Genome-wide characterization of copy number variants in epilepsy patients

Jean Monlong[1,2,10], Simon L. Girard[1,3,4,10], Caroline Meloche[4], Maxime Cadieux-Dion[4], Danielle M. Andrade[5], Ron G. Lafreniere[4], Micheline Gravel[4], Dan Spiegelman[6], Alexandre Dionne-Laporte[6], Cyrus Boelman[7], Fadi Hamdan[8], Jacques L. Michaud[8], Guy Rouleau[6], Berge A. Minassian[7], Guillaume Bourque[1,2,9,*], and Patrick Cossette[4,*]

[1]Department of Human Genetics, McGill University, Montréal, H3A 1B1, Canada

[2]Canadian Center for Computational Genomics, Montréal, H3A 1A4, Canada

[3]Departement des sciences fondamentales, Université du Québec à Chicoutimi, Chicoutimi, G7H 2B1, Canada

[4]Centre de Recherche du Centre Hospitalier de l'Universite de Montréal, Montréal, H2X 0A9, Canada.

[5]Epilepsy Genetics Program, Division of Neurology, Toronto Western Hospital, University of Toronto, Toronto, Canada.

[6]Montréal Neurological Institute, McGill University, Montréal, H3A 2B4, Canada.

[7]Division of Neurology, The Hospital for Sick Children, Toronto, M5G 1X8, Canada.

[8]CHU Sainte-Justine Research Center, Montréal, H3T 1C5, Canada.

[9]McGill University and Génome Québec Innovation Center, Montréal, H3A 1A4, Canada

[10]These authors contributed equally to this work

[*]Correspondence: guil.bourque@mcgill.ca (GB), patrick.cossette@umontreal.ca (PC)

October 6, 2017

Abstract

Epilepsy will affect nearly 3% of people at some point during their lifetime. Previous copy number variants (CNVs) studies of epilepsy have used array-based technology and were restricted to the detection of large or exonic events. In contrast, whole-genome sequencing (WGS) has the potential to more comprehensively profile CNVs but existing analytic methods suffer from limited sensitivity and specificity. We show that this is in part due to the non-uniformity of read coverage, even after intra-sample normalization. To improve on this, we developed PopSV, an algorithm that uses multiple samples to control for technical variation and enables the robust detection of CNVs. Using WGS and PopSV, we performed a comprehensive characterization of CNVs in 198 individuals affected with epilepsy and 301 controls. We found an enrichment of rare exonic events in epilepsy patients, especially in genes with predicted loss-of-function intolerance. Notably, this genome-wide survey also revealed an enrichment of rare non-coding CNVs near previously known epilepsy genes. This enrichment was strongest for non-coding CNVs located within 100 Kbp of an epilepsy gene and in regions associated with changes in the gene expression, such as expression QTLs or DNase I hypersensitive sites. Finally, we report on 21 potentially damaging events that could be associated with known or new candidate epilepsy genes. Our results suggest that comprehensive profiling of CNVs could help explain a larger fraction of epilepsy cases.

1 Introduction

Structural variants (SVs) are defined as genetic mutations affecting more than 50 base pairs and encompass several types of rearrangements: deletion, duplication, novel insertion, inversion and translocation. Deletions and duplications, which affect DNA copy number, are 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]. Numerous diseases are caused by SVs with a demonstrated detrimental effect[4,5]. While cytogenetic approaches and array-based technologies have been used to identify large SVs, whole-genome sequencing (WGS) has the potential to uncover the full range of SVs both in terms of type and size[6,7]. SV detection methods that use read-pair and split read information[8] can detect deletions and duplications but most CNV-focused approaches look for an increased or decreased read coverage, the expected consequence of a dupli-

cation or a deletion. Coverage-based methods exist to analyze single samples[9], pairs of samples[10] or multiple samples[11,12,13] but the presence of technical bias in WGS remains an important challenge. Indeed, various features of sequencing experiments, such as mappability[14,15], GC content[16] or replication timing[17], have a negative impact on the uniformity of the read coverage[18].

Epilepsy is a common neurological disorder characterized by recurrent and unprovoked seizures. It is estimated that up to 3% of the population will suffer from a form of epilepsy at some point during their lifetime. Although the disease presents a strong genetic component that can be as high as 95%, typical "monogenic" epilepsy is rare, accounting for only a fraction of cases[19,20]. For the last thirty years, many teams have focused on finding genetic susceptibility factors associated with the disease. Genome-wide association studies have had only moderate success and revealed only a few marginal associations[21,22]. A meta-analysis combining multiple epilepsy cohorts did find positive associations with the disease, the strongest in SCN1A, a gene already associated with the genetic mechanism of the disease[23]. Thanks to array-based technologies, surveys of large CNVs (>50 Kbp) have shown the importance of large and *de novo* CNVs as well as identified a few associations with specific genes[24,25,26,27,28]. Rare genic CNVs were typically found in around 10% of epilepsy patients[25,29,28] and CNVs larger than 1 Mbp were significantly enriched in patients compared to controls[30,29,31]. Unfortunately, small CNVs and other types of SVs could not be efficiently or consistently detected using these technologies, hence much remains to be done.

To more comprehensively characterize the role of CNVs in epilepsy, we performed whole-genome sequencing of epileptic patients from the Canadian Epilepsy Network (CENet), the largest WGS study on epilepsy to date. In the present study, we assessed the frequency of CNVs in epileptic individuals using 198 unrelated patients and 301 healthy individuals. Using this data, we showed that technical variation in WGS remains problematic for CNV detection despite state-of-the-art intra-sample normalization. To correct for this and to maximize the potential of the CENet cohorts, we developed a population-based CNV detection algorithm called PopSV. Our method uses information across samples to avoid systematic biases and to more precisely detect regions with abnormal coverage. Using two public WGS cohorts[32 33], and additional orthogonal validation, we showed that PopSV outperforms other analytical methods both in terms of specificity and sensitivity, especially for small CNVs. Using this tool, we built a comprehensive catalog of CNVs in the

CENet epilepsy patients and studied the properties of these potentially damaging structural events across the genome.

2 Material and Methods

Epilepsy patients and sequencing Patients were recruited through two main recruitment sites at the Centre Hospitalier Universitaire de Montréal (CHUM) and the Sick Kids Hospital in Toronto as part of the Canadian Epilepsy Network (CENet). The ethics committees of both sites approved this study. Before their inclusion in this study, patients or parents (when needed) had to give written informed consents. The main cohort of this study was constituted of 198 unrelated patients with various types of epilepsy; 85 males and 113 females. The mean age at onset of the disease for our cohort was 9.2 ($\pm$6.7) years. Supplementary Table S1 presents a detailed description of the clinical features for the various individuals recruited in this study. 301 unrelated healthy parents of other probands from CENet were also included in this study and used as a control cohort. DNA was exclusively extracted from blood DNA.

Libraries were generated using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina) and paired-end reads of size 125 bp were sequenced on a HiSeq 2500 to an average coverage of 37.6x $\pm$ 5.6x. Reads were aligned to reference Homo_sapiens b37, aligned with BWA[34]. Finally, Picard was used to merge, realign and mark duplicate reads. For more details, see Supplementary Information.

