Population-Based Approaches to Characterize Copy Number Variation from Whole-Genome Sequencing in Healthy Individuals and Disease Cohorts

Jean Monlong

Faculty of Medicine
Department of Human Genetics
McGill University, Montreal, Canada
July, 2018

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Abstract

Copy number variation (CNV) affects genomic regions from 50 bp up to entire chromosomes. In addition to being one of the major forms of genomic variation during recent evolution, CNV is implicated in many genetic disorders, complex traits and cancers. Whole-genome sequencing (WGS) makes it possible to interrogate the genome for different types of variation: single nucleotide variants, small insertion-deletions, copy-number variant and other structural variants. However, technical bias remains a challenge for CNV detection, especially in repeat-rich regions or to detect small or somatic variants. The vast majority of CNV detection methods analyze one sample at a time or only aggregate evidence across samples. In this work, I present a different approach that uses a large set of reference samples to correct for technical variation. This population-based approach is used on three different applications. First at the chromosome arm level, I used WGS data across 93 blood samples to detect somatic CNV in paired kidney cancer samples. The population-based approach was sensitive enough to detect somatic loss or gain of chromosome arms despite weak signal in the bulk samples. We further studied tumors from male patients and found that somatic loss of chromosome Y was frequent and resulted in down-regulation of important genes such as *KDM5D* and *KDM6C*, two tumor suppressors previously associated with cancer. Next, a method was implemented to identify CNVs across the genome following a similar population-based strategy. After an extensive comparison with existing methods and experimental validation, we found that our method, PopSV, was more sensitive than other methods. Using PopSV and WGS data for 198 individuals with epilepsy and 301 controls, we studied the distribution of small CNVs across the genomes of epilepsy patients. In addition to the known enrichment in large rare exonic CNVs, we found a significant enrichment of rare exonic CNVs smaller than 50 Kbp in epilepsy patients, especially in genes predicted to be intolerant to loss of function variants. More interestingly we observed, for the first time, a strong enrichment of non-coding CNVs close to known epilepsy genes. Finally, we used PopSV to investigate copy number variation in low-

mappability regions. Thanks to its population-based strategy, PopSV's performance was stable across different repeat profiles and we further analyzed the genomes of 640 healthy individuals. In contrast to existing CNV databases, we found a large amount of CNVs in repeat-rich regions and identified regions with recurrent CNVs that were absent from existing CNV catalogs, many of which were located within or near protein-coding genes. Independently from the known enrichment in segmental duplications, we found strong CNV enrichments in low-mappability regions, DNA satellites, short-tandem repeats and specific families of transposable elements. Thanks to the ever-reducing cost of sequencing, large-scale WGS datasets are becoming more and more common. By using information across several samples, this work shows that variant detection can be dramatically improved and benefit CNV studies in cancer, complex disease or in challenging genomic regions.

Résumé

Les variabilités du nombre de copies (VNCs) sont des variations génomiques affectant 50 nucléotides ou plus. Les VNCs ont fortement contribué à l'évolution humaine récente mais jouent aussi un rôle important dans de nombreuses maladies génétiques et autres caractères complexes. Le séquençage du génome permet d'étudier différent types de variations génomiques: les substitutions d'un nucléotide, les petites insertions/délétions ainsi que les VNCs et autres variants structuraux. Cependant la présence de biais techniques limite la détection des VNCs, en particulier dans les régions répétées du génome ou pour détecter les variants les plus petits ou somatiques. La plupart des méthodes de détection de VNCs analysent chaque échantillon séparément ou accumulent naïvement le signal dans plusieurs échantillons. Dans cette étude, je présente une nouvelle approche qui vise à utiliser un groupe d'échantillons comme référence pour intégrer la variation d'origine technique. Cette approche est appliquée dans le cadre de trois études génomiques. Dans un premier temps au niveau chromosomique, j'utilise des données de séquençage de 93 échantillons de sang pour détecter des VNCs somatiques dans des échantillons de tumeur du rein provenant des mêmes individus. Grâce à l'utilisation d'échantillons de référence, les pertes ou gains de chromosomes somatiques dont le signal est faible ont aussi pu être détectées. Dans cette étude, nous nous concentrons ensuite sur la perte somatique du chromosome Y dans les tumeurs des patients hommes. Entre autre nous montrons que la perte somatique du chromosome Y est associée à une diminution de l'expression de ses gènes, dont *KDM5D* et *KDM6C*, deux gènes suppresseurs de tumeurs. Dans un second temps, j'ai développé une méthode utilisant une approche similaire pour détecter des VNCs dans le génome. À l'aide de données de séquencage et de validation expérimentale, nous montrons que notre méthode, PopSV, est plus sensible que les méthodes existantes. Nous étudions ensuite la distribution des VNCs dans 198 individus atteints d'épilepsie et 301 contrôles. Nous retrouvons l'enrichissement connu des VNCs larges, rares et exoniques mais nous montrons que les VNCs codants plus petits que 50,000 nucléotides sont aussi en-

richis dans les malades, notamment dans les gènes prédis pour être intolérants aux variants perte de fonction. Nous observons aussi pour la première fois un enrichissement de VNCs non-codants proches de gènes associés à l'épilepsie. Dans un troisième temps, j'utilise PopSV pour étudier la distribution des VNCs dans les régions répétées du génome de 640 individus sains. Malgré la difficulté inhérente à ces régions, la performance de notre approche reste stable. Nous trouvons de nombreux VNCs dans les régions répétées et identifions des régions qui contiennent fréquemment des VNCs mais absentes des catalogues publics de VNCs, notamment proches de gènes. De plus, nous décrivons un enrichissement dans les régions de faible mappabilité et dans certaines familles de satellites, microsatellites et éléments transposables, indépendemment de l'enrichissement connu dans les duplications segmentales. Ces résultats démontrent les bénéfices de l'utilisation d'échantillons de référence pour détecter les VNCs à partir de données de séquencage et pour étudier le profile génomique de cancers, maladies complexes ou génomes d'individus sains.

Table of Contents

List of Abbreviations

aCGH: **a**rray **C**omparative **G**enomic **H**ybridization.

AT: **A**denine or **T**hymine.

BAM: **B**inary **A**lignment/**M**ap.

BLAST: **B**asic **L**ocal **A**lignment **S**earch **T**ool.

bp: **b**ase **p**air.

cDNA: **c**omplementary **DNA**.

CENet: **C**anadian **E**pilepsy **Net**work.

ccRCC: **c**lear **c**ell **R**enal **C**ell **C**arcinoma.

CNV: **C**opy-**N**umber **V**ariation or **C**opy **N**umber **V**ariant.

CRISPR: **C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats.

CTG: **C**entromere, **T**elomere, **G**aps.

DGV: **D**atabase of **G**enomic **V**ariants.

DNA: **D**eoxyribo**N**ucleic **A**cid.

DNase: **D**eoxyribo**N**uclease.

dsDNA: **d**ouble-**s**tranded **DNA**.

eQTL: **e**xpression **Q**uantitave **T**rait **L**ocus.

ERV: **E**ndogenous **R**etro**V**irus.

FDR: **F**alse **D**iscovery **R**ate.

FISH: **F**luorescent **I***n* **S***itu* **H**ybridization.

FoSTeS: **Fo**rk **S**talling and **Te**mplate **S**witching.

GC: **G**uanine or **C**ytosine.

GoNL: **G**enome **o**f **N**ether**L**ands.

GTEx: **G**enotype-**T**issue **Ex**pression project.

HERV: **H**uman **E**ndogenous **R**etro**V**irus.

HIV/AIDS: **H**uman **I**mmunodeficiency **V**irus/**A**cquired **I**mmune **D**eficiency **S**yndrome.

ICGC: **I**nternational **C**ancer **G**enome **C**onsortium.

IQR: **I**nter**Q**uartile **R**ange.

Kbp: **K**ilo **b**ase **p**air, i.e. 1,000 bp.

LOX: **L**oss **O**f (chromosome) **X**.

LOY: **L**oss **O**f (chromosome) **Y**.

Mbp: **M**ega **b**ase **p**air, i.e. 1 million bp.

Gbp: **G**iga **b**ase **p**air, i.e. 1 billion bp.

MEI: **M**obile **E**lement **I**nsertion.

MMEJ: **M**icrohomology-**M**ediated **E**nd **J**oining.

NHEJ: **N**on-**H**omologous **E**nd **J**oining.

NAHR: **N**on-**A**llelic **H**omologous **R**ecombination.

OMIM: **O**nline **M**endelian **I**nheritance in **M**an.

PCR: **P**olymerase **C**hain **R**eaction.

QTL: **Q**uantitave **T**rait **L**ocus.

RCC: **R**enal **C**ell **C**arcinoma.

RD: **R**ead-**D**epth, also called read coverage or depth of coverage in the literature.

RNA: **R**ibo**N**ucleic **A**cid.

RPKM: **R**eads **P**er **K**ilobase per **M**illion mapped reads.

RT-PCR: **R**eal-**T**ime **PCR**.

sCNV: **s**omatic **C**opy-**N**umber **V**ariation.

sLOY **s**omatic **L**oss **O**f (chromosome) **Y**.

SNP: **S**ingle **N**ucleotide **P**olymorphism.

SNV: **S**ingle **N**ucleotide **V**ariant.

SV: **S**tructural **V**ariation or **S**tructural **V**ariant.

STR: **S**hort **T**andem **R**epeat.

SVA: **S**INE/**V**NTR/**A**lu.

TAD: **T**opologically **A**ssociating **D**omains.

TE: **T**ransposable **E**lement.

WGS: **W**hole-**G**enome **S**equencing.

List of Figures

List of Tables

Acknowledgments

I would first like to thank my supervisor, Dr. Guillaume Bourque, for having me as his PhD student, supporting me over the years and providing a great environment for me to grow as a scientist. I learned a lot about genomics, critical thinking, problem solving, science and academia from our discussions. I will leave Canada with a greater passion for research. In addition to the mentoring, I'm extremely grateful for the all the opportunities I was given: for sending me to many conferences, for including me to national and international projects, for encouraging my short stays abroad visiting his collaborators, and for helping me find a post-doc. These have been extremely valuable experiences. Also, thank you for giving me the freedom and trust to organize my time. I'm glad I was able to have teaching experiences or to explore ideas that were not directly relevant to my main project.

I would like to thank the members of my supervisory committee, Dr. Simon Gravel and Dr. Mathieu Blanchette, for their guidance, their availability, and their valuable thesis and career advice.

I would like to thank the former and current members of the Bourque lab who have made my day-to-day life as a PhD student so easy. Special thanks to Trish for being my partner in crime in the lab, Louis for showing me what to aim for in term of knowledge and skills, Mathieu for the discussions and for including me in exciting projects, Simon for his guidance and trust with his projects, Francois and Toby for the R enthusiasm, Eric for PhD advice, and everyone for the repeated support and faith in my project. It was always a pleasure to work in the lab, during the studious moments and during the breaks, debating over the student life vs staff life, having "important meetings" on Friday afternoon, or when someone was dressed up as the Queen to celebrate their citizenship... Good times.

I would also like to thank the Human Genetics department for the opportunities I benefited from and their support of the student society. I had a great experience as part of the executive committee of the student society and it is partly because of the recognition and support of the department. A special mention to Ross who was always so understanding and on top of the administrative work, especially during my troubles with the study permit renewal...

Then, I would like to thank my friends in Montreal. My PhD was also an awesome period of my life because of the friendships that grew in 692, the soccer team, the volley ball team, at the HGSS events, in the HGSS Exec, at conferences and other travels, at the Genome Center and around. Special thanks to Emma, Renata, Patricia, Marion and Vini; I knew I could be fully be myself with you. I am also thankful to my compatriotes Thibault, Richard and Stan with which I had great times during my first years here.

Enfin, je voudrais remercier ma famille pour leur soutien malgré la distance. J'ai hâte d'entrer dans le club familial des Drs. Monlong. Je remercie ma soeur Julie qui m'a montré la voie et a placé la barre bien haute. Et bien sur mes parents, Patrick et Claire, qui ont fait de moi la personne que je suis aujourd'hui et qui ont cultivé ma passion pour la science depuis tout petit. Je réalise à peine la chance que j'ai eu de grandir dans cette famille. J'ai une pensée toute particulière pour ma mère qui j'espère serait fière de moi.

Format of the Thesis

This thesis consists of 6 chapters. Chapter 1 is an introduction to the different topics relevant to my PhD work and presents its hypotheses and objectives. Chapter 2-4 are original research chapters. Chapter 2 contains a manuscript that was published in *Scientific Reports*[1]. Chapter 3 contains a manuscript that was published in *PLoS Genetics*[2]. The manuscript in Chapter 4 was published in *Nucleic Acids Research*[3]. Chapter 5 is a general discussion about the benefits of population-based approaches for CNV detection in whole-genome sequencing data. Chapter 6 concludes the thesis and describes future directions for population-based approaches of structural variant detection.

Appendix A lists other publications to which the thesis author have contributed during the thesis period. Appendix B, Appendix C and Appendix D contain supplementary material for chapters 2, 3 and 4, respectively.

Contributions of Authors

Chapter 1 contains a literature review covering sequencing technology, structural variation, genomic repeats, cancer and epilepsy. Chapter 5 and 6 contains a general discussion and future directions about population-based approaches and whole-genome sequencing studies. These chapters were written by the thesis author under the supervision of Dr. Bourque.

Chapter 2 represents a manuscript authored by Madeleine Arseneault, Jean Monlong, Naveen S. Vasudev, Ruhina S. Laskar, Maryam Safisamghabadi, Patricia Harnden, Lars Egevad, Nazanin Nourbehesht, Pudchalaluck Panichnantakul, Ivana Holcatova, Antonin Brisuda, Vladimir Janout, Helena Kollarova, Lenka Foretova, Marie Navratilova, Dana Mates, Viorel Jinga, David Zaridze, Anush Mukeria, Pouria Jandaghi, Paul Brennan, Alvis Brazma, Jorg Tost, Ghislaine Scelo, Rosamonde E. Banks, Mark Lathrop, Guillaume Bourque and Yasser Riazalhosseini. The thesis author developed the computational methods for WGS analysis, analyzed WGS data and generated the figures. YR conceived the study and designed the experiments with contribution from GB. NV, PH, LE, IH, AB, VJ, HK, LF, MN, DM, ViJ, DZ, AM, PB, GS and REB were responsible for patient selection, sample collection, sample preparation and pathological reviews. MA and NN prepared DNA, performed experiments to detect loss of chromosome Y and analyzed the data with contribution from PJ. PP, RSL, PJ and AlB performed the gene expression analysis. ML provided critical advice on data analysis and statistical approaches. MS and PJ performed functional analysis with cell line models. MA and YR wrote the manuscript, with assistance from JT, GS and REB.

Chapter 3 contains a manuscript by Jean Monlong, Simon L. Girard, Caroline Meloche, Maxime Cadieux-Dion, Danielle M. Andrade, Ron G. Lafreniere, Micheline Gravel, Dan Spiegelman, Alexandre Dionne-Laporte, Cyrus Boelman, Jacques L. Michaud, Guy Rouleau, Berge A. Minassian, Guillaume Bourque and Patrick Cossette. The thesis author implemented the method, performed the analysis with SLG, ADL and DS, designed the experimental validation with CM and SLG, and wrote the manuscript with SLG and GB. SLG, GB and PC conceived and designed the study. DMA, MG, CB, JLM, GR, BAM, RGL, FH performed the clinical recruitment. CM and MCD performed the experimental validation.

The manuscript in Chapter 4 was authored by Jean Monlong, Patrick Cossette, Caroline Meloche, Guy Rouleau, Simon L. Girard and Guillaume Bourque. The thesis author and GB conceived and designed the study. The thesis author performed the analysis, designed the experimental validation with CM and SLG, and wrote the manuscript with GB. PC, GR and SLG provided data and resources for the method validation. CM performed the experimental validation.

The contribution of the thesis author to other manuscripts is described in Appendix A.

Chapter 1

Introduction

1.1 Structural Variation and Copy-Number Variation

1.1.1 Types of Structural Variants

Structural variants (SVs) are defined as genetic variation of more than 50 base pairs. The different canonical forms of SV include deletion, duplication, novel insertion, inversion and translocation[4]. Deletions and duplications of a genomic region, which affect DNA copy number, are collectively known as copy number variants (CNVs). A duplication can be broadly defined as a gain in copy number of a region, either in tandem configuration (tandem duplication) or in a distant locus. In contrast, inversion and translocation are considered balanced rearrangements: no DNA sequence is lost or gain. In reality, small deletion or duplication are often present around their breakpoints[5,6]. Transposable elements retrotransposition creates mobile element insertion (MEI). Because these elements are present in the genome, polymorphic MEI are often considered CNVs. In general, a "novel" insertion involves the insertion of a DNA sequence absent from the genome, e.g. viral DNA, but the term is also used in the MEI literature to describe a new insertion of a transposable element.

Complex SVs involve a combination of canonical forms at the variant level[7]. In a recent study using high-depth long-insert and linked reads sequencing[6], thousands of SVs were found to be complex. Most of these complex SVs (84.4%) involved

inversions, consistent with previous studies that had noticed small deletions and duplications at inversions breakpoints[5,6,8]. More extreme genomic events can create complex SVs that combine dozens of canonical forms and span large regions or several chromosomes. An example is chromothripsis, also called chromosome shattering, which creates a highly fragmented profile with dozens of segments recombined in a different order resulting in a patchwork of duplicated/deleted/inverted regions. While originally though to be rare, recent surveys showed a higher than expected prevalence of somatic and germline chromothripsis. For example, a pan-cancer study found chromothripsis in 38.9% of the glioblastomas and in 8.7% of other cancer types[9]. In the recent study of 689 individuals with autism spectrum disorder and other developmental abnormalities, two cases harbored germline chromothripsis[6].

While SVs are intuitively defined in relation to the ancestral state of the genome, it is important to note that in practice the reference genome is used as baseline. As a result, a variant is a difference in sequence compared to the reference genome but not necessarily compared to the ancestral genome. For example, a recent mobile element insertion might be present in the reference genome but when absent, i.e. in the ancestral state, it is often called a deletion. Similarly, rare deletions of unique regions in the reference genome would resemble novel insertions.

CNVs and in particular deletions have been widely studied. One reason is technological as large CNVs have been routinely studied before the advent of high-throughput sequencing, for example using karyotyping or hybridization approaches (see section 1.2.1). In addition, CNVs, and in particular deletions, are though to have a stronger functional impact compared to balanced variants. A deletion disrupts an entire region and potentially several genes while balanced SVs or insertions might affect only the regions around the variant boundaries or insertion site.

The gain or loss of a full chromosome, also called aneuploidy, is particularly rare in normal cells due to the large phenotypic effects of a dosage change in hundreds to thousands of genes. However, aneuploidy is a hallmark of cancer and observed frequently across cancer types, as described in Section 1.7.2. With whole-genome

doubling, full or arm-level chromosomal CNVs are at the high end of the variant size spectrum.

1.1.2 Mechanism of Formation

The mechanisms of SV formation are diverse and result in a heterogeneous distribution of SV across the genome, both in term of size and location[4,10,11]. New variants can occur during DNA repair, recombination, replication or through retrotransposition.

Non-homologous end joining (NHEJ) is a DNA repair mechanism that often results in deletions. In the presence of double-strand breaks, the two ends slowly denature until the arrival of the repair machinery that joins the two ends. Occasionally, misalignment of the overhanging ends lead to small insertions. Larger sequences can also be incorporated during the repair, leading to large insertions.

Microhomology-mediated end joining (MMEJ) is a type NHEJ which repair the double-strand DNA breaks using micro-homology (5-25 bp) between the broken ends. MMEJ often result in deletions of the sequence between the micro-homology regions but can also create translocation and more complex variants.

Homologous recombination is another repair mechanism that uses a template, usually another chromatid, to repair double strand breaks. By aligning a template, homologous recombination can repair accurately a double strand break even if part of the original nucleotides were lost. Mis-alignment, potentially due to the presence of repeats, results in repair between non-allelic regions and leads to deletions.

Similarly, non-allelic homologous recombination (NAHR) occurs when sister chromatids are not correctly aligned during recombination. Depending on the misalignment configuration, NAHR results in deletion, duplication or inversion. The chromatid misalignment is often caused by the presence of highly similar sequences. Genomic repeats like segmental duplications and transposable elements are frequent templates for NAHR. The majority of NAHR in recent human evolution involved L1 elements although Alus are enriched around older rearrangements[12,13]. NAHR

can also occur during mitosis[14].

