on challenging regions. A call was considered in a *low coverage* region if more than 90% of its sequence was annotated as so.

In addition to this *per-sample* concordance, we compute a *per-region* concordance estimate by pooling all the calls from the samples. When most of the normal/tumor pairs are consistently called in a specific bin, it is classified as *concordant*. Then the bins can be grouped according to their GC content or repeat content to test that the quality of the calls is similar. This approach is particularly useful to verify that the proportion *concordant* bins is similar in bin extreme GC content or repeat content. Finally we compute a null distribution with same approach but using randomly selected samples instead of the sample called in each bin. Dividing the proportion of *concordant* bin by its null equivalent gives an idea of the significance of the observation. This fold-change from the null is used to compare the different methods. In addition, we use the *per-region* concordance to estimate the amount of the genome that can be correctly tested. Here the genome is fragmented in 1 Mbp windows and we count how many show more *concordant* regions than *concordant* by chance. The 1 Mbp fragmentation is used in order to avoid biases from segmentation behavior. If the regions were used as-is, a segmentation that tend to locally call longer segments will look largely superior even-though it calls the same variants. The fragmentation of the genome in large windows limits this bias and allows for fair comparison between the different methods. By counting how many 1 Mbp windows can be called correctly, we estimate how much of the genome can be correctly tested by each method.

Concordance between paired normal and tumor samples The same approach as described previously when comparing pairs of twins was applied in the renal cancer dataset, on pairs of normal/tumor samples. Here true germline calls should be also found in the paired tumor sample, and concordance is computed for each normal sample. Again both *per-sample* and *per-region* estimates are computed and compared between methods.

Concordance between different bin sizes We compared the calls using small bins (500 bp) and the calls using larger bins (5 Kbp). First we counted how many small bin calls supported any large bin call. These metrics were separated according to the size of the large bin call. To investigate large bin calls with no supporting small bin call, we display the average Z-scores in the small bins overlapping large bin calls. This is useful to test if the lack of support is due to lower confidence or real discordance between the different runs. If the Z-scores in the small bins deviates from 0 in the correct direction we conclude that they support the large bin call.

Then we checked which of the small bin call overlapped large bin calls. More specifically, we grouped them by size to verify that large enough small bin calls are present in the large bin calls. This analysis is used to both test the sensitivity of PopSV with a particular bin size, and its resolution to variants smaller than the bin size. Indeed this framework allow us to ask questions such as: how much of the variants spanning only half a bin are detected ?

Experimental validation The 23 variants chosen for experimental validation were randomly selected among both one-copy and two-copy deletions. We selected both small ($\sim$ 700 bp) and

large ($\sim$ 4 Kbp) variants in each class. In addition, 3 deletions in low-mappability regions were also randomly selected and included. The coverage at base pair resolution was visually inspected for each deletion in order to map the breakpoints. PCR primers were designed to target the whole deletion region. We performed long-range PCR followed by gel electrophoresis. We then compared 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 and could lead to non-validation when the variant is actually present.

8.5 Genomic patterns of CNVs

Merging results using two different bin sizes Small bins gives better resolution for smaller variant. 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 to define breakpoints and other information (e.g. copy number estimate), the proportion of overlap was used. If call(s) from the small bin run 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 retrieve 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.

Computing global estimates of copy number variation In Table 1, a call in low coverage region is completely located within *extremely low coverage* regions (as defined by our coverage tracks). The amount of sequence affected in a genome is computed by merging all the variants (e.g. if several samples are combined) and counting the number of bases in this merged set. After the merging step, each base of the genome either overlapped a merged variant or not. Hence each affected base is counted only once, even if it overlaps CNVs in several samples, or with large copy number differences.

Comparing with 1000 Genomes SV set The SV catalog from[26] was downloaded and parsed into our preferred BED-like format. We first checked that we could reproduce the numbers in the main SV paper. Then we retrieved the set of autosomal deletion, duplication and CNVs. We then removed deletions smaller than 300 bp as well as variants with high frequency ($> 80\%$). This sub-set of SV 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.

Distance to centromere, telomere and assembly gaps The centromeres, telomeres and assembly gaps (CTGs) are annotated in the gap track from UCSC[49]. 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, meaning 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. Thanks to this null distribution we are able to see if variants are closer/farther to CTG than we would expect by chance.