Testing for technical biases in WGS To investigate the bias in read depth (RD), we fragmented the genome in non-overlapping bins of 5 Kbp and counted the number of properly mapped reads. In each sample, we corrected for GC bias and removed bins with extremely low or high coverage (see Supplementary Information). Then, read counts across all samples were combined and quantile-normalized. Using simulations and permutations, we constructed two control RD datasets with no region-specific or sample-specific bias. We computed the mean and standard deviation of the coverage in each bin across samples. 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. The same analysis was performed monitoring the lowest coverage.

PopSV 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. The genome is first segmented into bins and the number of reads with proper mapping in each bin was counted for each sample. In a typical design, the genome is segmented in non-overlapping consecutive windows of equal size, but custom designs could also be used. 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 bins to normalize read coverage (see Supplementary Information). Our targeted normalization was compared to global approaches that adjust for the median coverage, or quantile-based approaches. After normalization, the value observed in each bin is compared with the profiles observed in the reference samples and a Z-score is calculated (Figure 1b). 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. Consecutive or nearby abnormal bins are merged and considered as one variant.

Validation and benchmark of PopSV We compared PopSV to FREEC[10], CNVnator[9] and cn.MOPS[11], three popular RD methods that can be applied to WGS datasets, to identify CNVs using a publicly available dataset from a Twin study[32]. We also ran LUMPY[8] which uses an orthogonal mapping signal: the insert size, orientation and split mapping of paired reads. For LUMPY, all the CNVs (deletions and duplications) and intra-chromosomal translocations (labeled as 'BND' in Lumpy's output) larger than 300 bp were kept for the upcoming analysis. First, we compared the frequency at which a region is affected by a CNV using the calls from the different methods. To investigate the presence of systematic calls in each method, we compute how many of the calls in a typical sample are called at different frequencies in the cohort. For example, on average, how many calls in one sample are called in more than 90% of the samples. Then, for each method, the samples were clustered using the CNV calls and hierarchical clustering with different linkage criteria (see Supplementary Information). The Rand index estimated the concordance between the

clustering and the known pedigree. Next, we measured the number of CNVs identified in each twin that were also found in the matching twin. We removed calls present in more than 50% of the samples to ensure that systematic errors were not biasing our replication estimates. Hence, a replicated call is most likely true as it is present in a minority of samples but consistently in the twin pair. The number and the proportion of replicated calls are used as a surrogate to explore sensitivity and specificity, respectively.

The approach described previously comparing pairs of twins was also applied in a renal cell carcinoma dataset[33], on pairs of normal/tumor samples. In this case, a replicated call is found in the normal sample and in the paired tumor sample. Finally, we compared calls using small bins (500 bp) and calls using larger bins (5 Kbp). This comparison explores the quality of the calls, the size of detectable events and the resolution for different bin size. First, we counted how many small bin calls supported any large bin call. We then looked at the proportion of small bin calls of different size that were also found in the large bin calls.

Validation by Taqman RT-PCR We first selected CNV calls in epilepsy patients that spanned at least 2 consecutive bins. We kept CNVs overlapping a Taqman probe and prioritized exonic CNVs. A second batch of CNVs, containing small non-coding CNVs, was also sent for validation. Hundreds of non-coding CNVs spanning only one bin were randomly selected. When possible the breakpoints were manually fine-tuned from manual inspection of a base-pair level coverage representation or using IGV; the breakpoints remained unchanged when they couldn't be refined. Finally, we kept regions overlapping a Taqman probe.

Probes were selected using the assay search tool on the Thermofisher website. All probes were tested for patients and controls that were called in PopSV as well as an additional 10 control individuals to ensure the validity of the probe. For each CNV, one assay was chosen in the middle of the genomic region of interest and located in an exon when possible. All reactions with TaqMan Copy Number Assays were performed in duplex using the FAM dye label based assay for the target of interest (Taqman copy number assay, Made to order, #4400291, Applied Biosystems by Life Technologies) and the VIC dye label based TaqMan Copy Number Reference Assay for RNase P (4403326, Life technologies). Amplification reactions ($10\mu L$), which were performed in quadru-

plicate, consisted of: 10 ng gDNA, 1X TaqMan Copy Number Assay, 1X TaqMan Copy Number Reference Assay, RNase P, 1X TaqMan Genotyping Master Mix (4371355, Life Technologies) or 1X SensiFAST Probe Lo-ROX Kit (BIO-84020, Froggabio). PCR was performed with an Applied Biosystems QuantStudio7 flex Real-Time PCR system using the standard curve settings and the default universal cycling conditions: 95 °C 10 minutes followed by 40 cycles: 95 °C 15 seconds, 60 °C 60 seconds. Data were analyzed with QuantStudio Real-Time PCR system software v1.2 (Applied Biosystems by Life Technologies) using autobaseline and manual Ct threshold of 0.2. Results export files were opened in CopyCaller™ Software v2.0 for sample copy number analysis by the relative quantitation method. The median ΔCt was used as the calibrator sample in the analysis settings.

CNVs enrichment in exonic regions CNV were called using PopSV and 5 Kbp bins, using all the samples as reference. For each cohort, we retrieved the CNV catalog by merging CNV that are recurrent in multiple samples. Hence, the CNV catalog represents all the different CNVs found in each cohort. To control for the population size, we sub-sampled 150 samples in each cohort. For each sub-sampling and each cohort, control regions were selected to fit the size distribution of the CNV catalog and the overlap with centromeres, telomeres and assembly gaps. The fold-enrichment represents how much more/less of the CNVs overlap an exon compared to the control regions. The same was performed for exons from genes that are likely loss-of-function intolerant[35] (see Supplementary Information). The significance was assessed by computing the fold-enrichment after permuting a thousand times the cohort labels. The same approach was used with rare CNVs, i.e. being present in less than 1% of PopSV calls in the Twins and renal cancer datasets, and in four public datasets (see Supplementary Information).

To test for a difference in deletion/duplication ratio among rare CNVs, we compared the numbers of rare deletions and duplications in the epilepsy patients and controls using a χ^2 test. The same test was performed after downsampling the controls to the sample size of the epilepsy cohort.

In each cohort, we then retrieved the CNV catalog of rare exonic CNVs. We evaluated the proportion of the CNVs in the catalog that are private (i.e. seen in only one sample), or seen in 2 samples or more, 3 samples or more, etc. The control cohort was down-sampled a thousand times

to the same sample size as the epilepsy cohort to provide a confidence interval and empirical P-value (see Supplementary Information). We performed the same analysis after removing the top 20 samples with the most non-private rare exonic CNVs (Figure S10). The analysis was also replicated using French-Canadian individuals only.