Fork stalling and template switching (FoSTeS) occurs during DNA replication when a strand is detached from its current fork and continues replicating in another strand. Depending on the sequence of switches, FoSTeS can result in a translocation, deletion, duplication or inversion.

Slippage during DNA replication can lead to small deletions or duplications, creating and maintaining tandem repeats. Short tandem repeats are particularly susceptible to shrinkage or expansion using this mechanism. While each slippage might only affect a few base pairs, sequential events lead to polymorphic alleles that can differ by hundreds of base pairs between two genomes.

To retrotranspose, a mobile element is first transcribed into a RNA copy which is then converted back to a DNA. The DNA copy then inserts itself at another location of the genome. The DNA sequence of autonomous TEs, such as L1s, code for proteins responsible for the reverse transcription and insertion into the genome. Other TEs use the machinery from autonomous elements to retrotranspose. A similar mechanism is responsible for the insertion and retrotransposition of viral DNA. Once inserted in the host genome, the viral DNA can often copy itself in other genomic locations or in other cells. Similar to retrotransposons, new insertions can then be considered as duplication events.

The mechanism of formation is often inferred from the sequence around the variant boundaries[11]. Segmental duplication or large repeats flanking a variant suggest NAHR. Micro-homology at the boundaries is a sign of MMEJ. No homology points at either NHEJ or FoSTeS.

Aneuploidy arise from problems with the chromosome migration during mitosis. The main mechanism behind arm-level losses or gains are fusion of chromosomes after pericentromeric breakage. Breaks near centromeres can happen in fragile sites, which tend to break under certain conditions, or due to merotelic attachment, an abnormal attachment of sister chromatids during mitosis[15].

1.1.3 Association with Disease and Functional Impact of CNV

CNV and disease Individuals suffering from numerous diseases including obesity[16], schizophrenia[17], autism[18], epilepsy[19], Crohn's Disease[20], cancer[21] and other inherited diseases[22,23], carry SVs with a demonstrated detrimental effect[24,25,26]. First, a few Mendelian disorders are exclusively caused by CNV in specific regions. For example, Williams-Beuren Syndrome which typically presents facial dysmorphies and intellectual disability, is caused by deletions at 7q11.23. As another example, the deletion of the *PMP22* gene is the most common mutation responsible for hereditary neuropathy with liability to pressure palsies. In the early 1990s, Lupski et al. were surprised to find that a duplication in the same region segregated perfectly with hereditary neuropathy Charcot-Marie-Tooth type 1A[27]. The region had been identified using linkage analysis but the idea of a gene-dosage mechanism for the disease was so unexpected that both *Nature* and *Science* refused to review the paper.

CNVs resulting in gene-dosage changes have often milder effects but many have been associated with complex traits or susceptibility to disease. Frequent deletions in the *GSTM1* gene were identified as a risk factor for asthma in independent studies across different populations[28]. Another example of common disease-associated CNV involve the *DEFB4* gene. The median copy number of this gene is 4 in healthy individuals. A lower number of copies has been associated with Crohn disease[29] and higher copy number with psoriasis[30]. Deletions and duplications of the *CCL3L1* gene are also associated with distinct phenotypes. Deletions increase HIV/AIDS susceptibility[31] while duplications increase the risk to develop rheumatoid arthritis[32]. In the examples above, variation in the copy number of the entire gene is affecting the gene dosage resulting in gene expression changes. Although genes with common CNVs are assumed to be tolerant to dosage changes, gene expression tend to change with the number of gene copies in the genome. For example, Handsaker et al.[33] studied multi-copies CNVs and showed that the resulting gene dosage

changes correlated with gene expression.

CNV and gene expression Quantitative trait loci (QTL) and more precisely expression QTLs (eQTLs) are genomic variants that are associated with changes in gene expression. While most of the eQTLs tested and found are single nucleotide variants (SNVs), WGS has allowed the detection of hundreds of SV-eQTLs. Among the first to look for SV-eQTLs, Stranger et al. identified dozens of CNVs in four human populations that were associated with gene expression[34]. Around half of the associated CNVs were located outside of the affected gene or only overlapped partially, hinting at an alternative to the gene dosage mechanism. Later, Lower et al. characterized deletions that affected the expression of a gene located 300 Kbp away, *NMEA*, by analyzing gene expression and using conformation capture to demonstrate physical contact between the two distant regions[35]. Combining their WGS data with RNA sequencing across 462 individuals, the most recent SV catalog from the 1000 Genomes Project identified 54 eQTLs whose lead variant was a SV and 166 additional SVs that were in linkage disequilibrium with SNV-eQTLs[5]. Most of these SV-eQTLs overlapped coding sequence but some were located in non-coding regions upstream of the affected gene. Only 0.56% of the eQTLs were attributed to SV but this number might be an underestimation because of the higher noise in SV calling compared to SNV calling. To improve on this, a recent study used deep WGS to more reliably call SVs and investigated SV-eQTLs in multiple tissues from the GTEx dataset[36]. Using state-of-the-art approaches to infer causal variants, they estimated that 3.5-6.8% of eQTLs could be attributed to SVs. Although less abundant than SNV-eQTLs, SVs had a larger effect size. The comprehensive analysis of the location and effect of these causal SV-eQTLs nicely clarified the relation between SV and gene expression. When overlapping coding regions, SV-eQTLs affected gene expression following the gene dosage model, that is deletion leading to down-regulation and duplication to up-regulation. Non-coding SV-eQTLs, which represented the vast majority of SV-eQTLs (89%), were enriched in or close to regulatory regions (e.g. exons, transcription start site, transcription factor binding sites,

enhancers, gene 3' end) and all types of SV could lead to both higher or lower gene expression. Finally, the effect of rare SVs on gene expression was also explored. Despite the challenge of analyzing rare variants in a cohort of only 147 individuals, a clear enrichment of rare SVs was found around genes that showed outlier expressions in the cohort. These gene-altering rare SVs included cases from both the gene dosage and regulatory region disruption model.

Several recent studies elegantly shed light into a mechanism by which non-coding CNV alter gene expression called enhancer hijacking: that is, a regulatory region inducing the ectopic expression of a gene it normally doesn't regulate because of a CNV-mediated re-positioning. A first example was comprehensively described in individuals with limb malformation[37]. Using conformation capture sequencing and by recreating SVs in mice with CRISPR/Cas genome editing, they elegantly showed that SVs crossing the boundaries of topologically associating domains (TADs) could lead to strong phenotypes. TADs are 3D domains (mean size 830 Kbp) that confine regulatory elements with their targets. The deletions, tandem duplications and inversions resulted in ectopic interactions between a cluster of enhancers and genes located in the neighboring TAD. Ectopic interactions were responsible for ectopic expression of these genes during limb development in mice whose genome had been edited to recreate the SVs. With additional genome editing and conformation capture experiments, this study concluded that the crossing of the TAD boundary was the crucial factor rather than simply the distance between enhancers and genes. Enhancer hijacking might also be important in cancer where a single CNV might lead to a strong expression of oncogenes. To explore this, Weischenfeldt et al.[38] developed a method that detects associations between somatic CNV breakpoints overlapping several TADs and gene over-expression. In their study, Weischenfeldt et al.[38] first described a known cancer gene, *TERT*, which had been already found to be upregulated by such mechanisms. Interestingly, both deletion and duplication resulted in over-expression. They further described two genes, *IRS4* and *IGF2*, using orthogonal experiments to support how the presence of somatic CNVs lead to changes in

chromatin state and physical contact. For example, somatic deletions downstream of *IRS4* overlapped a TAD boundary and resulted in 25-400 fold over-expression of the gene in several cancer types. In contrast, the ectopic expression of *IGF2* was due to single tandem duplication of *IGF2* and a super-enhancer in the neighboring TAD that created a novel chromatin domain with both. Additional experiments showed that the region was active and in contact with the gene promoter in tumor with the duplication. These enhancer hijacking events are important because the change of a single copy can lead to large over-expression. In contrast, dosage effect due to full-gene CNV tend to be as strong as the amplification.

1.2 Whole-Genome Sequencing

1.2.1 SV Detection Before High-Throughput Sequencing

Early cytogenetic techniques were able to detect aneuploidy and extremely large SVs. Thanks to banding, each chromosome in a karyotype can be uniquely identified which facilitated the detection of trisomies and their associated disorder, such as Down syndrome (trisomy 21) and Edwards syndrome (trisomy 18). Furthermore, the bands can be used to identify translocations and large inversions or CNVs. SVs need to span several millions of bases, typically more than 10-20 Mbp, to have a chance to be visible in the karyotype.

Fluorescent *in situ* hybridization (FISH) was developed in the 1980s. Fluorescent probes bind to specific genomic regions by hybridization, i.e. through DNA sequence complementarity. The presence or absence of the DNA sequence was assessed by inspecting the fluorescence in the cells or tissue samples.

In array comparative genomic hybridization (aCGH) experiments, DNA from a test sample and reference sample are labeled using different fluorophores and hybridized to several thousand probes. The probes, which usually tag most of the known genes and tile non-coding regions of the genome, are printed on a glass slide. The fluorescence of each probes is used to estimate the amount of DNA sequence in

the test sample compared to the reference sample. Using this method, CNV down to approximately 100 Kbp of DNA sequences can be detected. Arrays also can be designed specifically to target regions of interest, for example with recurrent CNVs. These custom arrays don't cover the genome uniformly but can detect smaller CNVs in the regions with of high probe density. This technology is not able to detect balanced chromosomal imbalances such as translocations or inversions.

1.2.2 A New Hope

While large SVs have been identified by cytogenetic approaches and array-based technologies, whole-genome sequencing (WGS) could in theory discover SVs of all sizes or types[39]. The vast majority of studies follow a re-sequencing strategy where short DNA fragments (or reads) are sequenced and aligned (or mapped) to the reference genome. Furthermore, both ends of a DNA fragment are often sequenced and this pair information can be used to improve alignment to the reference genome and variant calling. The reads and their alignment are then used to find single-nucleotide variants (SNVs), small insertions/deletions (indels) but also small SVs across the genome. Array-technology required dense representation of the hybridization probes in a region of interest to be able to detect CNVs smaller than 100 Kbp. With WGS, the sequencing depth is now the main limiting factor, although even early experiments could detect thousands of small SVs. For example, the most recent survey of the 1000 Genomes Project used WGS with a sequencing depth of 7x and identified more than four thousands variants per individual with a median variant size below 40 Kbp for the six different SV types analyzed[5]. In contrast to aCGH, the sequencing reads can also be used to detect balanced variants such as inversions, translocations and novel insertions. Although the detection of such variants is more challenging than single-nucleotide variant (SNV) calling, WGS is a one-fit-all experiment that greatly increases the resolution of SV detection.

To detect SVs from WGS, methods analyze either read-depth (RD) variation[40,41,42], paired-end information[43,44], breakpoints detection through split-read approach[45] or

de novo assembly[46]. Methods are described in more details in section 1.3, with a particular focus on CNV detection.

Another unique aspect of WGS is the possibility of pooling experiments to increase the detection power of common variants. Instead of analyzing each experiment separately, the sequencing reads can be pooled across several samples. For example, it is sometimes challenging finding several reads spanning a SV breakpoint within a single sample. By pooling several experiments, the number of supporting reads increases if a SV is shared by several samples. This approach was used across hundreds of samples of the 1000 Genomes Projects and greatly increased the number of SVs discovered in the population[11,47].

1.2.3 The Technical Bias Strikes Back

Although it represents a considerable improvement in term of resolution, WGS is affected by technical biases that remain an important challenge. Indeed, it has been shown that various features of sequencing experiments, such as mappability, GC content or replication timing, have a negative impact on the uniformity of the coverage[48,49,50,51,52]. In addition to its effect on read coverage, repeated sequences lead to confusion in read mapping, creating SV-like patterns and thus false-positives when calling variants.

GC content is a well-known source of bias although not completely understood. Reduced efficiency of PCR amplification explains a large fraction of this bias and more robust protocol were proposed[53,54]. Still other steps of the sequencing protocols adds substantial bias and GC bias persists even with optimized protocols[53]. The bias patterns tend to be different from a sequencing center to the other[50], suggesting an effect of the library preparation or sequencing machinery. For these reasons, it has been challenging to correct for this source of bias. With PCR-free libraries, the effect of GC bias is reduced but still needs to be corrected for when comparing read coverage across the genome.

DNA replication also affects the distribution of reads across the genome. Al-

though sequencing of bulk samples, i.e. of many cells, should minimize the effect of replication patterns, systematic increase might be present in regions that tend to replicate earlier. For example, Koren et al. estimated replication timing across the genome using WGS of cells in S and G1 phases[55].

Finally, the mappability of the sequence affects how many reads can be confidently mapped to the reference genome. The presence of repeats and other similar regions lead to multi-mapping, i.e. several positions where a read could have originated from. Hence, when using reads with unique mapping in the genome, the coverage in repeat-rich regions drops considerably. The challenges and proposed solutions associated with mappability are described in more detail in section 1.4.

Unfortunately, the variability in term of read distribution is difficult to model and correct for because it involves various factors, including some that vary from an experiment to another and others that are still unknown. This issue particularly impairs the detection of SV supported by weaker signal, which is inevitable in regions of low-mappability, for smaller SVs or in cancer samples with stromal contamination or cell heterogeneity.

1.2.4 The Return of the Long Reads

Sanger sequencing, invented in 1977[56], was used for the original sequencing of the human genome[57] and is still used today to sequence DNA fragment 500 to 1000 bp long. The technology that followed in the 2000s is capable of sequencing shorter reads but much more efficiently resulting in a cost order of magnitudes lower. However, many of the challenges faced by WGS is a result of the short size of the sequenced read. Recently, new technologies have been developed to perform WGS using much longer reads, in the range of 10-100 Kbp. PacBio was the first and has been successfully applied to several human genomes[58,59,60]. Nanopore sequencing is becoming efficient with the first human WGS sample just released publicly[61]. Although the cost and rate of sequencing errors remains high compared to short-read sequencing, the benefit for genome assembly or SV detection is clear.

1.3 Existing CNV Detection Methods

1.3.1 Different Strategies to Detect SV and CNV

The vast majority of SV detection methods rely on evidence from the read mapping on a reference genome: changes in read depth, B-allele frequency, discordant paired-end mapping, or split-reads. *De novo* genome assembly could also be used to identify SVs but its application using short read sequencing remains challenging.

Read depth Changes in the copy number in a region should lead to changes in the number of reads mapped to this region in the reference genome. By modeling read depth, sometimes called read coverage or depth of coverage, one approach is to identify regions with significantly more reads (duplication) or fewer reads (deletion) than in the genome or the flanking regions. Only CNV, i.e. imbalanced SVs, can be detected by these approaches. CNV detection methods that use read depth are described in more details in the next section.

B-allele frequency The proportion of reads supporting heterozygous SNVs can help identify CNVs too. The loss of heterozygozity within deletions or the deviation from the 50% coverage of the alternate allele can complement coverage signal. This approach was inspired from CNV detection strategies developed for SNP-array, such as in the ASCAT method[62]. Here the intensity of the probe and the so-called B-allele frequency was use in concert to call CNVs. Thanks to sequencing, both the coverage information and SNVs are more densely represented and lead to a better resolution. Still, the B-allele information is relevant only for CNVs large enough to span several heterozygous SNVs. Methods such as ERDS[63], Control-FREEC[64] or Sequenza[65] integrate the B-allele frequency information to call germline or somatic CNVs.

Paired-end mapping The distance between the two mapped reads in a pair and their orientation can also help identify SVs. Because the majority of the reads are

expected to map correctly, they can be used to estimate the expected distribution of the distance between paired read. With this distribution, one strategy is to retrieve pairs that are significantly too close or too far from each other. Read pairs might map close to each other in the reference genome because of an insertion in the sequenced DNA somewhere between the reads. More typically, reads that map far from each other suggest a deletion in the sequenced DNA or potentially a translocation. Finally, tandem duplications or inversions should lead to some read pairs mapping in the incorrect orientation relative to each other. All those reads with discordant paired-end mapping are typically retrieved and clustered together. Each cluster of reads is then disentangled to predict the most likely variant and the location of its breakpoints.

Split-reads The strategies described above use either reads within the variant or around the variant's boundaries. In contrast, the split-read approach looks for reads exactly spanning a variant's breakpoint. Generally, one read is mapped uniquely to the genome and serves as an anchor while its pair is split in two pieces which are then aligned separately. This split-mapping can be computationally expensive. To limit the computational cost, methods analyze only pairs with one unmapped reads or restrict the range searched for the split-mapping[45,66]. Split-reads can be searched specifically to complement candidates variants identified from discordant paired-end mapping. These additional supporting reads are tallied and used to assess the final supporting evidence in methods such as LUMPY[67] or DELLY[68].

Assembly Local assembly of reads around candidate variants has been used as in silico validation and to characterize the breakpoint sequence[69,70]. Going further, recent methods have been using local read assembly as their main SV detection strategy, especially in cancer[71,72]. If *de novo* assembly keeps improving, for example thanks to longer reads, SV could also be called by directly comparing assembled genomes or to the reference genome. For example, Assemblytics has been recently developed to align two assembled genomes and to annotate SVs that differentiate

them[73].

1.3.2 CNV Detection Using Read Depth

Single-sample methods The first methods that used read-depth signal to call CNVs assumed a uniform read coverage across the genome and attempted to segment it along the chromosomes. The segments produced by these approaches represent regions with similar copy number. The circular binary segmentation, adapted from aCGH analysis[74], was one of the first and remains a popular segmentation algorithm.

Subsequent methods offered better correction of technical biases and more modern segmentation techniques. For example, CNVnator[41] corrects for the GC bias, masks repeat-rich content and uses a mean-shift segmentation approach inspired from the image recognition field. CNVnator has been used extensively in both germline and somatic CNV surveys[5,36]. FREEC[40] is another popular approach that can correct for both GC bias and mappability using precomputed tracks. It segments the corrected read-depth signal with a LASSO-based segmentation approach. FREEC has been extended to Control-FREEC to include the B-allele frequency in its CNV detection process.

Methods inspired from aCGH offer to use another sample as control. Although variants in the control sample might create problems down the line, it is particularly sensible when studying tumors whose tumoral and normal tissues has been sequenced. By using the normal sample (usually blood) as control, the CNV detection is naturally reduced to the detection of somatic CNVs, i.e. present in the tumor but absent from the normal sample. In practice the methodology is similar to the single-sample approaches described above but using the read-depth ratio of the tumor versus normal tissues. A few methods have further been implemented specifically with cancer in mind and estimate the tumor ploidy and/or stromal contamination before or during CNV calling[65,75].

Multi-samples methods To improve the sensitivity of the variant detection and model the region-specific pattern of read depth, a few methods have been developed to jointly analyze multiple samples together.

cn.MOPS considers simultaneously several samples and detects copy number variation using a Poisson model and a Bayesian approach[42]. By jointly analyzing samples, cn.MOPS calls variants based on the strength of the read-depth signal across the samples. Even if the signal-to-noise ratio is small, the presence of a consistent pattern in several samples provides further evidence that the region contains a CNV in those samples.

The second version of GenomeSTRiP models the read depth across hundreds of samples as a mixture of Gaussian distributions[33]. It is particularly useful to genotype multi-copies variants in the genomes, i.e. regions that have more than two copies in most individuals. Multi-copies variants create different groups of samples that translate into different RD distributions. By deconvoluting the mixture of distributions, the relative difference between RD modes help associate each distribution (and sample) to a copy-number estimate. As it relies on full signal (i.e. around integer values in the model) and a simple read-depth normalization, it is still limited in regions with low coverage and for small or rare variants. The power to detect a variant increases with the frequency of polymorphisms as more individuals populate the different genotype groups, improving the copy-number estimation. GenomeSTRiP 2.0 was successfully applied to 849 individuals from the 1000 Genomes Project and unmasked the population variation of hundreds of multi-copies variants.