Variation rate computation The variation rate represent how much a sample varies from the reference genome. The first estimate is the average number of variant per Mb, the second is the average number of affected nucleotide per Mb. These estimates are computed for each samples. Afterwards the distribution across samples is visualized.

We computed these estimates separately for regions of *low coverage*, *expected coverage* and *high coverage* (as defined by our coverage tracks). These three classes of regions might have different false positive rate (FPR). In order to be sure that the differences in variation rate are not due to FPR differences, we adjust our estimates. We use the estimates from the *per-region* validation and sub-sampled the calls in each class with the corresponding proportion. Sub-sampling the variants and recomputing the variation rate was preferred over a crude multiplication of the original variation rate by the FPR.

Simulating control regions Control regions are simulated to have the same size distribution and same overlap with specified genomic features. In practice, this was used to control for the distance to centromere, telomeres and assembly gaps, as well as the overlap with regions of low mappability. Hence the patterns observed afterwards are not caused by the over-representation of region in low/high coverage regions or their proximity to CTGs. To simply control for the distance to a genomic feature, we created new annotation with regions flanking the regions in question and control for the overlap with these.

First, thousands of bases are randomly chosen in the genome. The distance between each base and the genomic features is then computed. At this point, simulating a region of a specific size and with specific overlap profile can be done by randomly choosing as center one of the bases that fit the profile :

$$\{b, \forall\ feature\ f, O_f(d_f^b - \frac{S_r}{2}) < 0\}$$

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

Hence for each input region, a control region would be selected as described. In practice, the input regions are first grouped by overlap profile (i.e. O_f) and then split up according to their size (e.g. 30 different size classes). Instead of using the exact desired size S_r, we now use the median size of the size class and simulate regions in chunk, which speeds up the computations while providing satisfactory results.

The number of random bases is important : a low number might result in duplicated regions in the output, but a high number is more computationally demanding. In practice, we perform the simulation twice with 10^5 random bases. The second run is used to simulate again all the duplicated regions from the first run.

Of note, when desired we control for the distance to CTG using this approach. We actually correct for the overlap with CTGs and regions flanking CTGs. We used CTGs, 500 Kbp regions flanking CTGs and 3 Mbp regions flanking CTGs. These gave satisfactory distribution of the distance to the nearest CTG in the control regions.

Enrichment in genomic features Regions of interest are overlapped with genomic features. We then compute the proportion of regions overlapping each feature. The same is done with control regions, constructed from the regions of interest (see previous paragraph).

If sample information is available (e.g. when analyzing variants), the proportion of overlap and the control regions are computed separately for each sample. Hence the control region fits perfectly the profile of the variants in each sample and is not simply a reflection of the majority of the samples. For each sample (and each feature), the enrichment measure is the difference between the proportion in the original and control regions. A Wilcoxon test on this measure assesses how significant is the potential deviation from 0. The fold-enrichment is the ratio between overlap proportion between original and control regions.

Eventually, we display how much of a variant overlap a feature of interest. This distribution is useful to get a sense if the genomic feature overlap completely the variants or just a small fraction of it.

Somatic variant definition Somatic variants were defined as variant in a tumor samples with no or 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. Here a tumor variant was defined as somatic if it overlapped less than 10% of any variant in the paired normal.

Frequency distribution The frequency at which a region is affected is computed using calls from our 640 samples. The cumulative proportion of affected genome is shown for each frequency in the frequency curve. In addition, frequency curves are computed using small or large variants, exonic or non-exonic variants, and deletions or duplications.

Eventually, we perform the same analysis with the set of comparable CNVs extracted from the 1000 Genomes catalog. Of note, the CNV set was down-sampled to 640 random samples in order to give comparable frequency curves.

CNV impact Exons of protein-coding genes and promoter regions (10 Kbp upstream of the transcription start site) were extracted from the Gencode annotation v19. We counted how many different genes had their exons, exons + promoter and exon + promoter + introns hit by a CNV, in a sample or in the entire dataset.

The number of genes hit by at least one of the CNV in at least one sample increases with sample size. We performed a saturation analysis by down-sampling our dataset to lower sample size. For each down-sampled set the number of genes hit is computed. Afterwards the number of affected genes is plotted against the sample size. A plateau at higher sample size would mean that a set of CNV-tolerant genes has been almost completely discovered.

Finally, we also computed the number of GWAS hits overlapping a CNV, per sample or in the entire dataset.