CNV enrichment in and near epilepsy genes We used the list of genes associated with epilepsy from the EpilepsyGene resource[36] which consists of 154 genes strongly associated with epilepsy. For a particular set of CNV we count how many of the genes hit are known epilepsy genes. To control for the gene size of epilepsy genes and CNV-hit genes, we randomly selected genes with sizes similar to the genes hit by CNVs and evaluated how many of these were epilepsy genes. After ten thousand samplings, we computed an empirical P-value (see Supplementary Information). Using this sampling approach we tested different sets of CNVs: deletion or duplications of different frequency in the epilepsy cohort, control individuals and samples from the twin study.

To investigate rare non-coding CNVs close to known epilepsy genes, we counted how many patients have such a CNV at different distance threshold. We compared this cumulative distribution to the control cohort, after down-sampling it to the sample size of the epilepsy cohort. We performed the same analysis on rare non-coding CNVs that overlap an eQTL associated with the epilepsy genes[37] or a DNase I hypersensitive site associated with the promoter of epilepsy genes[38]. A Kolmogorov-Smirnov test was used to test the difference in distribution. We computed the odds ratio of having such a CNV for different distance thresholds between epilepsy patients and controls.

Putatively pathogenic CNVs Exonic CNVs larger than 10 Kbp and found in less than 1% of the 301 controls were first selected. We further retained either CNVs overlapping the exon of a known epilepsy-associated gene[36] or deletions overlapping the exon of a loss-of-function intolerant gene[35], or CNVs present in two or more of our epilepsy patients. All the putatively pathogenic CNVs were validated by Taqman RT-PCR.

3 Results

Technical bias in read coverage We sequenced the genome of 198 unrelated individuals affected with epilepsy and 301 unrelated healthy controls. We compared the coverage between samples and across the genome to look for technical biases (e.g. GC correction, mappability filtering). In contrast to simulated datasets, we found that the inter-sample mean coverage in each bin varied between genomic regions even after stringent corrections and filters (Figure 1a). Supporting this observation, the bin coverage variance across samples was also lower than expected and varied between regions (Figure S1a). We also observed experiment-specific biases. In particular, some samples consistently had the highest, or the lowest, coverage across large portions of the genome (Figure S1b). These observations were not unique to our data and could also be observed in two public WGS datasets (Figure S2). This fluctuation of coverage has implications for CNV detection approaches that assume a uniform distribution[10,9,39] after standard bias correction and will lead to false positives.

CNV detection with PopSV To better control for technical bias, we developed PopSV, a new SV detection method. PopSV uses read depth across the samples to normalize coverage and detect change in DNA copy number (Figure 1b). The normalization step here is critical since most approaches will fail to give acceptable normalized coverage scores (Figure S1b). Moreover, with global median/variance adjustment or quantile normalization, the remaining subtle experimental variation impairs the abnormal coverage test (Figure S3a). The targeted normalization used by PopSV was found to have better statistical properties (Figure S3b). In order to assess the performance of our tool, CNVs calls were obtained from several algorithms on a twin cohort[8,9,10,11]. We found that PopSV performed as well or better in different aspects. First, for several algorithms, a large proportion of the detected events in a typical sample were also identified in almost all samples (60% of the calls found in >95% of the samples, Figure S4). PopSV's calls were better distributed across the frequency spectrum, hence more informative as we expect the relative frequency of disease-related variants to be rare. In addition, the pedigree structure was more accurately recovered when the CNVs were used to cluster the individuals in the Twins cohort (Figure S5). The agreement with the pedigree was computed by the Rand index after clustering the individuals with three hierar-

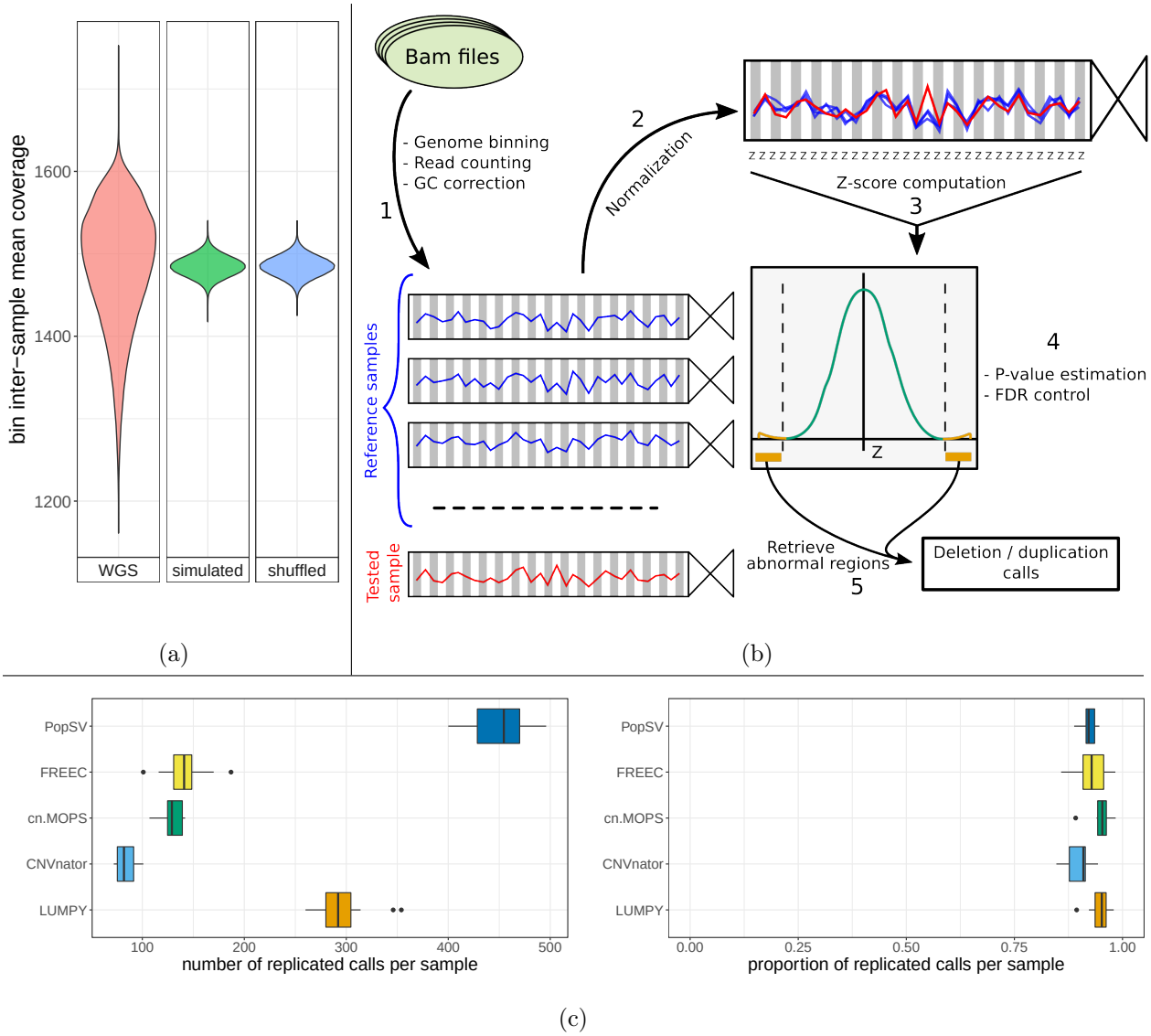
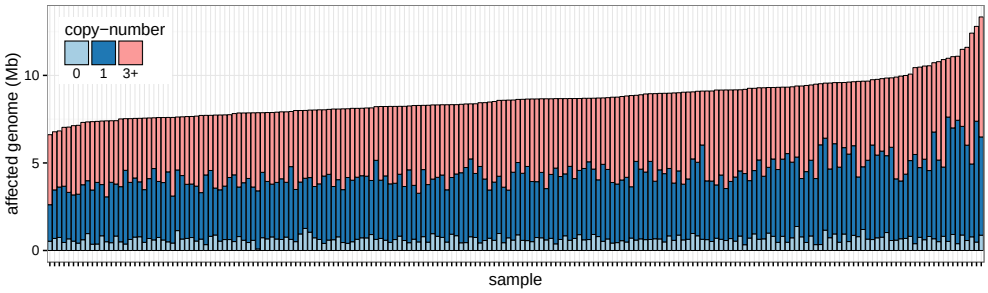