Both cn.MOPS and GenomeSTRiP use the additional samples to find further support for a variant, which is particularly efficient at detecting common CNVs. However, both methods define models with full copy number changes which has limited power when dealing with CNVs with partial signal (e.g. small variants, somatic variants, or variants in low-mappability regions). In contrast, the approach described later in this work uses the multiple samples differently: they are used to define a baseline for the technical variation, not to aggregate evidence supporting

a variant. When the coverage diverges enough from this baseline, no matter the frequency or the strength, a CNV is called. In theory such approach will be able to better detect rare variants, small variants, somatic variants and variants in low-mappability regions.

These approaches are also called population-based methods because they jointly study a population sample of sequencing experiments, i.e. a group of samples representative of the technical variation across experiments[33]. Of note, the term population in this context refers to a statistical population rather than human populations.

1.4 Low-Mappability Regions

1.4.1 The Different Classes of Human Repeats

1.53 Gbp of the human genome is annotated as a repeat when considering elements identified by Repeat Masker[76] and segmental duplications. Repeats are classified based on their size, sequence and mechanism of formation.

Segmental duplications (SDs) are large regions (>1 Kbp) with high similarity (>90%). Usually the results of NAHR, segmental duplication are known hotspots of structural variation. SDs can be nested, i.e. duplications within duplications. These class of repeat is thought to have boosted recent human evolution[77]. Humans experienced a high rate of SD creation in recent evolution which contributed to the expansion of important gene families. Large families of genes involved in immune response, cell adhesion and brain development cluster within SDs.

Transposable elements (TE) represents approximately 45% of the human genome. TEs are interspersed in the genome: if we cut the reference genome in consecutive windows of 500 bp, 70.5% of the windows would overlap transposable elements. Their wide distribution make them a popular template for NAHR. Some TE families tend to cluster in fragile regions of the genome, which are regions that tend to break under certain stress conditions. A small fraction of TEs of the Alu, L1 and SVA families are still active in the human genome[78]. Retrotransposition of these

elements contributes to novel MEI.

Satellites consist of sequences that are repeated, most of the time in a tandem configuration, and span large regions. The size and composition of the repeated sequence, also called unit, define the different classes of satellites. Most of the macro-satellites are present close to centromeres, the main family being alpha satellites whose 171 bp long unit is repeated to span on average 5.6 Kbp. The sequence of the unit varies from chromosome to chromosome. In addition to the tandem repetition of the unit sequence, higher order structure is present. The sequences can be duplicated in the same orientation or in an inverted conformation.

Short tandem repeats (STRs), also called micro-satellites, have sequence units ranging from 1 to 10 bp. In addition to the tandem duplication that they share with SDs and satellites, short tandem repeats can vary because of slippage during the replication process. As a result, STRs are one of the most polymorphic class of variant in the human population, making them particularly useful for forensics and parentage tests. STRs have been recently linked to gene expression regulation[79,80] and potentially to polygenic disorders[81].

Low complexity regions are regions with high AT or GC content that, unlike satellites, have no apparent structure. Very little is known about their mechanism of formation and variation. Longer stretches of similar sequences might be present by chance in those regions which might promote homology-based mechanisms of CNV. Low complexity sequences might also favor secondary DNA structure, promoting replication slippage.

1.4.2 Impact of Repeats on Mappability

The presence of repeats can confuse read mapping and decrease the number of uniquely mapped reads in certain regions. These low-mappability regions can contain any class of repeats, e.g. segmental duplication, transposable elements, short tandem repeats or low-complexity sequences. The effect of repeats on the mapping of a read is specific to each region: it depends on the density and nature of the

repeats present. As a result, this technical variation is difficult to model. Existing methods either remove the signal in these regions or smooth the signal to avoid spurious variation (see Tackling the Repeat Challenge).

Furthermore, the multi-mapping of reads between similar repeat instances resembles signal supporting certain SVs. For instance, translocation are supported by pairs of reads mapped far from each other. When one read of the pair spans a repeat, it is sometimes aligned to another repeat element in the genome, far from its paired read. To minimize this issue, read aligners could favor configurations where both pairs map together but forcing read pairs to map together would impair the detection of real variants. In practice, some aligners output the best alignment as well as other secondary alignments where the reads could have originated from. This multi-mapping information or other mapping metrics are used by SV callers to identify legitimate variants or flag others that could be caused by mapping confusion. Nonetheless, even with paired-end aware mappers many reads overlapping repeats show this incorrect mapping and can hinder SV calling. Because a low-mappability sequence highly resembles another sequence in the genome, a variant or a sequencing error in the read might lead to a better (although incorrect) mapping in the incorrect location. This multi-mapping confusion can also occur locally and incorrectly support other types of SVs such as deletion and tandem duplications.

To minimize the effect of multi-mapping, algorithms tend to use uniquely mapped reads only. Fewer reads support the presence of a variant but they are of better quality. In some cases, repeats flank a unique region and can mask potential CNVs from paired-end or split-read approaches that use uniquely mapped reads only. Indeed, because of the repeats, reads around the breakpoints that would support the CNV don't map uniquely. If the flanking repeats are long and highly similar, these CNVs can only be detected by changes in the read coverage.

1.4.3 Tackling the Repeat Challenge

Because of the mappability and other technical biases, existing approaches suffer from limited specificity and sensitivity[11,39], especially in specific regions of the genome, including regions of low-complexity and low-mappability[48,49].

Approaches that use paired-read information and split-read mapping are difficult to modify to deal with the presence of repeats. Oftentimes, repeated regions or low-confidence mapping are simply filtered (or flagged) when calling SVs. The integration of the multi-mapping information could always be improved but the mapping patterns are often region-specific and difficult to model. Despite the challenges, some attempts were made for specific types of variants. For example, Hormozdiari et al.[82] modeled transposon insertion and He et al.[83] proposed a way to handle the multi-mapped reads when searching for tandem duplication.

Approaches relying on read coverage are relatively more robust because they use signal across the whole variant rather than the breakpoint regions. A deletion or duplication between repeated sequences might be difficult to call confidently using paired-end or split-read information but the change in RD across a variant is less affected by repeats around the breakpoints. The presence of repeats along the entire variant still remains a challenge for RD approaches. To deal with regions of high repeat content, repeats were originally masked before CNV detection to avoid problems from multi-mapping of the sequencing reads[84,85]. Another approach used bins of variable length designed specifically to provide uniform coverage of uniquely mapped reads across the genome[74]. For example, a region with repeats was extended as much as necessary to contain, on average, a similar read coverage as in unique regions. While it simplified the methodology of the CNV calling, it was not ideal. First, repeat-rich regions often remained problematic and were still specifically filtered out of the output. Indeed, repeat-rich regions are not only less covered by uniquely-mapped reads but also more variable. The variable-length bin design adjusted the mean read coverage but the repeat-rich regions remained more variable and challenging for most segmentation approaches. Second, the contribution of

the repeat-rich regions in the extended bin becomes minimal because of their low coverage. The repeats are not masked but they are not represented by the RD of these regions either. Although it helps mitigate the effect of repeats, it doesn't help address variation in repeat-rich regions.

Other methods keeps repeats unmasked and use bins of equal size but transform the coverage signal to reduce the unwanted effect in repeat-rich regions. For example, CNAseg[86] and CNVnator[41] use a smoothing step that reduce the effect of outliers and smooth the signal using flanking regions. Similar to the variable-length bin strategy, the signal in the repeat regions is traded off for a easier calling and segmentation. After smoothing, CNVs in repeat-rich regions but also small CNVs might become invisible.

In an attempt to tackle the problem at its roots, Alkan et al. developed a read aligner to better deal with reads mapping to several locations in the genome[84]. Using mrFAST they were able to better detect and genotype copy-number variants in some large and highly similar segmental duplication. They showed that alignment could be improved for many segmental duplications regions to the point where accurate CNV detection was possible. However, this effort cannot be replicated for the vast majority of the repeat-rich regions of the human genome. Alignment algorithms performance in low-mappability regions are now mainly limited by the size of the sequencing reads.

1.4.4 Disease-Associated CNV in Repeats

CAG repeat expansion in a coding region of the *HTT* gene causes Huntington disease in a dominant and fully penetrant manner[87,88]. When the short tandem repeat is large enough, typically larger than 36 units, the mutant protein is responsible for an increase in the decay rate of neurons. Fragile X syndrome is caused by the expansion of a CGG repeat in the 5' untranslated region of the *FMR1* gene[89]. The length of the repeat varies from 15 to 60bp (20 units) in the healthy population while repeats larger than 600 bp cause the disease. Repeats with intermediate

size increase the disease risk for the offsprings. ICF syndrome (Immunodeficiency, Centromeric instability and Facial anomalies) is characterized by an extension of pericentromeric satellites. Facioscapulohumeral muscular dystrophy was associated to the contraction of a satellite DNA in the sub-telomeric region of chromosome 4[90].

CNVs involving repeats are also widespread in cancers. L1 retrotransposition is a common phenomenon in some cancer types such as epithelial tumors[91,92,93]. The first instance of disruptive insertion was documented in the tumor suppressor gene *APC* in colon cancer[94]. Microsatellite instability is important in some cancers, such as colorectal and endometrial cancers, and usually coincides with the disruption of DNA repair mechanisms[95]. It results in extensive copy number changes in microsatellites. In colorectal cancers, micro-satellite instability is typical of a specific sub-group, Lynch syndrome, representing around 15% of cases and associated with better prognosis[96]. Fragile sites are also enriched in somatic SVs. Furthermore, fragile sites are often unstable in cancer and are enriched in low-complexity sequences and satellites[97,98]. Transposable elements tend to cluster in these regions as well, sometimes taking advantage of the DNA breaks to insert new copies. Satellite instability and increased retrotransposition suggest that repeated region might be more fragile or more variable than in normal genomes.

In addition to variation in the repeat sequence, some repeated regions favor the formation of CNVs. For instance, segmental duplications and TEs provide templates for NAHR. Alu or L1 elements are the most abundant and most frequent templates. Alu-mediated deletions in the *LDLR* gene were among the first to be described in a patient with familial hypercholesterolemia[99]. A recombination event between HERV-I copies has been linked to male infertility by causing a ∼800 Kbp deletions containing the azoospermia factor gene on chromosome Y[100]. Cancer genes such as *MLL-1*, *VHL* and *BRCA1* seem to be experiencing CNVs resulting from NAHR between Alu elements[101]. Alu-mediated recombinations around *MSH2* are responsible for germline deletions that have been linked to susceptibility to hereditary nonpolyposis colon cancer[102]. CNV can be a byproduct of TE insertion as well.

For example, the insertion of a L1 lead to a 46 Kbp long pathogenic deletion in the *PDHX* gene in an individual suffering from PDHc deficiency[103].

1.5 CNV Distribution in Normal Genomes

In healthy individuals, a higher proportion of the genome is estimated to be affected by SVs as compared to single nucleotide polymorphisms (SNPs)[104]. Several databases and studies have cataloged CNVs in the human genome and described their distribution. The CNV enrichment in segmental duplications has been extensively documented.

1.5.1 Public CNV Catalogs

The long-standing database for structural variation in healthy individuals is most likely the Database of Genomic Variants (DGV). It aggregates findings from more than 55 studies and annotates more than 200,000 different regions[105]. Although it represents the largest aggregation of variants, DGV should be used carefully. For example, the variant information, such as frequency, breakpoint resolution or genotype comes from each original study and might not be directly comparable. Variant frequency is particularly important for disease studies but the frequency in DGV might not be representative of the population frequencies. Indeed, the different studies used different technologies, some of which might not have the resolution to detect a variant of interest. The low sample size of some studies might also inflate the frequency estimates.

A few studies using aCGH across hundreds of healthy individuals provided a more representative distribution of large CNVs in the population. Redon et al. found that a larger than expected fraction of the genome was affected by CNVs, cumulatively affecting more bases than single nucleotide polymorphisms[106]. They described CNVs across four human populations and their overlap with genes, disease loci, functional elements and segmental duplications. Using arrays with higher density and a larger

sample size, Conrad et al. described similar patterns for common and rare CNVs in the human genome[24].

Using high-throughput sequencing, large-scale projects were able to catalog unbalanced types of SVs and CNVs at a better resolution. The 1000 Genomes Project was the first to produce such a catalog, analyzing 179 individuals across 4 human populations[11]. Its most recent catalog[33] analyzed 2,504 individuals from 26 populations and contains 42,279 deletions, 6,025 duplications, 16,631 mobile element insertions and hundreds of other SVs types such as inversions and translocations. In this project, nine different methods were combined in an ensemble approach in order to detect different types of variants and increase the confidence of each call. Extensive low-throughput validation was used to decide how to combine the output of the different methods and ensure high quality calls. Such a strategy increases the specificity of the variant detection but is less sensitive. In order to study a large number of individuals in a cost-effective manner, the sequencing depth of most experiments was kept low, around 7x. With these settings, the frequency estimates of the variants that could be detected are accurate but some types of variants might have been missed, for example rare variants, small CNVs or variants in low-mappability regions. In this study, as in many others, repeat-rich regions and other problematic regions were masked or smoothed at some step of the analysis to produce more accurate calls. In a follow-up study on the 1000 Genomes Project data, Handsaker et al.[33] studied the allele distribution of multi-copies CNVs across individuals. They identified 185 genes which overlapped with CNVs that were present in more than two copies in most of the individuals. They also describe variants that show specific distribution patterns in the African population, possibly because of different evolutionary constraints. A similar ensemble approach was used by the Genome of Netherlands (GoNL) consortium to analyze 750 individuals sequenced with medium depth of 13x[107]. 10 methods were combined and provided a SV and CNV catalog of the Dutch population.

Recently, long-read sequencing in on haploid cell lines[58] and a diploid human

sample[59] provided a better survey of SV across the genome. Challenging repeats are more often spanned by the long reads and many low-mappability regions could be analyzed. Due to its higher cost, few genomes have been sequenced yet. Nonetheless, a large fraction of the SV identified were novel. As expected, most of the novel variants were located in regions of low-mappability. Although devoid of population estimates, these catalogs contains hundreds of SV in low-mappability regions.

1.5.2 Enrichment in Segmental Duplication

The enrichment of CNVs in segmental duplication has been described early on in the first genome-wide array-based surveys[108,109]. Redon et al.[106] found that 24% of the 1,447 CNVs identified overlapped with segmental duplications. Wong et al.[110] noted a 5.7-fold enrichment of common CNVs in segmental duplications. Around 50% of the annotated segmental duplications are covered by CNVs in the Database of Genomic Variants[105]. Segmental duplications cover only ∼5.8% of the reference genome and the enrichment of CNVs has been replicated over the years with different technologies and resolution[11,58,84].

Several CNV hotspots are also located within segmental duplications[111]. These regions rearranged during recent human evolution and continue to experience copy number changes. Some of these hotspots, for example 15q13.3, 16p11.2 or 1q21.1, have been associated with diseases, particularly neurological disorders[112].

What creates this enrichment? As mentioned earlier segmental duplications are often templates for NAHR. Some segmental duplications might not be fixed yet and remain polymorphic in the population. Segmental duplications might also be more permeable to variation in general, as suggested by the higher substitution rate[113]. These regions might also be more fragile, i.e. more likely to experience DNA breaks that could then lead to SV.

1.6 CNV in Neurological Disorders and Epilepsy

1.6.1 CNV and Neurological Disorders

As mentioned previously, segmental duplications tend to sensitize some genomic regions to deletions and duplications, forming CNV hotspots regions. A subset of these hotspots have been associated with a number of neurological disorders such as autism, mental retardation, schizophrenia and epilepsy. In general, the affected genomic regions is unique, spans 50 Kbp to 10 Mbp and is flanked by long (>10 Kbp) and highly homologous (>95%) segmental duplications. For example, 16p11.2, 1q21.1 and 15q13.3 CNV hotspots have all been associated with autism, mental retardation and epilepsy[112].

An interesting case was described by Koolen et al. and involved an inversion that "activates" a CNV hotspot[114]. In an aCGH screen, Koolen et al.[114] identified recurrent *de novo* deletions in 17q21.31 in cases with mental retardation but not in controls. The region was flanked by inverted segmental duplications and an inversion needed to happen first in order to promote a NAHR-mediated deletion. The corresponding 900 Kbp inversion is present in around 20% of Europeans and predispose carrier for the deletion associated with mental retardation.

Rare CNVs, particularly deletions larger than 1 Mbp, have been associated in a number of neurological disorders. Early on, mental retardation and other developmental delay disorders have been linked to large CNVs which could explain >15% of the cases[112]. In schizophrenia, rare genic CNVs larger than 100 Kbp were present in 15% of cases, 3 times mores than in controls[115]. *de novo* CNVs were significantly associated with autism in a cohort of 118 patients and 196 controls[116]. 10% of the patients with sporadic autism had a *de novo* CNV versus only 1% of the controls. In a large cohort of 996 individuals with autism spectrum disorder, Pinto et al.[117] observed a higher burden of rare genic CNVs.

1.6.2 CNV and Epilepsy

Epilepsy is a neurological disease characterized by seizures. It has a prevalence of 1% in the population, and a lifetime incidence up to 3%. The phenotypes of epilepsy can be complex, and there are several types of epilepsy. In *generalized* epilepsy, sometimes called idiopathic or primary generalized epilepsy, the seizures affect the whole brain. *Absence* epilepsy, a sub-type of generalized epilepsy, is characterized by brief loss and return of consciousness. In contrast, patients with *focal* or partial epilepsy experience partial motor seizures. Focal seizures are more prevalent in children and teens and occur mostly during sleep. In the case of *epileptic encephalopathy*, affected individuals also exhibit severe cognitive and behavioral disturbances.

Both familial and sporadic cases of epilepsy seem to have a genetic component. The vast majority of the genetic variants associated with epilepsy comes from studies on one or just a few cases. Aggregating information across 818 studies, Ran et al. [118] compiled a list of genes associated with epilepsy. 154 high-confidence genes were identified from the recurrence and predicted impact of more than 3931 variants. For a gene to be included in the list, several losses of function and/or *de novo* variants had to be described in the literature.

As in other neurological diseases, the phenotypic heterogeneity remains a challenge when combining cases into large studies. In addition, incomplete penetrance or variable expressivity have been observed and complicates the identification of genes and pathogenic variants. Even in locus that are clearly associated with epilepsy risk, there are examples of unaffected carrier parents while other times, the same single-gene mutation can cause a wide range of seizure types[119].

Recurrent microdeletions were identified in up to 3% of patients with idiopathic generalized epilepsies and in 1% of focal epilepsies[120]. In Heinzen at al., 15q13.3 and 16p13.11 were the CNV hotspots with the most frequent microdeletions[121]. These variants were often transmitted from a healthy parent. Another study of 517 individuals with mixed types of epilepsy revealed that 2.9% of the patients had a deletion in the 15q11.2, 15q33.3 or 16.q13.11 hotspots[119]. In Mefford et al. [19], 315

patients with epileptic encephalopathy were also screened with aCGH, of which 1.6% had a CNVs in 16p11.2, 22q11 and 15q13.3 hotspots.

In addition to CNVs in hotspot regions, several studies observed a significantly higher number of rare large CNVs in epilepsy patients. In the discussion, Mefford et al.[19] mentioned a significant excess of large deletions in their cohort compared to the controls. 2.2% of the patients had rare CNVs larger than 1 Mbp while only 0.3% of the controls. However, the controls came from two independent studies that used different arrays. Using epilepsy cases and matched controls, Striano et al.[122] found no difference in term of number of rare CNVs between epilepsy patients and controls but that CNVs were significantly larger and affected more genes in the epilepsy patients. Other studies of CNV in epilepsy tend to focus on rare CNVs in the cases and only use controls to filter variants. Using this approach, up to 10% of epilepsy patients carry a unique and possibly pathogenic CNV. This numbers varies depending on the type of epilepsy studied and the additional criteria used to filter CNVs (Table 1.1). In Mefford et al.[119], 8.9% of the 517 patients with generalized and focal epilepsies had one or more rare genic CNVs that was absent from their 2493 controls. A similar study on epileptic encephalopathy found that 7.9% of the 315 patients had a rare genic CNV never seen in 4,519 controls, half of which were classified as pathogenic or likely pathogenic[19]. Deletions of known epilepsy genes or *de novo* deletions were considered pathogenic while *de novo* duplications and CNVs larger than 1 Mbp were considered likely pathogenic. In a more clinical setting, CNVs detected from aCGH could explain the phenotype of 5% of the 805 epilepsy patients screened[123]. The pathogenicity of the variants were assessed by clinicians based on the size, inheritance, gene, hotspot overlap and concordance between the clinical symptoms and the ones reported in the relevant literature. Another study on childhood epilepsies identified a large rare CNV in 71 patients out of 222, 33 of which were considered pathogenic or likely pathogenic based on the inheritance pattern or overlap with known pathogenic variants[120]. *de novo* variant are often of interest and 11 *de novo* CNV were identified in this study. All four variants larger

than 3 Mb were *de novo* CNVs. More recently, the exome of 349 trios with epileptic encephalopathy identified 18 *de novo* CNVs in 17 patients (4.8%), 10 of which were classified as likely pathogenic[124].