Figure 1: **PopSV approach.** a) Technical bias across the genome remains after stringent correction and filtering. The distribution of the bin inter-sample mean coverage in the epilepsy cohort (red) is compared to null distributions (blue: bins shuffled, green: simulated normal distribution). b) PopSV approach. 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). c) Number and proportion of calls from a twin that was replicated in the other twin.

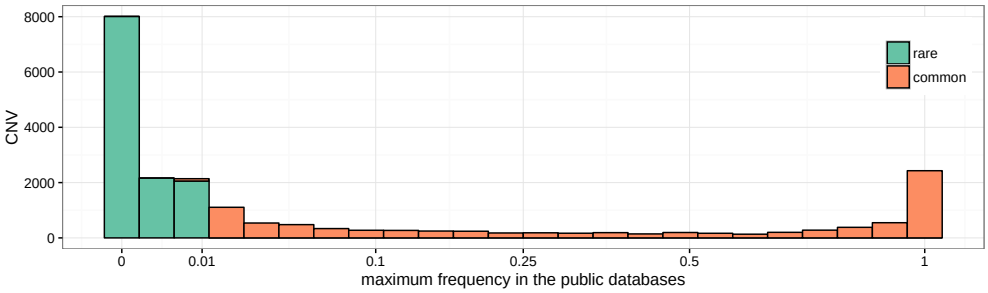
chical clustering approaches (see Supplementary Information). Looking at the replication between twins, PopSV detected more replicated CNVs compared to other methods, while maintaining similar replication rate (Figure 1c). Similar results were also obtained in a different dataset (Figure S6). Finally, we repeated the twin analysis using 500 bp bins and observed high consistency with the 5 Kbp calls (Figure S7). These results suggest that PopSV can accurately detect 75% of events that are as large as half the bin size used.

CNV in the CENet cohort Having demonstrated the quality of the PopSV calls, we applied our tool to the epilepsy cohort. The cohort comprises 198 individuals diagnosed with either generalized (n=160), focal (n=32) or unclassified (n=6) epilepsy. CNVs ranged from 5 Kbp to 3.2 Mbp with an average size of 9.98 Kbp. We observed an average of 870 CNVs per individual accounting for 8.7 Mb of variant calls (Figure 2a). This is around 9 times more variants and considerably smaller than in a typical array-based study[40] (Figure S8), such as the previous epilepsy surveys[25,29,26,28]. Then, we annotated each variant using four public SV databases[13,41,42,43] as well as PopSV calls in the two public datasets used earlier (see Supplementary Information). For each CNV, we derived the maximum frequency across these databases and defined as rare any region consistently annotated in less than 1% of the individuals (Figure 2b). In total, we identified 12,480 regions with rare CNVs in the epilepsy cohort including: 8,022 (64.3%) with heterozygous deletions, 21 (0.2%) with homozygous deletions and 4,850 (38.9%) with duplications. Although the overall amount of rare CNVs was not higher in epilepsy patients, the proportion of deletion was significantly higher compared to controls (χ^2 test: P-value 10^{-7}). Next, we selected 151 CNVs and further validated them using a Taqman CNV assay and RT-PCR. To explore PopSV's performance across different CNV profiles, we selected variants of different type, size and frequency. We found that the calls were concordant in 90.7% of the cases (Table S2 and S3). As expected, the estimated false positive rate was slightly higher for rare or smaller variants (12.1% for rare CNVs; 15.1% for CNV <20 Kbp).

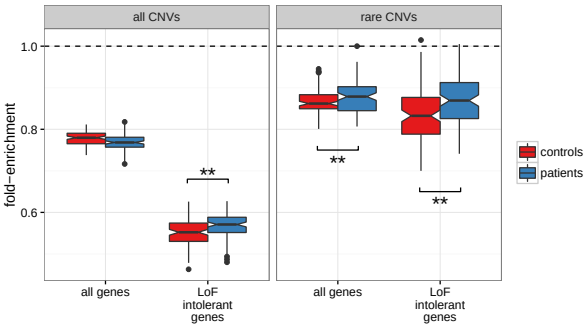
CNV enrichment in exonic regions To assess the role of CNVs in the pathogenic mechanism of epilepsy, we evaluated the prevalence of exonic CNVs in our epileptic cohort compared with healthy controls. First, focusing on CNVs larger than 50 Kbp, we observed the increased exonic burden



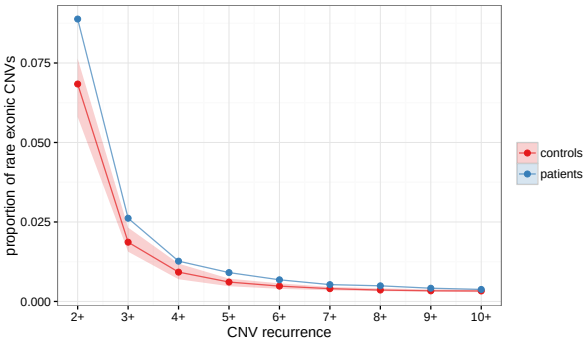
(a)



(b)



(c)

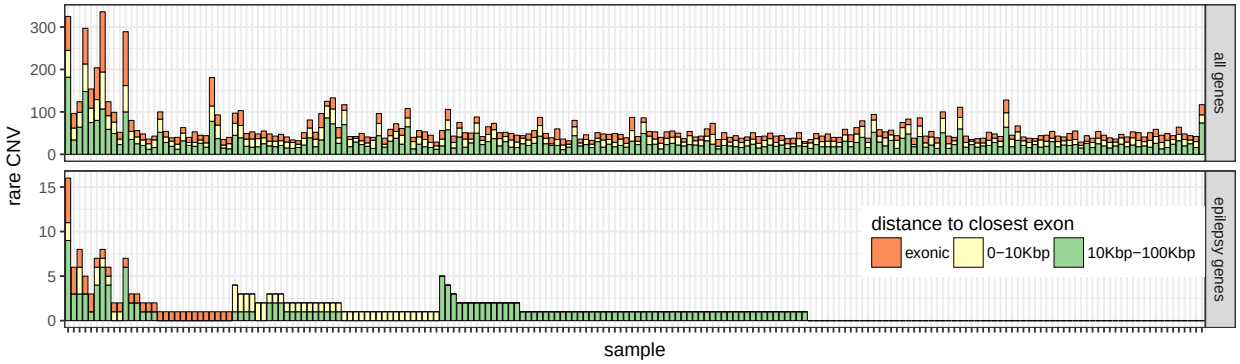


(d)

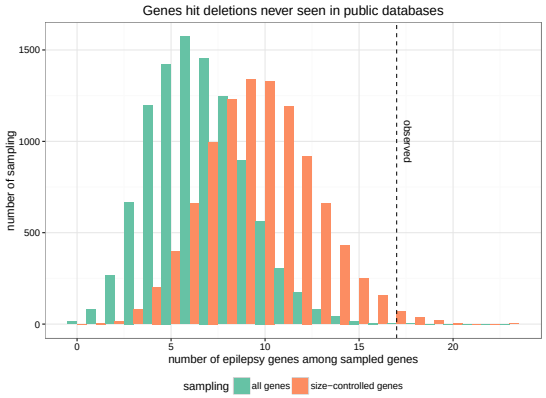
Figure 2: **CNVs in the epilepsy cohort.** a) Regions with a CNV in each epilepsy patient. b) Each CNV in the CNV catalog of the epilepsy cohort was annotated with its maximum frequency in five SV databases. c) Enrichment in exonic sequence for all CNVs and rare CNVs. d) Proportion of rare exonic CNVs (y-axis) seen in X or more individuals (x-axis). The ribbon shows the 5%-95% confidence interval.

described previously[30,29,31] (Figure S9). We observed fewer large CNVs overlapping exonic sequence than expected by chance but the fold-enrichment was higher in epileptic patients than in controls. When considering all CNVs however, there was no significant difference between epileptic patients and controls (Figure 2c). The number of CNVs overlapping exonic sequences of genes intolerant to loss-of-function mutations[35] was even lower. Interestingly, the coding regions of those genes were significantly more affected by CNVs in epileptic patients compared with controls (permutation P-value < 0.001). Because they are more likely pathogenic and of greater interest, we performed the same analysis using rare CNVs only. We observed a clear and significant enrichment of rare exonic CNVs in epileptic patients (permutation P-value < 0.001). The exonic enrichment for intolerant genes was also stronger (Figure 2c). In both cohorts, most of the rare exonic CNVs were private, i.e. present in only one individual. However, we observed that rare exonic CNVs were less likely private in the epileptic patients (permutation P-value < 0.001, Figure 2d). We also replicated this result using only individuals with a similar population background (French-Canadians, Figure S10). Overall we concluded that rare CNVs were not only enriched in exons but also affected exons more recurrently in the epilepsy cohort as compared to controls.

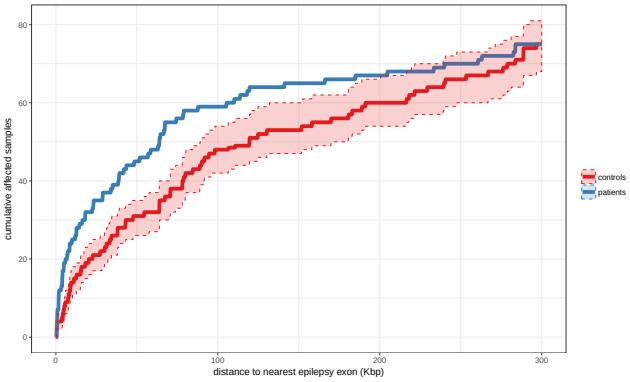
CNV enrichment in and near epilepsy genes We then sought to evaluate if there was an excess of CNVs disrupting epilepsy-related genes or nearby functional regions. We first retrieved genes whose exons were hit by rare deletions or duplications and evaluated how many were known epilepsy genes based on a list of 154 genes previously associated with epilepsy[36] (Figure 3a). Because epilepsy genes tend to be large, we controlled for the gene size when testing for enrichment (Figure S11a). In the epilepsy cohort only, we noted a clear enrichment for epilepsy genes hit by rare deletions (Figure S11b). Moreover, the enrichment became stronger for rare CNVs. For instance, the exons of 921 genes were disrupted in the epilepsy cohort when considering deletions completely absent from the public databases, 17 of which were epilepsy genes (P-value 0.015, Figure 3b). In addition, we observed significantly more epilepsy patients with a rare non-coding CNV close to an epilepsy gene compared to control individuals (Figure S12). To get a better idea of the functional regions close to epilepsy genes, we retrieved their associated eQTLs in the GTEx database[37] and the DNase hypersensitivity sites associated with their promoter regions[38]. Notably, focusing on



(a)



(b)



(c)

Figure 3: **CNVs and epilepsy genes.** a) Number of rare CNV in or close to exons of protein-coding genes (top) or epilepsy genes (bottom), in the epilepsy cohort. b) Number of genes hit by exonic deletions in the epilepsy cohort never seen in the public databases, compared to the expected distribution in all genes and matched-size genes. c) Rare non-coding CNVs in functional regions near epilepsy genes. The graph shows the cumulative number of individuals (y-axis) with a rare non-coding CNV located at X Kbp or less (x-axis) from the exonic sequence of a known epilepsy gene. We used CNVs overlapping regions functionally associated with the epilepsy gene (eQTL or promoter-associated DNase site).

rare non-coding CNVs overlapping these functional regions, the enrichment in epileptic patients was greatly strengthened and clearly present up to 100 Kbp from an epilepsy gene (Kolmogorov-Smirnov test: P-value 9×10^{-5}, Figure 3c). Comparing epilepsy patients and controls, the odds ratio of having such a CNV at a distance of 100 Kbp or less was 1.33 and gradually increased the closer to the exon (2.9 for CNVs at 5 Kbp or less, Figure S13). These non-coding CNVs were rare even in the epileptic cohort, but collectively represented an important fraction of affected patients. While 20 patients (10.1%) had exonic CNVs in epilepsy genes that were not seen in any control or public database, this number rose to 57 patients (28.8%) when counting non-coding CNVs in functional regions located at less than 100 Kbp of an epilepsy gene. These non-coding CNVs were never seen in the controls nor public databases and overlap with annotated enhancer of epilepsy genes. Although their functional impact remains putative, we believe these CNVs to be of high-interest for the identification of disease causing genes. Among these CNVs of high-interest, a duplication of a regulatory region 5 Kbp downstream of *CSNK1E* was detected and validated in two different patients but absent from our controls and the public databases (Figure S14a). Another example is a short deletion of an extremely conserved region downstream of *FAM63B*, detected in one patient and overlapping expression QTLs for this epilepsy gene (Figure S14b).