Study	Epilepsy type	rare CNVs	CNVs in hotspots	Sample size	CNV size	Filters
Mefford et al.[119] (2010)	generalized & focal	8.9%	2.9%	517	1.2 Mbp	rare genic
Mefford et al.[19] (2011)	epileptic encephalopathy	7.9%	1.6%	315	2.26 Mbp	rare genic
Striano et al.[122] (2012)	generalized & focal	9.3%	3.4%	265	1.9 Mbp	rare
Olson et al.[123] (2014)	clinical epilepsy	5% pathogenic	2.9%	805	18 Kbp - 142 Mbp	-
Helbig et al.[120] (2014)	childhood	31.9%	4.9%	222	102 Kbp - 12.7 Mbp	rare genic >100 Kbp
Mefford[124] (2015)	epileptic encephalopathy	4.8% *de novo*	-	349	377 Kbp	rare
Addis et al.[125] (2016)	absence	10.4%	2.8%	144	26 Kbp - 2.8 Mbp	rare genic >20 Kbp

Table 1.1: **CNV surveys of epilepsy patients.** The second and third columns represent the proportion of cases with a rare CNVs or a CNV in a known hotspot region, respectively. The *Sample size* represents the number of epilepsy cases in each study. The *CNV size* shows the average size of the CNVs detected (or the size range).

1.7 Somatic CNV in Cancers

1.7.1 Methodological Challenges

Contamination of a sequenced tumor by normal cells reduces the strength of the signal from somatic variants. For CNV, the copy number changes seem only partial because only a fraction of the cells share the variant. Cellular heterogeneity further reduces the strength of the CNV signal for the same reason. Hence, if the purity is too low or the somatic variant is present in a minor clone, the signal-to-noise ratio is reduced compared to germline variants. A higher ploidy can also result in a weaker CNV signal. For example after genome doubling, a one-copy deletion corresponds to a reduction of one quarter of the average coverage.

One strategy is to estimate the purity and ploidy in the sample before or at the same time as the CNV calling. Using information about the coverage deviation and B-allele frequency changes, methods such as Sequenza[65] and TITAN[75] can predict the most likely ploidy and purity of the tumor sample and adjust copy number estimates. TITAN further model cell heterogeneity in the tumor cells by testing the presence of minor clones with CNV signatures.

1.7.2 Chromosomal Aberrations and Aneuploidy

Somatic CNVs (sCNVs) are common in tumors from almost all cancer types[21,126]. Tumors sometimes harbor whole-genome doubling or, more frequently, chromosome arm-level gains or losses. Aneuploidy, i.e. chromosome gain or loss, is seen in almost all cancer types although at varying frequencies[126,127,128].

Aggregating aCGH data across 5,918 epithelial tumors, Baudis[127] identified frequent arm-level gains in 8q, 20q, 1q, 3q, 5p, 7q and 17q; and frequent losses in 3p, 4q, 13q, 17p and 18q. The median number of arm-level aberration per sample ranged from 0 in squamous skin neoplasias to 12 in small cell lung carcinomas[127]. To further interrogate the distribution of somatic CNVs in different types of cancer, Beroukhim et al.[21] analyzed 3,131 cancer genomes from 26 different cancer types. Although the frequency of sCNV decreases with their size, arm-level aberrations stood out as recurrent aberrations for both losses and gains and across all 26 cancer types studied. Arm-level aberrations were around 30 times more frequent than expected from the frequency-size relationship of other sCNVs. In this pan-cancer survey, 25% of a typical cancer genome was affected by arm-level sCNVs. In contrast to some focal sCNVs, arm-level sCNVs resulted in low-amplitude changes, most of the time a single copy loss/gain. Interestingly, it appeared that chromosome arms had either somatic losses or gains, but rarely both. Similar observations were made in a larger meta-analysis of 8,227 cancer CNV profiles from 107 aCGH studies[128]. Chromosome arms preferentially showed losses or gains, with 1q, 5p, 7p, 3q and 20q most frequently gained and 4q, 6q, 8p, 13q and 17p most frequently lost. Cancer types were clustered based on their arm-level aberration frequencies and formed three groups with similar developmental origins. Arm-level aberrations in gene-rich arms tended to co-occur when concordant (gain-gain, loss-loss)[128]. Using a single detection platform across 4,934 cancers and 11 cancer types, The Cancer Genome Atlas attempted to time the occurrence of different types of sCNVs[126]. For example, whole-genome doubling was observed in 37% of the tumors and tended to occur prior to other types of sCNVs. The rate of both arm-level and focal sCNVs were higher in

tumor that experienced whole-genome doubling. In term of arm-level aberrations, Zack et al.[126] found a median of 3 gains and 5 losses per tumor.

Some arm-level aberrations have been linked to worse prognosis[129,130]. In some cases, the arm-level is thought to directly affect cancer drivers. For example, the frequent loss of chromosome 9p is associated with more aggressive tumors and poor survival through down-regulation of tumor suppressors[130]. These aberrations can also be used as markers for prognosis prediction.

The pan-cancer studies described above didn't include or comment on sex chromosomes in their analysis[21,126,127,128].

1.7.3 Gender Imbalance

According to cancer statistics for 2017, the lifetime probability to be diagnosed with invasive cancer is 40.8% for men and 37.5% for women[131]. Several cancer types show the same gender imbalance with different incidences for males and females. For example, incidence of liver cancer is three times higher in men than in women, and even more for esophagus, larynx and bladder cancers. In contrast, the incidence is higher in women for cancers of the thyroid, anus and gallbladder. Of note, differences in incidence rates do not always translate in differences in death rates. Conversely, death rates in males and females can be very different despite similar incidence rates. Although the death rates are comparable, thyroid cancer incidence is three times higher in women. Some evidence suggests that the overdiagnosis rate is higher in women resulting in more nonfatal tumors diagnosed in women[132]. The death rates of melanoma are more than double in male compared to female but the incidence rate is only 60% higher[131]. The sex disparities in survival rates is partly explained by the younger age at diagnosis for women but is still present when controlling for known factors[133]. In general, the earlier diagnosis of cancer in women could contribute to the lower overall death rates.

Difference in height might contribute, to some extent, to gender disparities in cancer incidence. Indeed, height has been positively associated with cancer inci-

dence and death in both men and women[134]. This suggests that hormonal and genetics factors associated to height might be involved in cancer development and progression. A study by Walter et al. tested how much height could explain the gender differences in cancer risk and found that around a third of the excess risk for men was explained by height differences[135]. Prognosis in childhood cancers is also worse in males than females, suggesting the presence of non-hormonal factors[136].

Some imbalances could be explained by mutations in the sex chromosomes. Chromosome X hosts several tumor suppressor genes such as *UTX* (*KDM6A*), *ZMYM3*, *AMER1* (also known as *WTX*), *KDM5C*. *KDM6A* has a homolog gene in chromosome Y, *KDM6C*. Loss of chromosome Y (LOY) was observed in a number of cancers that are more prevalent in males, such as prostate, bladder or liver cancers[137,138,139].

1.7.4 Clear Cell Renal Cell Carcinoma

Each year, more than 330,000 cases of kidney cancer are diagnosed in the world and over 140,000 deaths are caused by this cancer[140]. Most of the kidney cancers start in the cells that line the renal tubules and are called renal cell carcinoma (RCC). 90% of kidney cancer cases are RCCs, of which 60-80% are clear cell RCCs (ccRCCs). Cigarette smoking, increased body mass index and hypertension are risk factor for RCC[141]. As mentioned previously, kidney cancer incidence is higher for males than females, with a ratio around 2:1. Although factors such as tobacco smoking and hypertension might partly explain the gender disparities, the increased incidence in males is not fully understood.

Recent studies characterized the genomic landscape of ccRCC using exome[142] or whole genome sequencing[143] and methylation arrays[142]. The von Hippel–Lindau tumor suppressor gene (*VHL*) had been previously identified as an important driver gene with somatic point mutations or epigenetic changes present in around 80% of ccRCC[144,145]. Furthermore, the small arm of chromosome 3, in which *VHL* is located, is lost in about 90% of ccRCC tumors[146]. To a lower extent, *PBRM1*, *SETD2* and *BAP1* genes in chromosome 3 further harbored recurrent somatic point

mutations[147,148,149]. Other frequent arm-level aberrations include 7 or 5q gains and losses of 6q, 8p and 14q[127,146,150]. Recurrent somatic mutations in the *KDM5C* gene result in deregulation of H3K4 methylation which promotes genomic instability in ccRCC[143,151].

The *KDM6A* gene encodes a histone demethylase and is also recurrently mutated in ccRCC[151]. Like the *SETD2* gene, the *KDM5C* and *KDM6A* genes encode histone modifiers, highlighting the importance of epigenetic regulation in this type of cancer. As mentioned before, both these genes are located on chromosome X and might contribute to the gender disparities in incidence rates.

1.8 Hypothesis and Objectives

Technical variation in whole-genome sequencing data hinders the detection of challenging genomic variants such as somatic alterations in cancer, small CNVs and variation in low-mappability regions. These classes of variants are challenging to detect because the strength of the supporting signal is often comparable to technical noise. Although some corrections exists in order to minimize the known biases or mask the effects of repeats, technical bias remains. Moreover, repeat-rich regions are often discarded by existing methods, explaining their absence from public CNV databases. Current methods either analyze one genome at a time or pool several genomes to aggregate evidence rather than to control for technical variation.

We hypothesize that using a set of samples as reference to define and identify abnormal read coverage could increase the resolution at which CNV can be detected. We speculate that using large datasets could bypass the need to identify unknown biases in WGS data and provide a useful baseline for read coverage without modeling complex repeat structures. In addition to the benefits in term of sensitivity, population-based approaches could also address variation in repeat-rich regions.

The primary objective of this work is to demonstrate the power of population-based approaches to detect CNV. Thus, the same original idea of using reference

samples to correct for technical bias was applied to three studies, each addressing one of the following objectives. The first objective was to show how using reference samples could help identify somatic arm-level CNVs, even if present in only a fraction of the tumor cells sequenced. After variant detection, the goal was to describe the prevalence of somatic loss of chromosome Y in kidney cancer in the context of other arm-level aberrations. The second objective was to extend this approach to the detection of CNV in small genomic regions. If the population-based approach was successful for large somatic CNVs, we aimed at proving that a similar method could improve the detection of small germline CNVs. Once implemented and validated, we applied the method to a disease study of 198 epilepsy patients and 301 controls with the objective of test the importance of small CNVs as genetic factors of epilepsy. Finally, the last objective was to use this method to investigate CNV in low-mappability regions. Thanks to its robustness to technical variation, our population-based method was tested and validated on different repeat profiles before being used to detect CNVs across 640 genomes of healthy individuals. The goal here was to produce a genome-wide CNV catalog that was more representative by including low-mappability regions and to investigate the enrichment of different repeat families with CNV.

Chapter 2

Population-Based Detection of Somatic Loss of Chromosome Y in Cancer

Preface: Bridging Text between Chapters 1 and 2

In this chapter, we first aimed at detecting somatic arm-level CNVs in kidney cancer. Although arm-level CNVs are more straightforward to detect than small CNVs, the difficulty lies in the somatic nature of the variants. Indeed, somatic CNVs are often present in only a minority of the cancer cells that were sequenced. We also had a special interest in chromosome Y which is particularly rich in repeated sequences. To robustly detect somatic arm-level CNV, we propose a population-based approach to analyze 93 pairs of ccRCC tumors and peripheral blood. By using the read coverage in the 93 normal blood samples, we show that we can detect somatic CNVs in WGS even if shared by a small fraction of the tumor cells. We further investigate the functional impact of somatic loss of chromosome Y in tumors from male patients.

This study was published as an Article in *Scientific Reports*[1]. Appendix B contains supplementary tables and figures from this publication.

Loss of chromosome Y leads to down regulation of *KDM5D* and *KDM6C* epigenetic modifiers in clear cell renal cell carcinoma

Madeleine Arseneault[1,2,*], Jean Monlong[1,2,*], Naveen S. Vasudev[3], Ruhina S. Laskar[4], Maryam Safisamghabadi[1,2], Patricia Harnden[3], Lars Egevad[5], Nazanin Nourbehesht[1,2], Pudchalaluck Panichnantakul[1,2], Ivana Holcatova[6], Antonin Brisuda[7], Vladimir Janout[8], Helena Kollarova[8], Lenka Foretova[9], Marie Navratilova[9], Dana Mates[10], Viorel Jinga[11], David Zaridze[12], Anush Mukeria[12], Pouria Jandaghi[1,2], Paul Brennan[4], Alvis Brazma[13], Jorg Tost[14], Ghislaine Scelo[4], Rosamonde E. Banks[3], Mark Lathrop[1,2], Guillaume Bourque[1,2], Yasser Riazalhosseini[1,2,+]

[1]Department of Human Genetics, McGill University, 1205 Dr Penfield Avenue, Montreal, QC, H3A 1B1, Canada

[2]McGill University and Genome Quebec Innovation Centre, 740 Doctor Penfield Avenue, Montreal, QC, H3A 0G1, Canada

[3]Leeds Institute of Cancer and Pathology, University of Leeds, Cancer Research Building, St James's University Hospital, Leeds, LS9 7TF, UK

[4]International Agency for Research on Cancer (IARC), 150 cours Albert Thomas, 69008 Lyon, France

[5]Karolinska Institutet, Department of Pathology, SE-171 77 Stockholm, Sweden

[6]First Faculty of Medicine, Institute of Hygiene and Epidemiology, Charles University in Prague, Studničkova 7, Praha 2, 128 00 Prague, Czech Republic

[7]University Hospital Motol, V Úvalu 84, 150 06 Prague, Czech Republic

[8]Department of Preventive Medicine, Faculty of Medicine, Palacky University, Hnevotinska 3, 775 15 Olomouc, Czech Republic

[9]Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute and MF MU, Zluty Kopec 7, 656 53 Brno, Czech Republic

[10]National Institute of Public Health, Dr Leonte Anastasievici 1–3, sector 5, Bucuresti 050463, Romania

[11]Carol Davila University of Medicine and Pharmacy, Th. Burghele Hospital, 20 Panduri Street, 050659 Bucharest, Romania

[12]Russian N.N. Blokhin Cancer Research Centre, Kashirskoye shosse 24, Moscow 115478, Russian Federation

[13]European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK

[14]Laboratory for Epigenetics & Environment, Centre National de Génotypage, CEA-Institut de Génomique, 2 rue Gaston Crémieux, 91000 Evry, France

[+]Correspondence: yasser.riazalhosseini@mcgill.ca

[*]These authors contributed equally to this work

2.1 Abstract

Recent genomic studies of sporadic clear cell renal cell carcinoma (ccRCC) have uncovered novel driver genes and pathways. Given the unequal incidence rates among men and women (male:female incidence ratio approaches 2:1), we compared the genome-wide distribution of the chromosomal abnormalities in both sexes. We observed a higher frequency for the somatic recurrent chromosomal copy number variations (CNVs) of autosomes in male subjects, whereas somatic loss of chromosome X was detected exclusively in female patients (17.1%). Furthermore, somatic loss of chromosome Y (LOY) was detected in about 40% of male subjects, while mosaic LOY was detected in DNA isolated from peripheral blood in 9.6% of them, and was the only recurrent CNV in constitutional DNA samples. LOY in constitutional DNA, but not in tumor DNA was associated with older age. Amongst Y-linked genes that were downregulated due to LOY, *KDM5D* and *KDM6C* epigenetic modifiers have functionally-similar X-linked homologs whose deficiency is involved in ccRCC progression. Our findings establish somatic LOY as a highly recurrent genetic defect in ccRCC that leads to downregulation of hitherto unsuspected epigenetic factors, and suggest that different mechanisms may underlie the somatic and mosaic LOY observed in tumors and peripheral blood, respectively.

2.2 Introduction

Chromosomal aneuploidy is a common phenomenon in many cancers, and the analysis of copy number variations (CNVs) across multiple samples has helped identify relevant driver genes for human cancers. For example, several oncogenes including *MYC*, *EGFR*, *ERBB2* and *CCND1* are recurrently amplified through chromosomal or focal gains, while multiple tumor suppressors such as *ATM*, *PTEN* and *CDKN2A* are commonly deleted in different cancers[126].

Clear cell renal cell carcinoma (ccRCC), which accounts for 75–80% of all renal cell carcinomas, is characterized by loss of chromosome 3p in about 90% of the

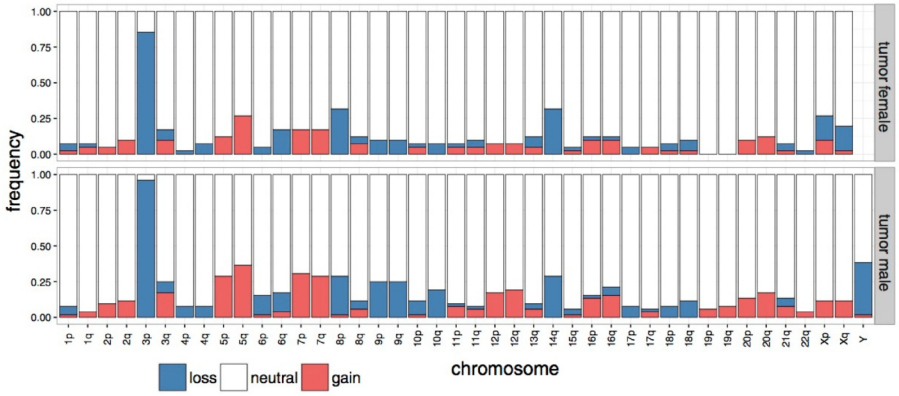
sporadic cases[146]. Remarkably, 3p harbors the four most commonly mutated genes in ccRCC whose cancer-driving activities have been established in the disease; *VHL*[144], *PBRM1*[147], *SETD2*[148], and *BAP1*[149], which are mutated in 80%, 40%, 19% and 12% of cases, respectively[151,152,153]. Inactivation of *VHL* leads to constitutive stabilization of the hypoxia inducible transcription factors (HIF), and abnormal activation of their downstream genes, which contribute to cancer development[154]. The remaining three genes encode proteins involved in chromatin remodeling and histone modifications, highlighting the important role of epigenome aberration in the disease[146]. While the incidence of ccRCC is increasing worldwide, the male-to-female incidence ratios are typically within the range of 1.5-2:1.0[155], arguing for a sex-specific analysis of the genomic abnormalities. Here, we set out to investigate the occurrence and the extent of germline and somatic CNVs in sporadic ccRCC in male and female patients separately, and to further characterize those affecting sex chromosomes.

2.3 Results and Discussion

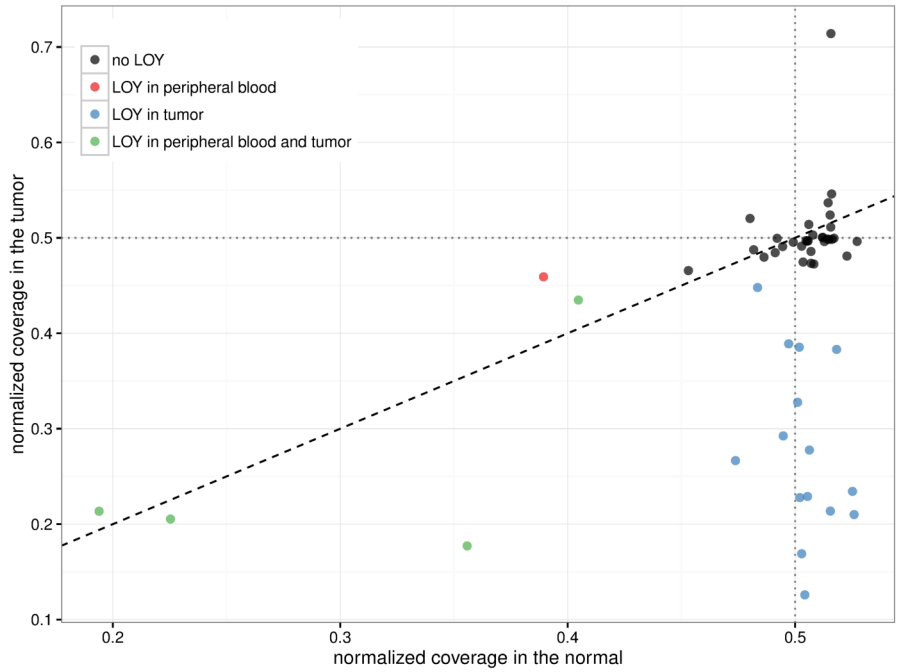
Loss of chromosome Y is common in ccRCC Using whole-genome sequencing (WGS) data of ccRCC and matched constitutional DNA sample pairs, which we have reported recently[143], we interrogated CNVs in DNA from 52 male and 41 female patients (discovery set; Supplementary Table S2.1) by analyzing coverage of sequencing reads mapped to each chromosome (see Methods). In line with previous literature, the most frequent somatic CNV was the loss of 3p detected in 91% of samples, followed by recurrent gains of chromosomes 5q (32%), 7 (23.6%), 12 (13%), and losses of chromosomes 14q (30%), 8p (29%) and 9 (16%). Overall, tumors from male patients exhibited higher prevalence for the recurrent chromosomal aberrations, in particular for gain of 7q (28% in males vs. 17% in females) and deletion of 9p (25% in males vs. 10% in females) (Fig. 2.1a). In contrast, we observed that loss of chromosome X (LOX) exclusively happens in female patients (17.1% of female cases). Given that several X-linked genes escape X-inactivation, and have therefore two functional copies in females but one in males, this observation suggests that

presence of a copy of chromosome X may potentially be essential for the survival of cancer cells. Curiously, whereas no tumors from male patients displayed LOX, loss of chromosome Y (LOY) was the second most frequent somatic chromosome aneuploidy in these tumors (36.5% of male subjects, N=19; Fig. 2.1a). The fraction of cells estimated to be affected by somatic LOY in these patients ranged from 11% to 75%, and in 14 patients somatic LOY was detected in at least 20% of the cells (Fig. 2.1b). Next, we examined the presence of CNVs in constitutional DNA isolated from peripheral samples collected from the same patients. Of significance, LOY was the only recurrent aneuploidy in constitutional DNA of our samples that was detected in 5 male patients (9.6%; Fig. S2.1), of which 4 showed the deletion in more than 20% of cells (Fig. 2.1b). Corroborating previous studies[156,157], the observed LOY was associated with older age in patients (P=0.04); the average age of patients with LOY in the peripheral blood was 68.9 year in comparison to 58.8 year in those without this abnormality. Notably, we did not observe any association between age of patients and extent of somatic LOY in tumors of the affected patients.

LOY is a whole-chromosome event Given the high prevalence of LOY in tumors and peripheral DNA of male patients, we further analyzed LOY in our sample series, particularly whether the observed LOY spans the whole chromosome or is focal. Analysis of sequencing read coverage along chromosome Y showed that the loss is observed throughout the chromosome in samples affected by LOY (Fig. 2.2a), suggesting that the deletion affects the whole chromosome. Based on availability of DNA, we subjected samples from seven of the patients affected with somatic LOY to verification by an orthogonal Y-Chromosome deletion detection assay surveying the presence of twenty specific regions of the Y chromosome by polymerase chain reaction (PCR) (see Methods). Somatic LOY at the chromosomal level was confirmed in all examined tumors, evident from an attenuated amplification of Y-chromosome-specific loci in DNA isolated from tumor samples compared to that of the matched constitutional DNA. This pattern was not observed in samples of other male patients who had not been identified as being affected by somatic LOY based on the



(a)



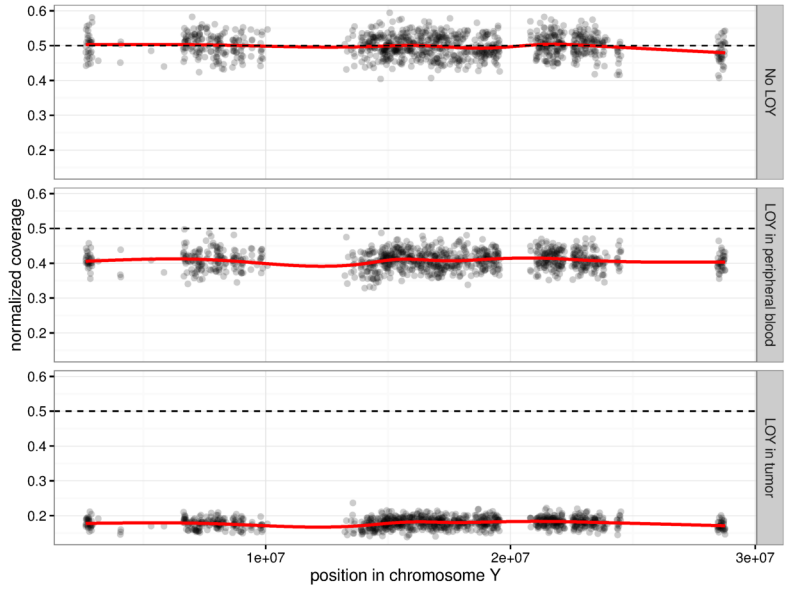
(b)

Figure 2.1: **Copy number analysis in ccRCC.** a) Bar graphs show the frequency of copy number variations across the genome in ccRCC tumors. Frequencies are presented in samples from female and male cases separately. b) Status of chromosome Y in DNA isolated from tumors (Y-axis) and patient-matched peripheral blood (X-axis) is shown for individual male subjects. In samples affected by LOY, the normalized coverage of chromosome Y, shown on Y and X axes for tumor and normal samples, respectively, is lower than the expected value of 0.5. The color codes define patient groups with different states for LOY.

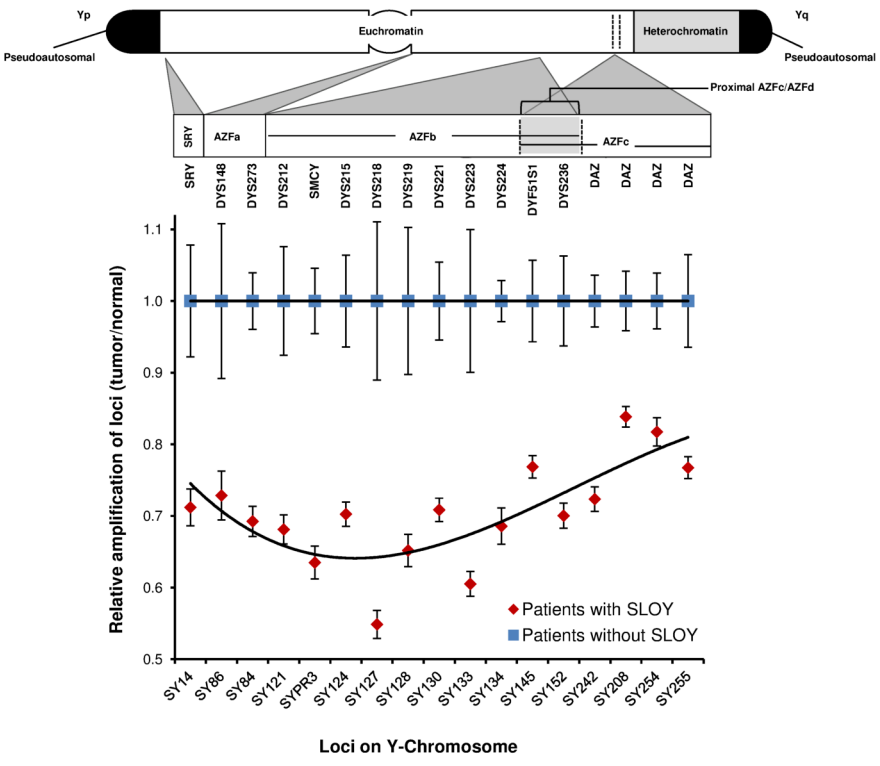
analysis of their WGS data (Fig. 2.2b).

To confirm these findings, we screened tumor and matched control DNA sample pairs of an additional 48 male ccRCC patients (validation set) for LOY using the above PCR-based assay. This analysis revealed somatic LOY in 20 (42.7%) of the validation sample set, demonstrating that this is a common genomic aberration in ccRCC, detected in 39.6% overall (discovery and validation sets; n=100) of male ccRCC patients (Fig. S2.2). Analysis of association between somatic LOY and clinical annotations including tumor stage or grade did not show any significant relationships.

LOY results in downregulation of epigenetic modifier genes We further examined the possible effect of somatic LOY at the RNA level by interrogating a RNA-Seq dataset on gene expression in normal and tumor samples from male patients within the discovery set[143]. We found that 11 genes had significantly different patterns of expression in tumors of the patients with and without somatic LOY (false-discovery rate (FDR)<0.01; Supplementary Table S2.2). These 11 genes were located on chromosome Y, and while expressed in normal kidney tissue, exhibited lower expression in tumors of patients harboring somatic LOY, indicating that this aberration may have functional consequences through deregulation of the affected genes. Moreover, the level of expression of each gene was found to be inversely correlated to the proportion of cells affected by LOY (Fig. 2.3). This observation was confirmed using gene expression data generated by microarrays, which was available for 29 tumors of the validation set[158] (Fig. S2.3). We surveyed the list of genes affected for potential functionally-relevant candidates. Among these genes, *TMSB4Y* has recently been identified as a tumor suppressor gene downregulated in male breast cancers[159], but not connected to ccRCC. Likewise, deletion of *KDM5D* has been detected in 52% of prostate cancers[160]. *KDM5D* encodes a lysine-specific histone H3 demethylase, which plays an important role in epigenetic regulation[161]. Furthermore, it has been shown that knockdown of *KDM5D* through RNA-interference (RNAi) increases cell proliferation and reduces apoptosis in prostate cancer[162], sug-



(a)



(b)

Figure 2.2: **LOY affects whole chromosome.** a) Sequencing coverage across chromosome Y is shown in constitutional DNA samples without (top) and with LOY (middle), and in a tumor sample with LOY (bottom). b) The cartoon on top depicts the location of the loci examined by PCR on Y chromosome. The dot graph on bottom shows average of relative amplification values (Tumor/normal samples of the same patient) for each locus in patients with (red) and without (blue) somatic LOY (SLOY). Error bars show the range across patients of each group.

gesting a tumor suppressor function for this gene. Intriguingly, *KDM5C*, the X-linked homologue of *KDM5D* is recurrently mutated in ccRCC[143,151,153], and its inactivation leads to genomic instability in ccRCC through deregulation of H3K4 methylation[163]. *KDM5D* shows 85% sequence identity to *KDM5C* and the products of these two genes possess a similar function in demethylating tri-methyl H3K4[161,163]. Given this functional similarity, we surveyed the mutational status of *KDM5C* in our discovery set, and investigated possible relationships between mutational status of *KDM5C* and *KDM5D* in tumors of male patients. In female patients, *KDM5C* was deleted in tumors of 7 cases through somatic LOX, and was affected by focal somatic deletions in two additional patients. Furthermore, somatic mutations of *KDM5C* were present in tumors of 3 patients who were also affected with LOX (P=0.003, Fisher's exact test, Fig. S2.4). Overall, *KDM5C* was affected with somatic genomic aberrations in 9 out of 41 (22%) female cases. As *KDM5C* escapes the X-inactivation[164], the concomitant mutations of *KDM5C* and LOX in the same tumors may suggest that this gene is a classical tumor suppressor affected with bi-allelic inactivation in ccRCC. In male cases, we identified *KDM5C* mutations in tumors of 3 patients (5.8%), of which one was also affected by somatic LOY (Fig. S2.4). We did not detect any mutation or a focal CNV affecting *KDM5D* in tumors of the male patients who did not exhibit LOY.

Our list of LOY-associated down-regulated genes (Supplementary Table S2.2) includes another epigenome modifier with an X-linked homologue that is also recurrently mutated in ccRCC; *UTY/KDM6C*. *KDM6C* demethylates H3K27, a function similar to that of *KDM6A*[165]. These genes also share over 83% in sequence similarity, resulting in highly conserved active sites in their products. Mutations of *KDM6A* leading to its inactivation have been recurrently observed in ccRCC[151,166], highlighting this gene as a potential key tumor suppressor in renal cancer. In addition to being affected by somatic LOX in 7 female patients, *KDM6A* was also affected by focal deletion in a female patient in our cohort.

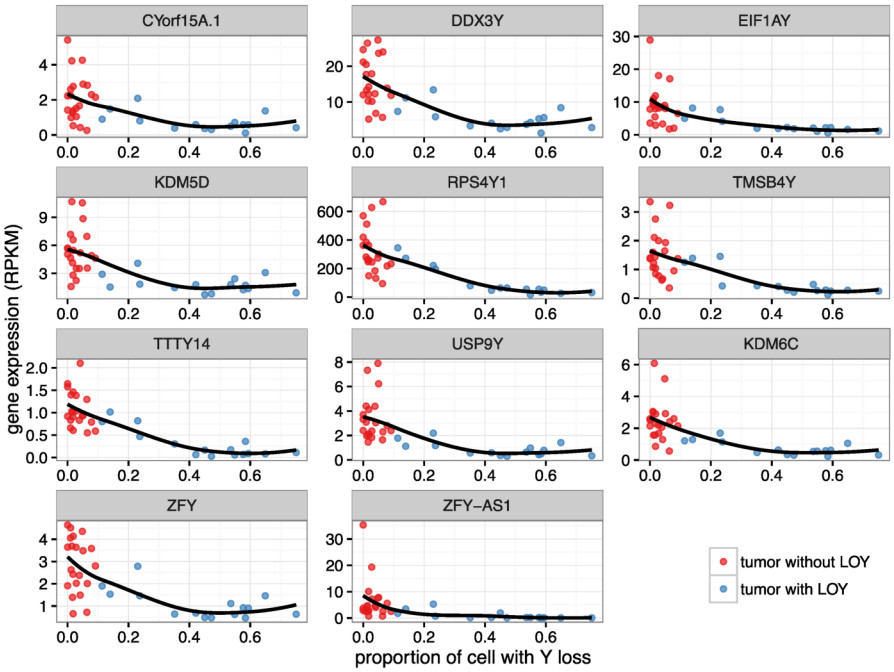


Figure 2.3: **Somatic LOY leads to downregulation of Y-linked genes.** Expression of Y chromosome genes downregulated in patients affected by somatic LOY is compared to the proportion of cells estimated to harbor somatic LOY in individual tumor samples.

***KDM5D* expression reduces viability of renal cancer cells** Given the reported tumor-suppressive function of *KDM5D* in prostate cancer[162], and of its X-link homolog *KDM5C* in renal cancer[163], we set out to examine whether *KDM5D* expression has an anti-tumor activity in renal cancer. We first evaluated *KDM5D* expression levels in several renal cancer cell lines, which have been derived from tumors resected from male patients. Amongst cell lines examined, ACHN cell line did not show any expression for *KDM5D* (Fig. 2.4a). This observation was in line with a previous study reporting the loss of chromosome Y in ACHN cell line[167]. We therefore selected this cell line for functional analysis of *KDM5D* expression. Ectopic expression of *KDM5D* cells reduced cell viability to 65% as compared to control transfection (Fig. 2.4b-c), suggesting the potential involvement of *KDM5D* depletion in renal cancer pathology.

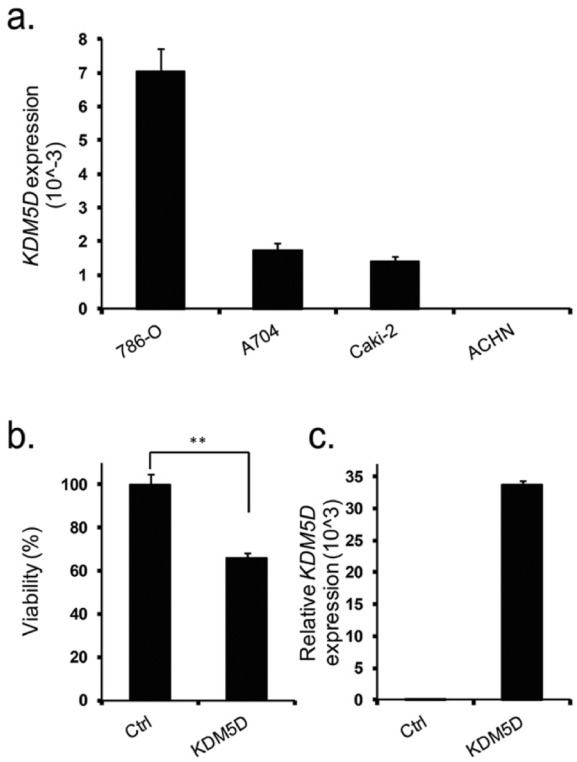


Figure 2.4: **Effect of *KDM5D* on viability of renal cancer cells.** a) Expression levels of *KDM5D* mRNA in renal cancer cell lines derived from tumors procured from male patients, as measured by qRT-PCR. *GAPDH* served as a housekeeping gene for measurement of relative gene expression. b) Over expression of *KDM5D* in ACHN cell line reduces cell viability. Values are the mean ± SD of six independent experiments. $^{**}P < 0.01$ when compared to the corresponding results from control (ctrl) (Mann-Whitney U test). c) Over expression of *KDM5D* following transfection was confirmed using qRT-PCR.

2.4 Conclusions

Emerging data emphasizes an association between LOY in peripheral blood and higher risk of cancer[168]. Likewise focal or chromosome-level somatic LOY occurs recurrently in different malignancies; however, current knowledge of mechanisms by which LOY may contribute to cancer is limited. Recent genomic studies of ccRCC have highlighted the importance of molecular aberrations that impair the function of chromatin remodeling and epigenetic modifiers in ccRCC development[148,163,169,170,171,172]. Our study expands these findings by highlighting the prevalence of somatic LOY among men affected by ccRCC, and suggesting a functional relevance for this aberration through down-regulation of previously unrecognized epigenetic modifiers *KDM5D* and *KDM6C*. Given the functional similarities between these genes and their X-linked homologs, it is plausible that down-regulation

of *KDM5D* and *KDM6C*, through somatic LOY, may contribute to ccRCC development or progression. Our in vitro data shows that over expression of *KDM5D* in cancer cells that are affected by LOY reduces cell viability. These findings indicate that down-regulation of *KDM5D* through LOY may contribute to the pathogenesis of renal cancer. However, further detailed analysis through future functional studies is warranted to understand the exact function and pathway context of *KDM5D* in renal cancer.

2.5 Methods

Patient samples and DNA isolation Clinical information for patients included in this study is presented in Supplementary Table S2.1. Patients undergoing nephrectomy for suspected renal cancer during the period December 2008 to March 2011 at St James's University Hospital in Leeds, UK; University Hospital Motol, Prague, Czech Republic; Masaryk Memorial Cancer Institute, Brno, Czech Republic; Th. Burghele Hospital, Bucharest, Romania; and N. N. Blokhin Cancer Research Centre, Moscow, Russia, were recruited to the study after informed consent was obtained. Recruitment in Central and Eastern Europe was coordinated by the International Agency for Research on Cancer (IARC). All experiments and methods were performed in accordance to the ethics guidelines from the International Cancer Genome Consortium (ICGC) and to the relevant national regulations and with sampling and clinical data collection being undertaken according to predefined standard operating procedures (SOPs) based on guidelines from ICGC. Ethical approvals were obtained from the Leeds (East) Local Research Ethics Committee, the IARC Ethics Committee, as well as from local ethics committee for recruiting centers in Czech Republic, Romania, and Russia. DNA from fresh-frozen tumor tissue samples and buffy coat was isolated using Autopure (Qiagen) as described previously[143], and were quantified by Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, ON, CAN).