Putatively pathogenic CNVs Next, we used an array of criteria to select the rare CNVs (less than 1% in 301 controls) with the highest disruptive potential in the epilepsy cohort. Priority was given to CNVs in genes already known to be associated with epilepsy. For CNVs in other genes, we also prioritize recurrent variants and deletions in genes highly intolerant to loss-of-function mutations. In total, we identified 21 such putative pathogenic CNVs (Tables 1-2 and Table S4). Out of these, 8 directly affected a gene previously associated with epilepsy[36] (Table 1). In particular, we identified a deletion resulting in the loss of more than half of the *DEPDC5* gene in a patient affected with partial epilepsy. A number of point mutations have previously been reported in this gene for the same condition[44,45]. We also identified two deletions and one duplication in *CHD2* gene (see Figure 4). The first deletion is large and affects a major portion of the gene while the second is a small 4.6 Kbp deletion of exon 13, the last exon of *CHD2*'s second isoform (Figure S15). No exon-disrupted CNVs were reported in any individuals from the control cohort. This gene

was previously associated with patients suffering from photosensitive epilepsy[46]. Interestingly, all three patients carrying the CNVs in *CHD2* have been diagnosed with eyelid myoclonia epilepsy with absence, the same diagnosis that was largely enriched in the Galizia *et al.* study. Other known epilepsy genes affected by deletions include *LGI1* and the 15q13.3 region.

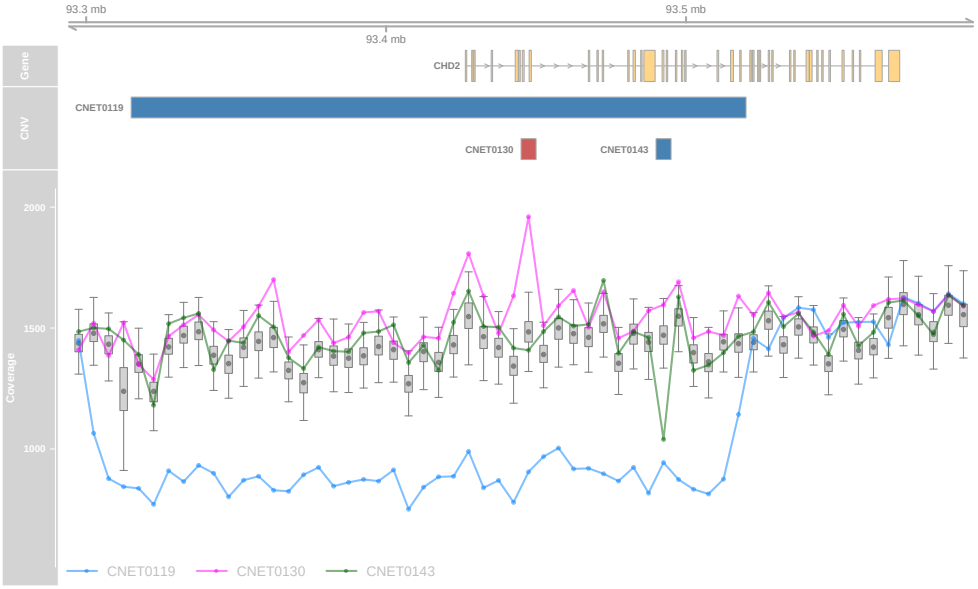


Figure 4: **Exonic CNVs in CHD2 detected by PopSV.** The 'CNV' panel shows the exonic deletions (blue) and duplications (red) called by PopSV. The 'Coverage' panel shows the read depth signal in the affected individuals (colored points/lines) and the coverage distribution in the reference samples (boxplot and grey point).

Four of the 21 putative pathogenic CNVs were found in more than one individual (see Table 2 for precise numbers). To assess their global prevalence we tested them in an additional cohort of 325 epileptic patients and 380 ethnically matched controls (Table 2). Two regions were replicated: the first region in chromosome 2 consists of duplication of the genes *TTC27*, *LTPB1* and *BIRC6*. In total, 4 patients carried this duplication and it was not reported in any of the two sets of controls. The second region was found on chromosome 16 and encompasses several genes. Two deletions were found in epileptic patients for this region and 1 epileptic individual and 1 control were also carrier of a duplication in the same region. Finally, the remaining putative pathogenic CNVs were also associated with a number of genes (see Table S4). However, as we lack additional evidence for

Patient	Epilepsy type	Syndrome	Copy number	Chr.	CNV start	CNV end	Epilepsy gene with exon disrupted	Taqman probe	Discovery		Replication	
									Patients	Controls	Patients	Controls
CNET0108	Generalized	Eyelid myoclonia epilepsy with absence	1	1	44195001	44460000	ST3GAL3	Hs05759463_cn	1 DEL	0	0	-
CNET0159	Generalized	Eyelid myoclonia epilepsy with absence	1	8	141925001	142010000	PTK2	Hs06202928_cn	1 DEL	0	0	-
CNET0093	Generalized	Juvenile onset; GTCs, Abs, Comp Partial	1	10	95525001	95545000	LGI1	Hs02682696_cn	1 DEL	0	0	-
CNET0140	Generalized	Idiopathic generalized epilepsies	1	13	35750001	35785000	NBEA	Hs05286691_cn	1 DEL	0	0	-
CNET0144	Generalized	Eyelid myoclonia epilepsy with absence	1	15	22745001	23275000	NIPA2	Hs04452887_cn	3 DEL	2 DEL	4 DEL (2DUP)	1 DEL (5 DUP)
CNET0009	Generalized	Idiopathic generalized epilepsies	1	15	30910001	32445000	CHRNA7	Hs03909657_cn	1 DEL	0	3 DEL	(1 DUP)
CNET0119	Generalized	Eyelid myoclonia epilepsy with absence	1	15	93300001	93515000	CHD2	Hs05385106_cn	1 DEL	0	0	-
CNET0143	Generalized	Childhood absence epilepsy	1		93489776	93494317		Hs026436998_cn	1 DEL	0	0	-
CNET0130	Generalized	Eyelid myoclonia epilepsy with absence	3		93445001	93450000		Hs01379802_cn	1 DUP	0	0	-
CNET0074	Focal	Frontal Lobe Epilepsy	1	22	32125001	32255000	DEPDC5	Hs01632214_cn	1 DEL	0	0	-

Table 1: **Pathogenic profiles in known epilepsy genes.** The 198 epileptic patients and 301 controls represent the discovery set. The replication set contains 325 epileptic patients and 380 controls. Variants that were not tested are marked with "-".

315 those specific CNV regions, we propose that these genes should be assessed in independent epilepsy cohorts. Of note, one patient had a rare 170 Kbp deletion encompassing three exons of the *PTPRD* gene which is predicted to be highly intolerant to loss-of-function mutations (pLI=1)[35]. Rare deletions in this gene were previously found in four independent individuals with attention-deficit hyperactivity disorder[47] and associated with intellectual disability[48]. De novo deletions were also 320 found in an individual with autism[49] and more recently in a patient with epileptic encephalopathy[27].

4 Discussion

Although several tools exist for the detection of CNVs using WGS data, we found that none of them could efficiently account for technical biases, thus resulting in limited sensitivity. To improve on this, we developed a new tool, PopSV, which we demonstrated was able to accurately detect CNVs, 325 including rare and small events.

As in previous array-based studies[30,29,31], we observed an enrichment of large exonic CNVs in patients compared to controls. However, thanks to the resolution of WGS and PopSV, we found that the global distribution of CNVs in 198 unrelated epilepsy patients was also skewed towards rare

Patient	Epilepsy type	Syndrome	Copy number	Chr.	CNV start	CNV end	Gene with exon disrupted	Taqman probe	Discovery		Replication	
									Patients	Controls	Patients	Controls
CNET0184	Generalized	Lennox-Gastaut syndrome	3	2	32625001	33335000	TTC27;LTBP1;BIRC6	Hs03387774_cn	2 DUP	0	2 DUP	0
CNET0097	Generalized	Eyelid myoclonia epilepsy with absence	3									
CNET0020	Generalized	Juvenile myoclonic epilepsy	1	12	7995001	8125000	SLC2A3;SLC2A14	Hs04406005_cn	2 DEL	2 DEL	2 DEL	2 DEL
CNET0198	Focal	Frontal lobe epilepsy	1									
CNET0012	Generalized	Idiopathic generalized epilepsy	3	15	90845001	90955000	ZNF774;IQGAP1	Hs03895490_cn	2 DUP	0	(1 DEL)	0
CNET0167	Generalized	Childhood absence epilepsy	3									
CNET0063	Generalized	Idiopathic generalized epilepsies	3	16	15460001	16290000	KIAA0430;MPV17L;NPIPA5; C16orf45;ABCC6;NDE1; FOPNL;ABCC1;MYH11	Hs05396556_cn	1 DUP + 1 DEL	0	1 DEL	1 DUP
CNET0037	Generalized	Idiopathic generalized epilepsies	1									

Table 2: **Recurrent CNVs with a pathogenic profile.** The 198 epileptic patients and 301 controls represent the discovery set. The replication set contains 325 epileptic patients and 380 controls.

exonic CNVs. In addition, genes disrupted by rare deletions in patients were enriched for previously known epilepsy genes. These observations are driven by small variants (<10 Kbp) and could not have been detected in previous array-based studies.

We also observed a clear enrichment of non-coding CNVs in the neighborhood of previously implicated genes. When focusing on CNVs seen only in the epilepsy cohort and around epilepsy genes, 10.1% of epilepsy patients have an exonic CNVs and our results shows that up to 28.8% of patients harbor non-coding CNVs of high-interest in the proximity of epilepsy genes. These non-coding variants are present in the epilepsy cohort only and located in annotated regulatory regions associated to known epilepsy genes. Although it is challenging to directly test their functional impact, their frequency and location suggest a putative importance in the genetic mechanism of epilepsy and should be further investigated in the future.

Finally, to better understand the impact of these findings on an individual scale, we selected CNVs with the highest pathogenic potential within our patients. These CNVs highlighted known but also potentially new epilepsy genes. Using a second cohort, we were also able to identify two chromosomal regions that were recurrently disrupted by CNVs. These findings highlight the benefits of having a comprehensive survey of CNVs when trying to understand the genetic causes of a disease.

5 Data and code availability

PopSV R package and documentation are available at http://jmonlong.github.io/PopSV/. The code used for the analysis and to produce figures and numbers is documented at http://github.com/jmonlong/epipopsv and archived in https://doi.org/10.5281/zenodo.1002893. Necessary data, including the CNV calls, was deposited at https://figshare.com/s/20dfdedcc4718e465185. The raw sequence data will be deposited at EGA prior to publication.

6 Acknowledgments

This work was supported by a grant from Genome Canada/Génome Québec, the Canadian Foundation for Innovation (CFI-32462), the National Sciences and Engineering Research Council (NSERC-448167-2013) and a grant from the Canadian Institute for Health Research (CIHR-MOP-115090). PC holds a Canada Research Chair in Genomics of Epilepsy. SLG and GB are supported by the Fonds de Recherche Québec Santé (FRQS-29493 and FRQS-25348). SLG is supported by a grant from the Réseau de Médecine Génétique Appliquée, a research network from the FRQS. 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 consortium who provided the renal cancer dataset. We would like to thank Sylvia Dobrzeniacka for sample handling and lab work. We are grateful to Dr. Ledia Brunga for her work on the epileptic cohort and to Brianna Goldenstein and Claudia Moreau for revising this manuscript. Finally, we would like to thank Simon Gravel, Mathieu Blanchette, Mathieu Bourgey, Toby Dylan Hocking, Fadi Hamdan and Claudia Moreau for helpful discussions.

7 Author Contributions

SLG, GB and PC conceived and designed the study. DMA, MG, CB, JLM, GR, BAM, RGL, FH performed the clinical recruitment. JM implemented the method. JM, SLG, ADL and DS performed the analyses. JM, CM and SLG designed the experimental validation. CM and MCD performed the experimental validation. JM, SLG and GB wrote the manuscript.

References

[1] Hall, I. M. and Quinlan, A. R. (2012). Detection and Interpretation of Genomic Structural Variation in Mammals. In Methods in Molecular Biology volume 838 pages 225–248.

[2] Sharp, A. J., Cheng, Z., and Eichler, E. E. (2006). Structural Variation of the Human Genome. Annual Review of Genomics and Human Genetics *7*, 407–442.

[3] Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K. *et al.* (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59–65.

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

[5] Spielmann, M. and Klopocki, E. (2013). CNVs of noncoding cis-regulatory elements in human disease. Current Opinion in Genetics & Development *23*, 249–256.

[6] Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics *14*, S1.

[7] Pirooznia, M., Goes, F., and Zandi, P. P. (2015). Whole-genome CNV analysis: Advances in computational approaches. Frontiers in Genetics *6*, 1–9.

[8] Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biology *15*, R84.

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

[10] Boeva, V., Zinovyev, A., Bleakley, K., Vert, J. P., Janoueix-Lerosey, I., Delattre, O., and

Barillot, E. (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics *27*, 268–269.

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

[12] Glusman, G., Severson, A., Dhankani, V., Robinson, M., Farrah, T., Mauldin, D. E., Stittrich, A. B., Ament, S. A., Roach, J. C., Brunkow, M. E. *et al.* (2015). Identification of copy number variants in whole-genome data using reference coverage profiles. Frontiers in Genetics *5*, 1–13.

[13] Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., and McCarroll, S. A. (2015). Large multiallelic copy number variations in humans. Nature Genetics *47*, 296–303.

[14] Treangen, T. J. and Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics *13*, 36–46.

[15] Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S., and Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics *28*, 2711–2718.

[16] Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research *40*, e72–e72.

[17] Koren, A., Handsaker, R. E., Kamitaki, N., Karlić, R., Ghosh, S., Polak, P., Eggan, K., and McCarroll, S. A. (2014). Genetic variation in human DNA replication timing. Cell *159*, 1015–1026.

[18] Cheung, M. S., Down, T. A., Latorre, I., and Ahringer, J. (2011). Systematic bias in high-throughput sequencing data and its correction by BEADS. Nucleic Acids Research *39*, e103–e103.