Inference of LOY from WGS data WGS data of tumor and blood DNA samples studied here were reported previously[143]. To detect aneuploidy and LOY from WGS data, we first measured read coverage across the genome in 5 Kbp bins. In each sample, the coverage was normalized by the median coverage across the autosomes. We then estimated, for each sample, the median normalized coverage in each chromosome arm. The only exception was chromosome Y which was considered as a whole. In order to avoid noise due to mappability issues, we used only the top 1000 bins with the lowest median divergence from the expected baseline in the normal samples. We used this normalized median coverage per chromosome arm to test aneuploidy in each sample. For each chromosome arm (or chromosome Y), a mixture of two Gaussian distributions was fitted to the empirical distribution of the median normalized coverage across samples. The main Gaussian was used as the null distribution (Fig. S2.5) to derive P-values. A chromosome arm was flagged as aneuploid if the Bonferonni-adjusted P-value was smaller than 0.01 and at least 10% of cells were affected. The proportion of cell with aneuploidy was estimated as the proportion of missing/excess coverage. For LOY, we expect a normalized coverage of 0.5 and the proportion of cells with LOY was (0.5-coverage)/0.5.

We used a logistic regression to test the association of LOY with age. Finally, the CNVs used for *KDM5C* or *KDM6A* deletion investigation were detected by PopSV[173] using the normal samples as reference and 5 Kbp bins.

PCR-based detection of LOY To examine the status of LOY in DNA of tumor and blood samples, Y Chromosome Deletion Detection System assay, Version 2 (Promega, WI, USA) was used as instructed by the manufacturer. Briefly, 20 specific regions of the Y chromosome were amplified by PCR using 5 multiplex master mixes, and PCR products were loaded on a QIAxcel instrument (Qiagen, ON, CAN). Densities of PCR products were estimated by BioCalculator software (v.3.2) and a normalization was performed by the control primer pair included in each multiplex master mix to control the amplification efficacy. We also included samples from three male subjects without LOY and one female sample to control the performance of the

assay. Similar to the analysis on WGS data, the probes were first normalized by the median probe amplification value across the normal samples. Then the median of the normalized amplification was computed for each sample. It summarized the overall amplification of chromosome Y in each sample. These values were used to produce Fig. S2.2 and to identify LOY. Following the same analysis as for the WGS data, the mixture of Gaussian distributions was fitted on the normalized amplification of the normal samples. Samples which deviated significantly ($P<0.01$) from the expected amplification and with an estimated proportion of affected of cells >10% were flagged as being affected by LOY.

Gene expression analysis Transcriptome profiles of the tumor samples included in this study (previously reported in our earlier publication[143]), were used to examine differential gene expression between male subjects affected with somatic LOY and those without this abnormality. RNA-seq data was available for tumors of 34 patients, of which 21 had RNA-seq for matched normal kidney samples. Differentially expressed genes between tumors affected with somatic LOY and those without this abnormality were identified using Student's T-Test on log2-transformed RPKM data, and the Benjamini-Hochberg method was used to correct multiple testing. Genes with a FDR$<$ 0.01 were considered differentially expressed. A linear regression was used to test the association between the proportion of cells with somatic LOY and gene expression (RPKM).

Gene expression microarray data for 29 tumors of validation samples had previously been reported[158], and were used to confirm the anti-correlation between the proportion of cells with somatic LOY and gene expression levels (log2 intensity).

Cell viability assay Renal cancer cell lines 786-O, A704, Caki-2, ACHN were obtained from ATCC (Rockville, USA) and cultured in RPMI, EMEM and McCoy medium supplemented with 10% (v/v) fetal bovine serum (FBS), 100 U/ml penicillin and 100 lg/ml streptomycin. Cells were incubated at 37C and 5% (v/v) CO2. For viability assays, 5000 cells were transfected with 100 ng of either *KDM5D* cDNA-

expressing (courtesy of Dr. Stephane Richard) or control empty vector (Sigma, Oakville, Canada) in 96-well plates using Lipofectamine 3000 (Invitrogen) according to the manufacturer's instructions. CellTiter-Glo assay (Promega, WI, USA) was used to assess cell viability after 72 hours post-transfection.

Quantitative real-time PCR (qRT-PCR) Total RNA was extracted from cells using miRNeasy kit (Qiagen, Toronto, Canada) according to the supplier protocols. 1 μg RNA was reverse transcribed into complementary DNA (cDNA) using Transcriptor First Strand cDNA Synthesis Kit (Roche, Laval, Canada) following instructions provided by the manufacturer. Real-time PCR reactions were prepared using LightCycler 480 SYBR green I master kit (Roche), and were run on a LightCycler 480 instrument (Roche) according to the manufacturer's recommendations. Triplicate PCR reactions were performed for each sample to ensure reliability. Expression of *KDM5D* mRNA was normalized to the expression of the housekeeping gene *GAPDH*, and was reported as $2^{-\Delta Ct}$. All the primers were purchased from IDT (Coralville, IA, US). The sequences of primers were `CGTGGAAGGACTCATGACCA` (*GAPDH* forward), `GCCATCACGCCACAGTTTC` (*GAPDH* reverse), `CGCAGCTTTGAAGAGCTAAG` (*KDM5D* forward) and `CAGCTGTGGAGTGTCCATCC` (*KDM5D* reverse).

2.6 Acknowledgments

This study was supported by funding from EU FP7 under grant agreement number 241669 (the CAGEKID project, http://www.cng.fr/cagekid), and funding from Génome Québec as well as Ministère de l'Enseignement supérieur, de la Recherche, de la Science et de la Technologie (MESRST). We also acknowledge the support of Cancer Research UK Centre and ECMC infrastructure funding in Leeds for contributions to sample collection. The Czech-Brno group was supported by MH CZ - DRO (MMCI, 00209805). JM is supported by funds from National Sciences and Engineering Research Council (NSERC-448167-2013).

Chapter 3

Population-Based Detection of CNVs in Epilepsy Patients

Preface: Bridging Text between Chapters 2 and 3

In this chapter, the population-based approach used to detect arm-level CNV was extended to the detection of small CNV across the genome. Instead of testing one chromosomal arm at a time, the genome is tiled and each region is tested for CNV using the same strategy: a set of reference samples is used to deal with technical variation. By better integrating technical variation, our goal is to improve the robustness and sensitivity of the CNV detection. Notably, this chapter introduces PopSV, the CNV detection method implemented in the context of this thesis, and its application to a disease study. PopSV is later used in chapter 4 to investigate CNVs in low-mappability regions.

In this Research Article published in *PLoS Genetics*[2], we first introduce the rationale for the method: the presence of visible technical bias in WGS coverage data. After an overview of the method, we show that it is more sensitive than other methods and avoids systematic calls. Finally, we apply PopSV to WGS data from 198 epilepsy patients and 301 controls and show CNV enrichments in exons and non-coding regions that potentially explain an important fraction of patients. Appendix C contains supplementary tables, figures and information.

Global characterization of copy number variants in epilepsy patients from whole genome sequencing

Jean Monlong[1,2,*], Simon L. Girard[1,3,4,*], Caroline Meloche[4], Maxime Cadieux-Dion[4,5], Danielle M. Andrade[6], Ron G. Lafreniere[4], Micheline Gravel[4], Dan Spiegelman[7], Alexandre Dionne-Laporte[7], Cyrus Boelman[8], Fadi F. Hamdan[9], Jacques L. Michaud[9], Guy Rouleau[7], Berge A. Minassian[10], Guillaume Bourque[1,2,11,+], 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] Département des sciences fondamentales, Université du Québec à Chicoutimi, Chicoutimi, G7H 2B1, Canada
[4] Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Montréal, H2X 0A9, Canada
[5] Center for Pediatric Genomic Medicine, Children's Mercy Hospital, Kansas City, MO, USA
[6] Epilepsy Genetics Program, Division of Neurology, Toronto Western Hospital, University of Toronto, Toronto, Canada
[7] Montreal Neurological Institute, McGill University, Montréal, H3A 2B4, Canada
[8] Division of Neurology, BC Children's Hospital, Vancouver, V6H 3N1, Canada
[9] CHU Sainte-Justine Research Center, Montréal, H3T 1C5, Canada
[10] Division of Neurology, The Hospital for Sick Children, Toronto, M5G 1X8, Canada
[11] McGill University and Génome Québec Innovation Center, Montréal, H3A 1A4, Canada
[+]Correspondence: guil.bourque@mcgill.ca or patrick.cossette@umontreal.ca
[*]These authors contributed equally to this work

3.1 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 accuracy. 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. For both large and small

variants, 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 sequence-based profiling of CNVs could help explain a larger fraction of epilepsy cases.

3.2 Author Summary

Epilepsy is a common neurological disorder affecting around 3% of the population. In some cases, epilepsy is caused by brain trauma or other brain anomalies but there are often no clear causes. Genetic factors have been associated with epilepsy in the past such as rare genetic variations found by linkage studies as well as common genetic variations found by genome-wide association studies and large copy-number variants. We sequenced the genome of ∼200 epilepsy patients and ∼300 healthy controls and compared the distribution of deletion (loss of a copy) and duplication (additional copy) of genomic regions. Thanks to the sequencing technology and a new method that takes advantage of the large sample size, we could compare the distribution of small copy-number variants between epilepsy patients and controls. Overall, we found that small variants are also associated with epilepsy. Indeed, the genome of epilepsy patients had more exonic copy-number variants, especially when rare or affecting genes with predicted loss-of-function intolerance. Focusing on regions around genes that have been previously associated with epilepsy, we also found more non-coding variants in epilepsy patients, especially deletions or variants in regulatory regions. Finally, we provide a list of 21 regions in which we found likely pathogenic variants.

3.3 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[4,10,11]. Numerous diseases are caused by SVs with a demonstrated detrimental effect[24,26]. 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[174,175]. SV detection methods that use read-pair and split read information[67] can detect deletions and duplications but most CNV-focused approaches look for an increased or decreased read coverage, the expected consequence of a duplication or a deletion. Coverage-based methods exist to analyze single samples[41], pairs of samples[40] or multiple samples[33,42,176] but the presence of technical bias in WGS remains an important challenge. Indeed, various features of sequencing experiments, such as mappability[48,49], GC content[50], replication timing[51], DNA quality and library preparation[177], have a negative impact on the uniformity of the read coverage[52].

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[178,179]. Genetic factors have been associated with epilepsy in the past such as rare genetic variations found by linkage studies as well as common genetic variations found by genome-wide association studies[180,181] For example, a meta-analysis combining multiple epilepsy cohorts found positive associations with the disease[182], the strongest in *SCN1A*, a gene already associated with the genetic mechanism of the disease via linkage studies and subse-

quent sequencing[183] or more recently as harboring *de novo* variants[184]. Thanks to array-based technologies, surveys of large CNVs (>50 Kbp) first associated CNVs in genomic hotspots such as 15q11.2 and 16p13.11 with generalized epilepsy[185,186]. Other studies have further shown the importance of large and *de novo* CNVs as well as identified a few associations with specific genes[119,120,124,125,187,188]. Rare genic CNVs were typically found in around 10% of epilepsy patients[19,119,125] and CNVs larger than 1 Mbp were significantly enriched in patients compared to controls[19,121,122,188]. 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 datasets[143,189], 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.

3.4 Results

Technical bias in read coverage

We sequenced the genomes of 198 unrelated individuals affected with epilepsy and 301 unrelated healthy controls. Because CNV detection relies on read coverage we first investigated the presence of technical bias and the value of standard corrections and filters (e.g. GC correction, mappability filtering). The genome was fragmented in 5 Kb bins and we counted the number of uniquely mapped reads in each bin. 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 (Fig. 3.1a). Supporting this observation, the bin coverage variance across samples was also lower than expected and varied between regions (Fig. S3.1a). We also observed experiment-specific biases. In particular, some samples consistently had the highest, or the lowest, coverage across large portions of the genome (Fig. S3.1b). These observations were not unique to our data and could also be observed in two public WGS datasets, and persisted even after correcting the GC bias and mappability using the more elaborate model from the QDNAseq pipeline[190] (Fig. S3.2). Our results across multiple samples suggest that existing GC bias and mappability corrections[190] cannot correct completely the technical variation in read coverage. This fluctuation of coverage has implications for CNV detection approaches that assume a uniform distribution[40,41,191] 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 (Fig. 3.1b). The normalization step here is critical since most approaches will fail to give acceptable normalized coverage scores (Fig. S3.1b). Moreover, with global median/variance adjustment or quantile normaliza-

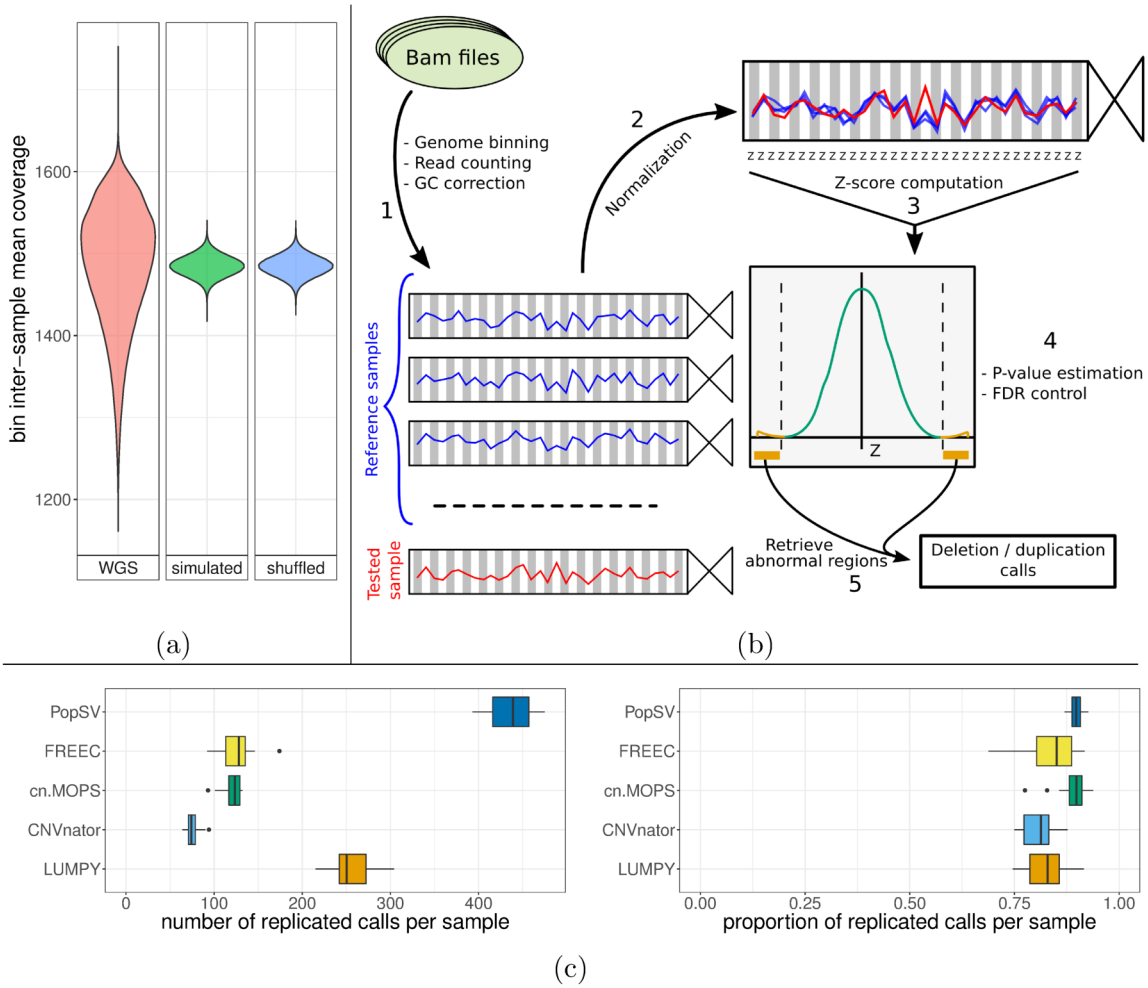


Figure 3.1: **PopSV approach.** a) Technical bias across the genome remains after stringent correction and filtering. The distribution of the bin inter-sample mean coverage in the epilepsy cohort (red) is compared to null distributions (blue: bins shuffled, green: simulated normal distribution). b) PopSV 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 monozygotic twin.

tion, the remaining subtle experimental variation impairs the abnormal coverage test (Fig. S3.3a). The targeted normalization used by PopSV was found to have better statistical properties (Fig. S3.3b). In order to assess the performance of our tool, we compared it to several algorithms[40,41,42,67] using a dataset that included monozygotic twins and also performed experimental validation of different types of predicted CNVs in the epilepsy cohort (see below). 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, Fig. S3.4). 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 dataset (Fig. S3.5). The agreement with the pedigree was computed by the Rand index after clustering the individuals with three hierarchical clustering approaches (see Supplementary Information). Looking at the replication between 10 pairs of monozygotic twins, PopSV detected more replicated CNVs compared to other methods, while maintaining similar replication rates (Fig. 3.1c). The CNV calls were further filtered with gradually more stringent significance thresholds and PopSV remained superior in term of number of replicated calls (Fig. S3.6). When investigating the overlap of calls between different methods, we noticed that PopSV was better recovering calls from CNVnator[41], FREEC[40], cn.MOPS[42] or LUMPY[67], especially if found by two or more methods (Fig. S3.7). For example, around 92% of the CNVs called by other methods were also found by PopSV when focusing on calls found in at least two methods. Similar results were also obtained in a cancer dataset where we looked for replicated germline CNVs in the paired tumor (Fig. S3.8). Finally, we repeated the twin analysis using 500 bp bins and observed high consistency with the 5 Kbp calls (Fig. S3.9). These results suggest that PopSV can accurately detect around 75% of events that are as large as half the bin size used (see Supplementary Information).

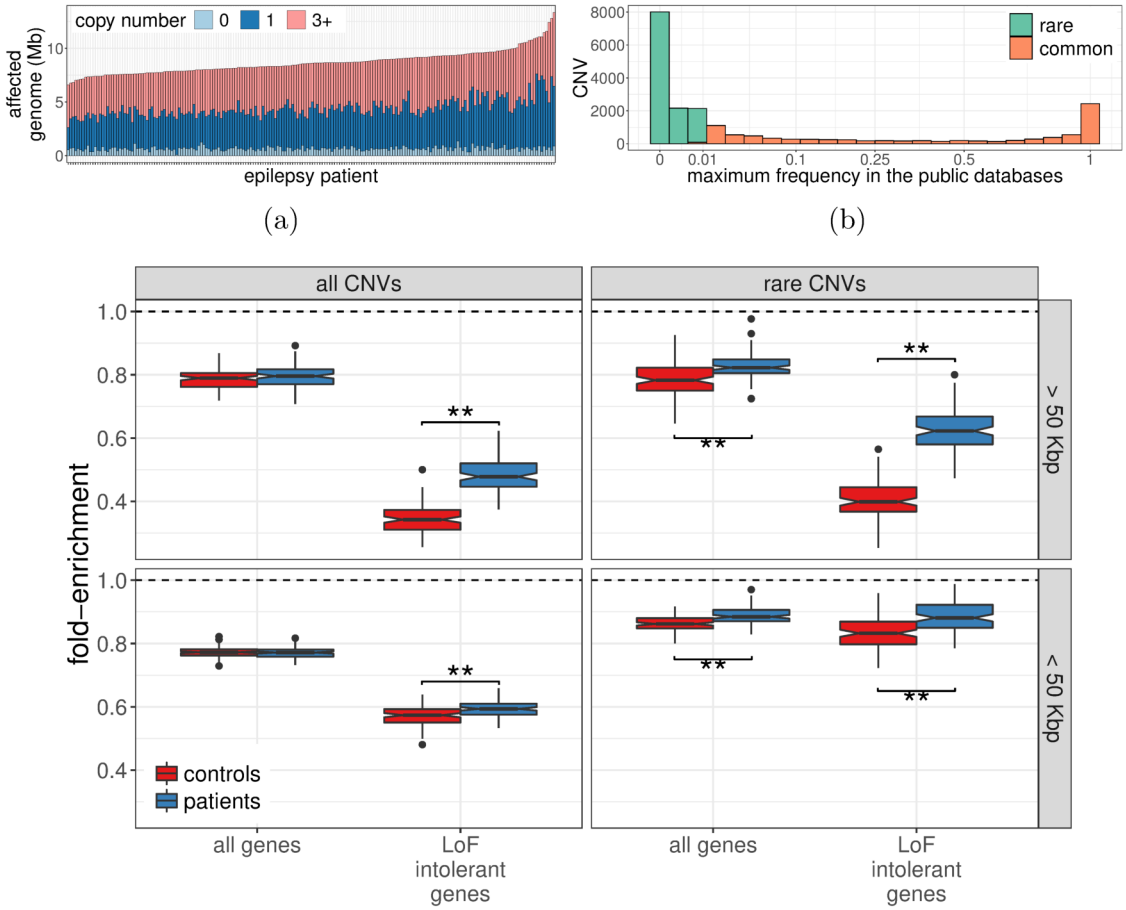
CNVs in the CENet cohorts and experimental validation

Having demonstrated the quality of the PopSV calls, we applied our tool to the epilepsy and control cohorts. The epilepsy 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 (Fig. 3.2a). This is around 9 times more variants and considerably smaller than in typical

array-based studies[106,111], such as the previous epilepsy surveys[19,119,120,125], although a similar size distribution was previously obtained using denser arrays[24] but were never applied to epilepsy (Fig. S3.10). Next, we annotated each variant using four public SV databases[5,33,107,192] as well as an internal database of the germline calls from PopSV 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 (Fig. 3.2b). 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 Real-Time PCR. To explore PopSV's performance across different CNV profiles, we selected variants of different types, sizes and frequencies. We found that the calls were concordant in 90.7% of the cases (Table 3.1 and S3.1). 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). Furthermore, we noted that calls supported by both PopSV and LUMPY (when available) had a similar validation rate as calls found by PopSV only (86.2% and 87.5% respectively).