[19] Berkovic, S. F., Howell, R. A., Hay, D. A., and Hopper, J. L. (1998). Epilepsies in twins: Genetics of the major epilepsy syndromes. Annals of Neurology *43*, 435–445.

[20] Zara, F., Bianchi, A., Avanzini, G., Di Donato, S., Castellotti, B., Patel, P. I., and Pandolfo, M. (1995). Mapping of genes predisposing to idiopathic generalized epilepsy. Human Molecular Genetics *4*, 1201–7.

[21] Guo, Y., Baum, L. W., Sham, P. C., Wong, V., Ng, P. W., Lui, C. H. T., Sin, N. C., Tsoi, T. H., Tang, C. S. M., Kwan, J. S. H. *et al.* (2012). Two-stage genome-wide association study identifies variants in CAMSAP1L1 as susceptibility loci for epilepsy in Chinese. Human Molecular Genetics *21*, 1184–1189.

[22] Kasperavičiüte, D., Catarino, C. B., Heinzen, E. L., Depondt, C., Cavalleri, G. L., Caboclo, L. O., Tate, S. K., Jamnadas-Khoda, J., Chinthapalli, K., Clayton, L. M. S. *et al.* (2010). Common genetic variation and susceptibility to partial epilepsies: A genome-wide association study. Brain *133*, 2136–2147.

[23] International League Against Epilepsy Consortium on Complex Epilepsies (2014). Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. The Lancet Neurology *13*, 893–903.

[24] Biervert, C. (1998). A Potassium Channel Mutation in Neonatal Human Epilepsy. Science *279*, 403–406.

[25] Mefford, H. C., Muhle, H., Ostertag, P., von Spiczak, S., Buysse, K., Baker, C., Franke, A., Malafosse, A., Genton, P., Thomas, P. *et al.* (2010). Genome-Wide Copy Number Variation in Epilepsy: Novel Susceptibility Loci in Idiopathic Generalized and Focal Epilepsies. PLoS Genetics *6*, e1000962.

[26] Helbig, I., Swinkels, M. E. M., Aten, E., Caliebe, A., van 't Slot, R., Boor, R., von Spiczak, S., Muhle, H., Jähn, J. a., van Binsbergen, E. *et al.* (2014). Structural genomic variation in childhood epilepsies with complex phenotypes. European Journal of Human Genetics *22*, 896–901.

[27] Mefford, H. (2015). Copy number variant analysis from exome data in 349 patients with epileptic encephalopathy. Annals of Neurology *78*, 323–328.

[28] Addis, L., Rosch, R. E., Valentin, A., Makoff, A., Robinson, R., Everett, K. V., Nashef, L., and Pal, D. K. (2016). Analysis of rare copy number variation in absence epilepsies. Neurology Genetics *2*, e56.

[29] Mefford, H. C., Yendle, S. C., Hsu, C., Cook, J., Geraghty, E., McMahon, J. M., Eeg-Olofsson, O., Sadleir, L. G., Gill, D., Ben-Zeev, B. *et al.* (2011). Rare copy number variants are an important cause of epileptic encephalopathies. Annals of Neurology *70*, 974–985.

[30] Heinzen, E. L., Radtke, R. A., Urban, T. J., Cavalleri, G. L., Depondt, C., Need, A. C., Walley, N. M., Nicoletti, P., Ge, D., Catarino, C. B. *et al.* (2010). Rare Deletions at 16p13.11 Predispose to a Diverse Spectrum of Sporadic Epilepsy Syndromes. American Journal of Human Genetics *86*, 707–718.

[31] Striano, P. (2012). Clinical Significance of Rare Copy Number Variations in Epilepsy. Archives of Neurology *69*, 322.

[32] Boivin, M., Brendgen, M., Dionne, G., Dubois, L., Pérusse, D., Robaey, P., Tremblay, R. E., and Vitaro, F. (2013). The Quebec Newborn Twin Study Into Adolescence: 15 Years Later. Twin Research and Human Genetics *16*, 64–69.

[33] Scelo, G., Riazalhosseini, Y., Greger, L., Letourneau, L., Gonzàlez-Porta, M., Wozniak, M. B., Bourgey, M., Harnden, P., Egevad, L., Jackson, S. M. *et al.* (2014). Variation in genomic landscape of clear cell renal cell carcinoma across Europe. Nature Communications *5*, 5135.

[34] Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

[35] Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B. *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

[36] Ran, X., Li, J., Shao, Q., Chen, H., Lin, Z., Sun, Z. S., and Wu, J. (2015). EpilepsyGene: A genetic resource for genes and mutations related to epilepsy. Nucleic Acids Research *43*, D893–D899.

[37] Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M. *et al.* (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648–660.

[38] Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J. *et al.* (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science *337*, 1190–1195.

[39] Xi, R., Hadjipanayis, A. G., Luquette, L. J., Kim, T.-M., Lee, E., Zhang, J., Johnson, M. D., Muzny, D. M., Wheeler, D. A., Gibbs, R. A. *et al.* (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proceedings of the National Academy of Sciences *108*, E1128–E1136.

[40] Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W. *et al.* (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

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

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

[43] Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe,

495 B. P., Baker, C., Nordenfelt, S., Bamshad, M. *et al.* (2015). Global diversity, population stratification, and selection of human copy-number variation. Science *349*, aab3761–aab3761.

[44] Dibbens, L. M., de Vries, B., Donatello, S., Heron, S. E., Hodgson, B. L., Chintawar, S., Crompton, D. E., Hughes, J. N., Bellows, S. T., Klein, K. M. *et al.* (2013). Mutations in DEPDC5 cause familial focal epilepsy with variable foci. Nature Genetics *45*, 546–551.

500 [45] Ishida, S., Picard, F., Rudolf, G., Noé, E., Achaz, G., Thomas, P., Genton, P., Mundwiller, E., Wolff, M., Marescaux, C. *et al.* (2013). Mutations of DEPDC5 cause autosomal dominant focal epilepsies. Nature Genetics *45*, 552–555.

[46] Galizia, E. C., Myers, C. T., Leu, C., de Kovel, C. G. F., Afrikanova, T., Cordero-Maldonado, M. L., Martins, T. G., Jacmin, M., Drury, S., Krishna Chinthapalli, V. *et al.* (2015). CHD2
505 variants are a risk factor for photosensitivity in epilepsy. Brain *138*, 1198–1208.

[47] Elia, J., Gai, X., Xie, H. M., Perin, J. C., Geiger, E., Glessner, J. T., D'arcy, M., DeBerardinis, R., Frackelton, E., Kim, C. *et al.* (2010). Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. Molecular Psychiatry *15*, 637–646.

510 [48] Choucair, N., Mignon-Ravix, C., Cacciagli, P., Abou Ghoch, J., Fawaz, A., Mégarbané, A., Villard, L., and Chouery, E. (2015). Evidence that homozygous PTPRD gene microdeletion causes trigonocephaly, hearing loss, and intellectual disability. Molecular Cytogenetics *8*, 39.

[49] Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T. R., Correia, C., Abrahams, B. S. *et al.* (2010). Functional impact of global rare copy
515 number variation in autism spectrum disorders. Nature *466*, 368–372.