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 found no difference between epileptic patients and controls (Fig. 3.2c). As expected, we observed fewer CNVs overlapping exonic sequence than expected by chance but similar levels for both groups. The number of CNVs overlapping exonic sequences of genes intolerant to loss-of-function mutations[193] was even lower. Interestingly, the coding regions of those genes were



(a) (b)

(c)

Figure 3.2: **CNVs in the epilepsy and control cohorts.** a) Regions with a CNV in each epilepsy patient. b) Each CNV in the CNV catalog of the epilepsy and control cohorts was annotated with its maximum frequency in five CNV databases. c) Enrichment in exonic sequence for all CNVs (left) and rare CNVs (right), larger than 50 Kbp (top) or smaller than 50 Kbp (bottom). The fold-enrichment (y-axis) represents how many CNVs overlap coding sequences compared to control regions randomly distributed in the genome.

significantly more affected by CNVs in epileptic patients compared with controls (permutation P-value<0.001, Figs 3.2c and S3.11). Because they are more likely pathogenic and of greater interest, we performed the same analysis using rare CNVs only. Here, we observed the increased exonic burden described previously for large rare CNVs[19,121,122]. In contrast to previous studies, we could also detect and compare small CNVs (<50 Kbp) in epileptic patients and healthy controls. We found similar enrichment patterns than for large CNVs (Figs 3.2c and S3.11), suggesting that small rare exonic CNVs are also associated with epilepsy. Indeed, there was no significant difference between epileptic patients and controls when considering all small CNVs

Table 3.1: **Real-Time PCR validation rates of PopSV calls.**

	Region	Validation rate
Total	151	0.907
CNV type		
Deletion	102	0.902
Duplication	49	0.918
Frequency in databases		
0	26	0.923
$(0, 0.01]$	24	0.833
$(0.01, 1]$	101	0.921
Carrier in CENet cohorts		
1	21	0.857
2	19	0.947
> 2	111	0.910
Size (Kbp)		
< 20	73	0.849
$(20, 100]$	38	0.974
> 100	40	0.950

Number and proportion of regions validated for CNVs of different types, sizes and frequencies.

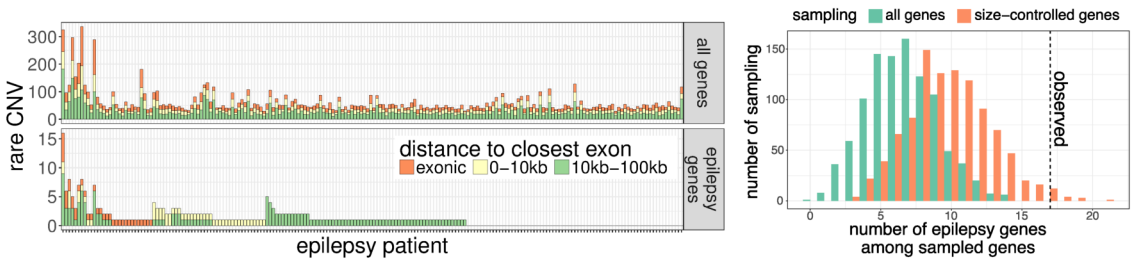
and all genes. The exonic enrichment was significant for genes with predicted loss-of-function intolerance and for rare variants (permutation P-value<0.001, Figs 3.2c and S3.11). In both cohorts, most of the rare exonic CNVs were private, i.e. present in only one individual. However, we observed that rare exonic CNVs were less likely private in the epileptic patients (permutation P-value<0.001, Fig. S3.12a). We replicated this result using only individuals with a similar population background (French-Canadians, Fig. S3.12b). 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

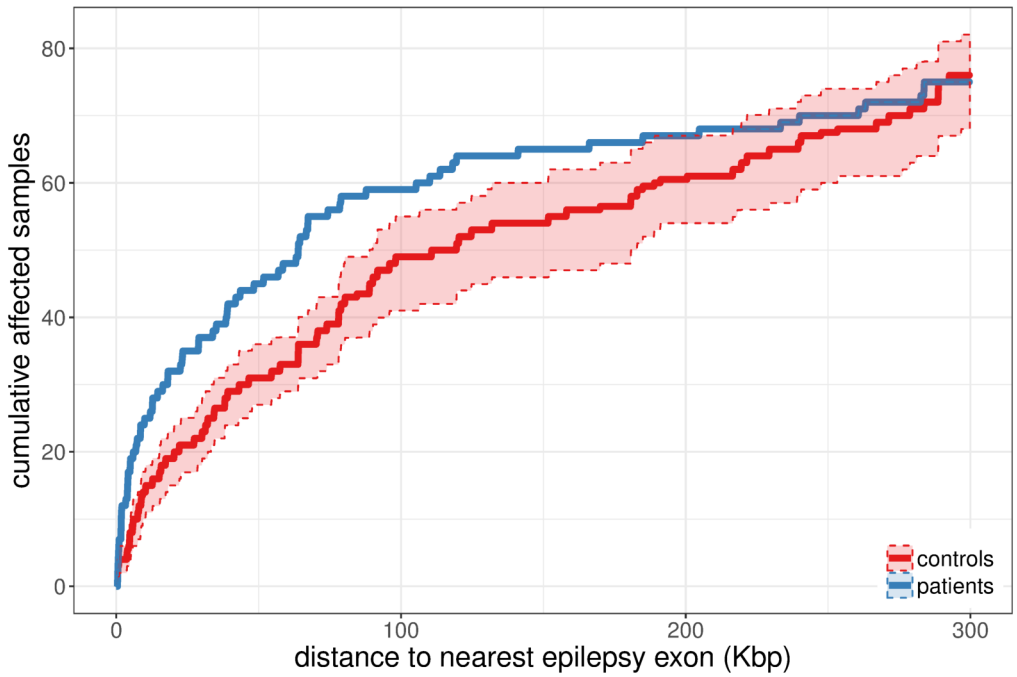
We then sought to evaluate if there was an excess of CNVs disrupting epilepsy-related genes or nearby functional regions. We first retrieved genes whose exons were hit by rare deletions or duplications and evaluated how many were known epilepsy genes based on a list of 154 genes previously associated with epilepsy[118] (Fig. 3.3a). Because epilepsy genes tend to be large, we controlled for the gene size when testing for enrichment (Fig. S3.13a). In the epilepsy cohort only, we noted a

clear enrichment for epilepsy genes hit by rare deletions (Fig. S3.13b). 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 and internal databases, 17 of which were epilepsy genes (P-value 0.015, Fig. 3.3b). In addition, we observed significantly more epilepsy patients with a rare non-coding CNV close to an epilepsy gene compared to control individuals (Fig. S3.14a). Interestingly, this enrichment was stronger for non-coding deletions (Fig. S3.14b). We further explored the distribution of rare non-coding deletions by testing each epilepsy gene for a difference in mutation load between patients and controls. The *GABRD* gene had the strongest and only nominally significant association with four non-coding deletions among the 198 epileptic patients and none in the 301 controls. *GABRD* encodes the delta subunit of the gamma-aminobutyric acid A receptor and has been associated with juvenile myoclonic epilepsy[194]. In our cohort, two of the four patients with a rare non-coding deletion close to *GABRD* had been diagnosed with this syndrome, including one patient with a 2.7 Kbp deletion located only 3 Kbp upstream of *GABRD*'s transcription start site (Fig. S3.15a). Although none survived multiple testing correction, we noted that the strongest associations were all in the direction of a higher mutation load in the epilepsy cohort rather than in the control cohort.

To get a better idea of the functional regions close to epilepsy genes, we retrieved their associated eQTLs in the GTEx database[195] and the DNase hypersensitivity sites associated with their promoter regions[196]. Notably, focusing on 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}, Fig. 3.3c). Comparing epilepsy patients and controls, the odds ratio of having such a CNV at a distance of 100 Kbp or less from an exon was 1.33 and gradually increased the closer to the exon (2.9 for CNVs at 5 Kbp or less, Fig. S3.16). These non-coding CNVs were rare even in the epileptic cohort, but collectively represented an important frac-



(a) (b)



(c)

Figure 3.3: **CNVs and epilepsy genes.** a) Number of rare CNVs in or close to exons of protein-coding genes (top) or epilepsy genes (bottom), in the epilepsy cohort. b) Number of epilepsy genes hit by exonic deletions in the epilepsy cohort and never seen in the public and internal databases (dotted line), compared to the expected distribution in all genes and size-matched genes (histograms). 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).

tion of affected patients. While 20 patients (10.1%) had exonic CNVs in epilepsy genes that were not seen in any control or in the public and internal databases, 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 the CNV 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 and internal databases (Fig. S3.15b). 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 (Fig. S3.15c).

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 exonic 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 3.2-3.3 and Table S3.2). Out of these, 8 directly affected a gene previously associated with epilepsy[118] (Table 3.2). 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[197,198]. We also identified two deletions and one duplication in *CHD2* gene (see Fig. 3.4). 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 (Fig. S3.17). No exon-disruptive CNVs were reported in any individuals from the control cohort. This gene was previously associated with patients suffering from photosensitive epilepsy[199]. 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.

Four of the 21 putative pathogenic CNVs were found in more than one individual

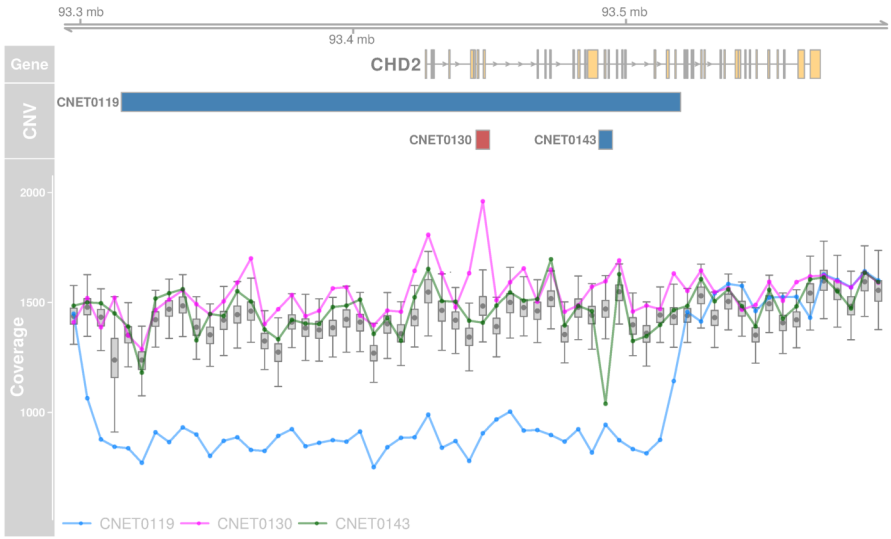


Figure 3.4: **Exonic CNVs in *CHD2* detected by PopSV.** The 'CNV' panel shows the exonic deletions (blue) and duplications (red) called by PopSV. The 'Coverage' panel shows the read depth signal in the affected individuals (colored points/lines) and the coverage distribution in the reference samples (boxplot and grey point).

Table 3.2: **Pathogenic profiles in known epilepsy genes.**

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	-

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 "-".

(see Table 3.3 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 3.3). 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 carriers of a duplication in the same region. This region corresponds

to a genomic hotspot whose deletions were previously associated with epilepsy[119] and other neurological disorders. Finally, the remaining putative pathogenic CNVs were also associated with a number of genes (see Table S3.2). However, as we lack additional evidence for 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)[193]. Rare deletions in this gene were previously found in four independent individuals with attention-deficit hyperactivity disorder[200] and associated with intellectual disability[201]. In addition, *de novo* deletions were found in an individual with autism[117] and more recently in a patient with epileptic encephalopathy[124]. A common intronic variant in *PTPRD* was also associated with remission of seizures after treatment in a clinical cohort of epilepsy patients[202].

Table 3.3: **Recurrent CNVs with a pathogenic profile.**

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									

The 198 epileptic patients and 301 controls represent the discovery set. The replication set contains 325 epileptic patients and 380 controls.

3.5 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, including rare and small events.

A key aspect of our approach is the use of a set of reference samples to identify

abnormal read coverage. In this context, the choice and number of reference samples will have an effect on the analysis. Results from running PopSV using different reference cohort sizes suggest that CNV calls are consistent across runs but that a higher number of reference samples increases the sensitivity and robustness of the CNV detection (Fig. S3.18). Based on these results, we recommend PopSV when 20 samples or more can be used as reference. In a given study, all samples can be used as a reference, or a subset of a few hundreds if the total sample size is extremely large. Although variants with frequency around 50% might not be detected, PopSV excels at detecting less frequent variants, smaller variants or variants in challenging regions such as repeat-rich regions. In a case/control design, the control samples could be used as reference in order to maximize the detection of case-specific variants. In the current study we used both epilepsy patients and controls as reference in order to be able to directly compare the observed CNV distributions. Finally, in a cancer project with paired normal and tumor samples, only normal samples should be used as reference such that PopSV can detect somatic CNVs of any frequency.

To maximize performance, the same library preparation, sequencing and data pre-processing should be employed on all the samples. To identify potential batch effects, a principal component analysis of read coverage was implemented as part of the PopSV package and is recommended to assess the homogeneity of the reference samples. The read length and aligner can lead to drastic changes in the read coverage and should be consistent across the cohort when analyzed with PopSV. This is particularly important in repeat-rich regions. Although the different datasets were produced by different sequencing and pre-processing protocols and showed varying degrees of technical bias (Fig. 3.1a, S3.1 and S3.2), the performance of PopSV was comparable when benchmarking the methods in the two public datasets and experimentally validating calls in the CENet cohort.

PopSV's approach does not require a uniform read coverage and integrate the coverage variation separately in each studied region. For these reasons, it would be straightforward to analyze targeted sequencing data, such as exome-sequencing.

PopSV could also be extended for the detection of other types of SVs such as balanced SVs. To do this, instead of counting properly mapped reads, the method could be modified to test for an excess of discordant reads. Finally, additional modules could be added to PopSV to help characterize the detected variants. For instance, instead of computing a copy-number estimate from the average coverage in the reference, a HMM approach including all samples could provide a better genotyping strategy. Similar to other approaches[41,50], an additional step in the pipeline could explore the effect of the bin size on the variation in read coverage across the population and suggest an optimal bin size.

As in previous array-based studies[19,121,122], we observed an enrichment of large rare exonic CNVs in patients compared to controls. However, thanks to the resolution of WGS and PopSV, we found that the global distribution of small CNVs (<50 Kbp) in 198 unrelated epilepsy patients was also skewed towards rare exonic CNVs. In addition, genes disrupted by rare deletions in patients were enriched for previously known epilepsy genes. These observations support the association of small CNVs with epilepsy 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 epilepsy 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.

3.6 Materials and Methods

Ethics Statement

This study was approved by the Research Ethics Board at the Sick Kids Hospital (REB number 1000033784) and the ethics committee at the Centre Hospitalier Universitaire de Montréal (project number 2003-1394,ND02.058-BSP(CA)). Before their inclusion in this study, patients or parents (when needed) had to give written informed consents.

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 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 (±6.7) years. Table S3.3 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 ± 5.6x. Reads were aligned to reference Homo_sapiens b37 with BWA[203]. Finally, Picard was used to merge, realign and mark duplicate reads. Raw sequence data has been deposited in the European Genome-phenome Archive, under the accession code EGAS00001002825. For more details, see Supple-

mentary Information.

Public WGS datasets

Two high-coverage public datasets were used to benchmark PopSV against existing methods.

A *Twin* study provided WGS sequencing data for 45 individuals, including 10 monozygotic twin quartets from the Quebec Study of Newborn Twins[189]. All patients gave informed consent in written form to participate in the Quebec Study of Newborn Twins. Ethic boards from the Centre de Recherche du CHUM, from the Université Laval and from the Montreal Neurological Institute approved this study. DNA was extracted from blood and 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[203] (bwa version 0.6.2-r126-tpx with threading enabled). The options were `'bwa aln -t 12 -q 5' and 'bwa sampe -t 12'`. 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).

A cancer dataset from a study of renal cell carcinoma[143] was also used. 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[203] backtrack to the GRCh37 reference genome. Picard was used to adjust pairs coordinates, flag duplicates and merge lanes. Finally, realignment was done with GATK. Raw sequence data has been deposited in the European Genome-phenome Archive, under the accession code EGAS00001000083. More details can be found in Scelo et al.[143].

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. This analysis was performed separately on the CENet dataset, the Twin dataset and the normal samples from the cancer dataset. On the Twin dataset, the same analysis was also run after correcting the read coverage following the QDNAseq pipeline[190] (see Supplementary Information).

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. PopSV was implemented in an R package (see Data and code availability). The genome is first segmented into bins and the number of reads with proper mapping in each bin is 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 (Fig. 3.1b). 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 abnormal bins are merged and considered as one variant. In PopSV's R package, circular binary segmentation[204] can also be used to merge bins into variant regions. Copy number was estimated by dividing the coverage in a region by the average coverage across the reference samples, multiplied by 2 (see Supplementary Information).

Validation and benchmark of PopSV

We compared PopSV to CNVnator[41], FREEC[40] and cn.MOPS[42], three popular RD methods that can be applied to WGS datasets. We also ran LUMPY[67] 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. These methods were run on the two publicly available datasets, using 5 Kbp bins for the RD methods.

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 dataset. For example, on average, how many calls in one sample are called in more than 90% of the samples. In the Twin dataset, the samples were clustered using the CNV calls from each method. Different linkage criteria were used for the hierarchical clustering (see Supplementary Information). The Rand index estimated the concordance between the clustering and the known pedigree (family-level). Next, we measured the number of CNVs identified in each twin that were also found in their monozygotic 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. For CNVnator, LUMPY and PopSV, the eval1/eval2 columns, number of supporting reads and adjusted P-values (respectively) were used to gradually filter low-quality calls and explore their effect on the replication metrics. In addition to their replication, we annotated the calls when their region overlapped a call found by other methods in the same sample. For calls found by at least two methods, we computed the proportion of calls from a method found by each of the other methods.

The approach described previously comparing pairs of twins was also applied in the cancer dataset, 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 sizes. 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 sizes that were also found in the large bin calls.

CNV detection in the CENet cohorts

CNVs were called using PopSV using 5 Kbp bins and all the samples from both the epilepsy and control cohorts as reference. We annotated the frequency of the CNVs using germline CNV calls from the Twin and cancer datasets (internal database) as well as four public CNV databases from the 1000 Genomes Project[5,33], the Genome of Netherlands[107] and the Simons Genome Diversity Project[192]. CNVs were annotated with the maximum frequency in the databases. Hence, a rare CNV is defined as present in less than 1% of the samples in each of the five CNV databases.

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.

Validation by Taqman RT-PCR

We first selected CNV calls in epilepsy patients that spanned at least 2 consecutive bins. We kept exonic CNVs of different sizes and overlapping a Taqman probe. A second batch of CNVs, containing small non-coding CNVs, was also sent for validation. Here, 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[205]; the breakpoints remained unchanged when they could not 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 quadruplicate, 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 was 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.

CNV enrichment in exonic regions

For each cohort (epilepsy and control), 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. Because the epilepsy and control cohorts have different sample sizes, the CNV catalogs for each cohort were built using 150 randomly selected samples. 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 (see Supplementary Information). The fold-enrichment represents how much more/less of the CNVs overlap an exon compared to the control regions. To robustly compare the two cohorts, we computed the median difference in fold-enrichment between the CNV catalogs from patients and controls across 100 sub-sampled catalogs. The cohort labels of the CNV catalogs were then permuted 10,000 times and the analysis repeated to derive a null distribution for the median difference in fold enrichment. A permuted P-value was computed from the observed difference and the null distribution.

Small (<50Kbp) and large (>50 Kbp) CNVs were analyzed separately. Exons from genes predicted to be loss-of-function intolerant[193] (probability of loss-of-function intolerance > 0.9) were also analyzed separately. The same analysis was repeated using only 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).

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). 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 also visualize the proportion of CNVs in the catalog seen in 2 samples or more, 3 samples or more, etc (Fig. S3.12a). We performed the same analysis after removing the top 20 samples with the highest number of non-private rare exonic CNVs. The analysis was also repeated 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[118] which consists of 154 genes strongly associated with epilepsy. We tested different sets of CNVs: deletion or duplications in the epilepsy cohort, control individuals and samples from the twin study, and using different threshold of maximum frequency. For each set of CNVs, we counted how many of the genes hit were known epilepsy genes. To control for the 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 were epilepsy genes. After sampling 10,000 gene sets, we computed an empirical P-value (see Supplementary Information).

To investigate rare non-coding CNVs close to known epilepsy genes, we counted how many patients have such a CNV at different thresholds of distance to the nearest exon. We compared this cumulative distribution to the control cohort, after down-sampling it to the sample size as the epilepsy cohort. We performed the same analysis using deletions only. Each epilepsy gene was also tested for an excess of rare non-coding deletions in patients versus controls using a Fisher test. Next, we restricted our analysis to rare non-coding CNVs that overlap an eQTL associated with the epilepsy genes[195] or a DNase I hypersensitive site associated with the promoter of epilepsy genes[196]. A Kolmogorov-Smirnov test was used to test the difference in distribution. Finally, using different values for the maximum distance to the nearest epilepsy gene, we computed the odds ratio of having such a CNV 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[118] or deletions overlapping the exon of a loss-of-function

intolerant gene[193], or CNVs present in two or more of our epilepsy patients. All the putatively pathogenic CNVs were validated by Taqman RT-PCR.

Data and code availability

The PopSV R package and its documentation are available at http://jmonlong.github.io/PopSV/. Scripts are provided to run the pipeline on different high performance computing systems. 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.1172181. Necessary data, including the CNV calls, was deposited at https://figshare.com/s/20dfdedcc4718e465185. Raw sequence data has been deposited in the European Genome-phenome Archive, under the accession code EGAS00001002825.

3.7 Acknowledgments

Data analyses were enabled by compute and storage resources provided by Compute Canada and Calcul Québec. We would like to thank Pascale Marquis at the Canadian Centre for Computational Genomics for processing the raw sequencing data to genomic variant calls and for her active participation in various quality assessment steps. The Canadian Centre for Computational Genomics (C3G) is a Node of the Canadian Genomic Innovation Network and is supported by the Canadian Government through Genome Canada. 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 Dobrzeniecka 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 and Claudia Moreau for helpful discussions.

Chapter 4

Population-Based Detection of CNVs in Low-Mappability Regions

Preface: Bridging Text between Chapters 3 and 4

The method implemented in the previous chapter was used to study CNVs in low-mappability region. The main challenge with low-mappability regions is the complex coverage profiles. Because of the repeated sequences, the number of uniquely mapped reads is reduced and fluctuates depending on which repeats are present in each region. Instead of attempting to model this complex sequence contexts, we use the population-based approach implemented in chapter 3, PopSV, to robustly call CNVs in these challenging regions. Indeed, the complex coverage profiles in low-mappability regions are consistent across samples from the same sequencing project. Hence, if the read coverage in a sample is different enough from the reference samples, it is likely due to a CNV.

This manuscript was published in *Nucleic Acids Research*[3]. First, we showed that PopSV performance was preserved in low-mappability regions by consistently showing a high sensitivity and a stable false positive rate across different repeat profiles. Using CNVs across 640 normal genomes we also showed that CNVs are enriched in low-mappability regions independently from the known enrichment in segmental duplication. Thanks to the population-based approach, we provided a catalog of germline CNV that includes repeat-rich regions, including thousands of regions with recurrent variants that were missing from existing CNV databases.

Appendix D contains supplementary tables, graphs and information.

Human copy number variants are enriched in regions of low mappability

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

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

4.1 Abstract

Copy number variants (CNVs) are known to affect a large portion of the human genome and have been implicated in many diseases. Although whole-genome sequencing (WGS) can help identify CNVs, most analytical methods suffer from limited sensitivity and specificity, especially in regions of low mappability. To address this, we use PopSV, a CNV caller that relies on multiple samples to control for technical variation. We demonstrate that our calls are stable across different types of repeat-rich regions and validate the accuracy of our predictions using orthogonal approaches. Applying PopSV to 640 human genomes, we find that low-mappability regions are approximately 5 times more likely to harbor germline CNVs, in stark contrast to the nearly uniform distribution observed for somatic CNVs in 95 cancer genomes. In addition to known enrichments in segmental duplication and near centromeres and telomeres, we also report that CNVs are enriched in specific types of satellite and in some of the most recent families of transposable elements. Finally, using this comprehensive approach, we identify 3,455 regions with recurrent CNVs that were missing from existing catalogs. In particular, we identify 347 genes with a novel exonic CNV in low-mappability regions, including 29 genes previously associated with disease.

4.2 Introduction

Genomic variation of 50 base pairs or more are collectively known as structural variants (SVs) and can take several forms including deletions, duplications, novel insertions, translocations and inversions[4]. Copy number variants (CNVs) are unbalanced SVs, i.e. affecting DNA copy number, and include deletions and any type of duplications (tandem duplications, triplications and other amplifications). A wide range of mechanisms can produce SVs and is responsible for the diverse SV distribution across the genome, both in term of location and size[4,10,11]. In healthy individuals, SVs are estimated to cumulatively affect a higher proportion of the genome as compared to single nucleotide polymorphisms (SNPs)[104]. SVs have been associated with numerous diseases including Crohn's Disease[20], schizophrenia[17], obesity[16], epilepsy[19], autism[18], cancer[21] and other inherited diseases[22,23], and many SVs have a demonstrated detrimental effect.

While large SVs have been first studied using cytogenetic approaches and array-based technologies, whole-genome sequencing (WGS) is in theory capable of detecting SVs of any type and size[39]. Numerous methods have been implemented to detect SVs from WGS data using either paired-end information[43,44], read-depth (RD) variation[40,41,42], breakpoints detection through split-read approach[45] or de novo assembly[46]. CNVs, potentially the most impactful SVs, can be detected by any of these strategies but are often resolved with a RD approach as it directly looks for signs of copy number changes. However, several features of WGS experiments result in technical bias and continue to be a major challenge. For example, GC content[50], mappability[48,49], replication timing[51], DNA quality and library preparation[177] have a detrimental impact on the uniformity of the RD[52]. Unfortunately, this variability is difficult to fully correct for as it involves different factors, some of which are unknown, that vary from one experiment to another. This issue particularly impairs the detection of CNV with weaker signal, which is inevitable in regions of low-mappability that represent around 10% of the human genome[206], for smaller CNVs or in cancer samples with cell heterogeneity or stromal contamination.

As a result, existing approaches suffer from limited sensitivity and specificity[11,39], especially in regions of low-complexity and low-mappability[48,49]. Even when problematic regions were masked and state-of-the-art bias correction[50,190] were applied, we showed that technical variation in RD could still be found across three WGS datasets studied[2].

To control for technical variation, we recently developed a CNV detection method, PopSV, which uses a set of reference samples to detect abnormal RD[2]. In each genome tested, the RD in a region is compared to the same region in the reference samples. PopSV differs from most previous RD methods, such as RDXplorer[207] or CNVnator[41], 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[40] or BIC-seq[191], we showed that they do not sufficiently control for experiment-specific noise as compared to PopSV[2]. Glusman et al.[176] does go further and normalize the RD with pre-computed RD profiles that fit the GC-fingerprint of a sample but this approach excludes regions with extreme RD and does not integrate the variance observed in individual regions. PopSV is also different from approaches such as cn.MOPS[42] and Genome STRiP[33] that scan simultaneously the genome of several samples and fit a Bayesian or Gaussian mixture model in each region. Those methods have more power to detect CNVs present in several samples but may miss sample-specific events. Moreover, their basic normalization of the RD and fully parametric models forces them to conceal a sizable portion of the genome and variants with weaker signal. Finally, another strategy to improve the accuracy of CNV detection has been to use an ensemble approach that combines information from different methods relying on different types of reads. Large re-sequencing projects such as the 1000 Genome Project[5,11] and the Genomes of Netherlands (GoNL) project[107,208] have adopted this strategy and have successfully identified many CNVs using an extensive panel of detection methods combined with low-throughput validation. Such a strategy increases the specificity of the calls at the cost of sensitivity.

Notably, with most of the tools and approaches described above, repeat-rich regions and other problematic regions of the genome are often removed or smoothed at some step of the analysis, to improve the accuracy of the calls. Although some methods[82,83] try to model ambiguous mapping and repeat structure, only particular situations are addressed and, as a consequence, low-mappability regions are just scarcely covered in the most recent CNV catalogs[5]. This is unfortunate given that CNVs in such regions have already been associated with various diseases[23,87,90,209,210] and that these regions are also more likely variable. Indeed, different types of genomic repeats are likely to contribute to CNV formation. For example, CNVs are known to be enriched in segmental duplications[10] and short and long tandem repeats are also known to be highly polymorphic[211,212]. Moreover, repeat templates, like segmental duplications or transposable elements, can facilitate the formation of CNV through non-allelic homologous recombination and other mechanisms[213].

Given these facts and the growing realization of the importance of repetitive regions in the genome[81,214], we wanted to investigate the performance of PopSV in low-mappability regions and explore the comprehensive CNV distribution across a large cohort of healthy individuals. After showing that population-based RD measures are better than existing mappability estimates to correct for variable coverage, we apply PopSV to 640 WGS individuals from three human cohorts. We compare the performance of PopSV on these datasets with existing CNV detection methods in regions of low-mappability and validate the quality of the predictions across different repeat profiles using PCR validation. Additionally, using publicly available long-read sequencing data and assemblies, we show that PopSV is able to detect some highly ambiguous CNVs. Next, having demonstrated the quality of the PopSV calls, we characterize the patterns of CNVs across the human genome and produce a CNV catalog where variants of different types are better represented compared to existing catalogs. We further find that CNVs are significantly enriched in regions of low-mappability and in different classes of repeats. Finally, we identify novel CNV regions in low-mappability regions that were absent from previous CNV catalogs

and describe their impact on protein-coding genes.

4.3 Material and Methods

Data Three publicly available WGS datasets were used. The first is a twin study[189] with an average depth of 40x across 45 French-Canadian individuals, including 10 families of parents and monozygotic twins. The second is a renal cell carcinoma dataset[143] (CageKid) with 95 tumor/normal pairs from four European countries and an average depth of 54x. The third contains 500 unrelated Dutch individuals from the GoNL[107] dataset with an average depth of 14x. In each study, the sequenced reads had been aligned using bwa[203]. See Supplementary Information for more details on access and read processing.

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

RD and mappability estimates To compare RD and mappability estimates in the Twin study, we first removed bins with extremely high RD if deviating from the median RD by more than 5 standard deviation. The RD across the different samples were then combined and quantile normalized. For each bin, we computed the average RD and standard deviation across the samples. We downloaded the mappability track for hg19[206] and computed the average mappability in each bin. We compared the RD in one randomly selected sample with the mappability estimates and with the inter-sample RD average. To correct for the variation explained by the mappability estimates we fitted a generalized additive model using a cubic

regression spline between the mappability estimates and RD in the sample (see Supplementary Information). With these estimations and the global standard deviation we computed a Z-score for each bin. A similar set of Z-scores was computed using the inter-sample average and standard deviation. The normality of these two Z-score distributions were compared in term of excess kurtosis and skewness. The Z-score distributions were also compared in different mappability intervals. Finally, 45 samples of each cohort were combined and their RD quantile normalized. The inter-sample RD mean and standard deviation were then computed separately in each cohort and compared with the mappability estimates and RD in the selected sample.

PopSV approach for CNV detection PopSV was first described and applied in a CNV analysis of epilepsy patients[2]. Briefly, a set of samples are chosen as reference and used to guide the normalization of each bin. After normalization the average RD and standard deviation in each bin are saved and used to transform the RD in all samples into Z-scores. CNVs are called in each sample when the RD is significantly higher or lower than in the reference samples. The Z-scores can be segmented using the circular binary segmentation[204] or after statistical testing at the bin level. As recommended, PopSV was run separately on each dataset to avoid false positives due to potential variation in sequencing protocols. More details are available in the original publication[2] and in the Supplementary Information. With PopSV there is no filtering, masking, smoothing or altering of repeat-rich regions: all the regions with properly mapped reads are analyzed.

Coverage track and low-mappability regions The average RD in the reference samples, a feature used during CNV calling, was used as a coverage track. Bins with a RD lower than 4 standard deviation from the median were classified as *low-mappability* (or *low coverage*). To highlight the most challenging region, we also defined *extremely low coverage* regions if the average RD was lower than 100 reads. We overlapped these regions with protein-coding genes and segmental duplications

(see Supplementary Information), and computed the distance to the nearest centromere, telomere or assembly gap. We also counted the number of protein-coding genes overlapping at least one low-coverage region.

CNV detection using other methods FREEC[40] and CNVnator[41] were run on each sample separately starting from the BAM files and using the same bin size as for PopSV (5 Kbp). cn.MOPS[42] was run on the same GC-corrected bin counts than for PopSV and samples from the same dataset were jointly analyzed. After retrieving split reads using YAHA[66], LUMPY[67] was run and we kept all the deletions and duplications larger than 300 bp. BND variants with both ends more than 300 bp apart in the same chromosome were also included as they could be CNVs lacking support to characterize their type properly. See Supplementary Information for more details.

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

Replication in twins For each twin and each method, a CNV call was defined as *replicated* if also found in the other monozygotic twin but in less than 50% of the population to remove systematic errors. The frequency was computed by counting samples with any overlapping CNVs. In order to avoid missing calls with borderline significance, we used slightly less confident calls for the second twin (see Supplementary Information). For each method, we computed the number and proportion of *replicated* calls per sample. We computed these metrics using all the calls, calls

in low-mappability regions only, calls in segmental duplications, calls overlapping annotated repeats and calls overlapping annotated satellites, all using a minimum overlap of 90% of the call's sequence. Finally, we computed the replication estimates for calls located at 1 Mbp or less from a centromere, telomere or assembly gap.

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

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

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

Analysis of CEPH12878 High coverage PCR-free Illumina WGS data for 30 samples, including CEPH12878, was downloaded from the 1000 Genomes Project (1000GP)[5] (see Supplementary Information). PopSV was run using 5 Kbp bins and all the samples as reference. Using the same coverage track as before we selected all deletions in CEPH12878 overlapping low-mappability regions (at least 90% of the call). We first looked for support in CEPH12878 assemblies that used Illumina short-read sequencing, BioNano Genomics genome maps and either single molecule sequencing from the Pacific Biosciences (PacBio) platform[59] or 10X Genomics linked-read sequencing[215]. For each selected deletion from PopSV, we aligned the flanking reference sequences to the assemblies using BLAST[216] (see Supplementary Information). When both flanks could be mapped to a contig, we visually inspected MUMmer plots[217] which either supported the deletion, the reference genome sequence or were too noisy to assess. We further annotated the selected calls if they overlapped with the deletions identified in Pendleton et al.[59] over a minimum of 1 Kbp. Finally, we downloaded the corrected PacBio reads and built a local assembly and consensus around each selected PopSV deletion (see Supplementary Information). We visually inspected MUMmer plots of the assembled and consensus sequences to confirm the presence of the deletion.

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

Comparison with public CNV catalogs We retrieved autosomal deletions, duplications and CNVs from four public CNV catalogs derived from large-scale WGS surveys: the 1000GP SV catalog[5], Genome STRiP's catalog from 847 individuals[33], Genome STRiP calls in 148 high-depth WGS genomes[36], and the GoNL SV catalog[107] (see Supplementary Information). To compare the amount of CNV with PopSV, we removed deletions smaller than 300 bp as well as variants with high frequency ($> 80\%$). We compared CNV frequency between the 620 unrelated samples and a down-sampled set of 620 randomly selected individuals from the 1000GP CNV catalog. The frequency was derived for all the nucleotide that overlaps at least one CNV as the proportion of individuals with a CNV in this locus. The frequency distribution was computed separately for the different CNV types.

Comparison with CNV catalogs from long-read studies The SV catalog from Chaisson et al.[58] was downloaded and overlapped with the CNV catalogs from 1000GP and PopSV results on our 640 genomes. Here, the 1000GP catalog contained deletions, duplications and CNVs of any size and frequency. Using control regions and logistic regression we tested for an enrichment of variants in the SV catalog from Chaisson et al.[58] (see Supplementary Information). The analysis was performed separately on deletions, duplications, low-mappability regions and extremely low-mappability regions. The same analysis was performed using the SV catalog from Pendleton et al.[59].

Novel CNV regions Using the 620 unrelated individuals across the three cohorts, we selected CNVs present in more than 1% of the population (7 individuals or more) and not overlapping any CNV in the 1000GP catalog[5]. We used deletions, duplications and CNVs of any size and frequency from the 1000GP. Novel CNVs were collapsed into novel CNV regions, i.e. contiguous regions in which each base is overlapped by at least one novel CNV. The novel CNV regions were annotated using the low-mappability and extremely low-mappability tracks. We also compared CNVs from the three other public CNV catalogs to the novel CNV regions.

Distance to centromere, telomere and assembly gaps The centromeres, telomeres and assembly gaps (CTGs) were retrieved from the gap track in UCSC[218]. In chromosomes with missing telomere annotation, we defined the telomere as the 10 Kbp region at the ends of chromosome. The distance from each variant to the nearest CTG was computed and represented as a cumulative proportion. Because this distribution changes with the size of the variants, we sampled random regions in the genome with similar sizes and computed the same distance distribution (see Supplementary Information). Thanks to this null distribution we were able to see if variants were located closer/further to CTG than expected by chance.

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

Overlap with gene annotation Exons of protein-coding genes and promoter regions (10 Kbp upstream of the transcription start site) were extracted from the Gencode annotation v19. We counted how many genes overlapped a CNV in the

population when considering exons only, exons and promoter region, or gene body and promoter region. In addition, we computed these numbers using only genes associated with a disease or phenotype in the OMIM Morbid Map (Online Mendelian Inheritance in Man; http://omim.org/). These numbers were also computed for CNVs that overlapped more than 90% of various classes of repeats. For example, Satellite-CNVs are CNVs with more than 90% of their region annotated as satellites.

4.4 Results

Modeling RD using population-based measures instead of mappability scores

When counting uniquely mapped reads, the mappability of a region is a major predictor of the observed RD. Theoretical mappability estimates[206] strongly correlated with the RD in a sample but many regions with intermediate mappability diverged from the predicted levels of RD (Fig. S4.1a). By computing the average RD across the 45 samples from the Twin study in each 5 Kbp bin we found that this divergence is consistent across samples and not simply due to a high RD variance (Fig. 4.1a). These mappability estimates only approximate RD variation and cannot explain the RD profile in numerous regions. In contrast, population-based metrics more directly estimate the expected RD level (Fig. S4.1b). Similarly to what was done in Monlong et al. in high-mappability regions[2], we hypothesized that population-based estimates of RD mean and standard deviation could be used directly and help analyze regions with reduced RD. To test this hypothesis, Z-scores corrected by the mappability-based estimates were compared to Z-scores derived from both the inter-sample mean and standard deviation. The population-based Z-scores better followed a Normal distribution with an excess kurtosis of 0.2 and skewness of 0.004 compared to 29.4 and -2.284 respectively for mappability-adjusted Z-scores (Fig. 4.1b). The distribution of the population-based Z-scores was also more stable across the mappability spectrum (Fig. 4.1c). When comparing samples from the three dif-

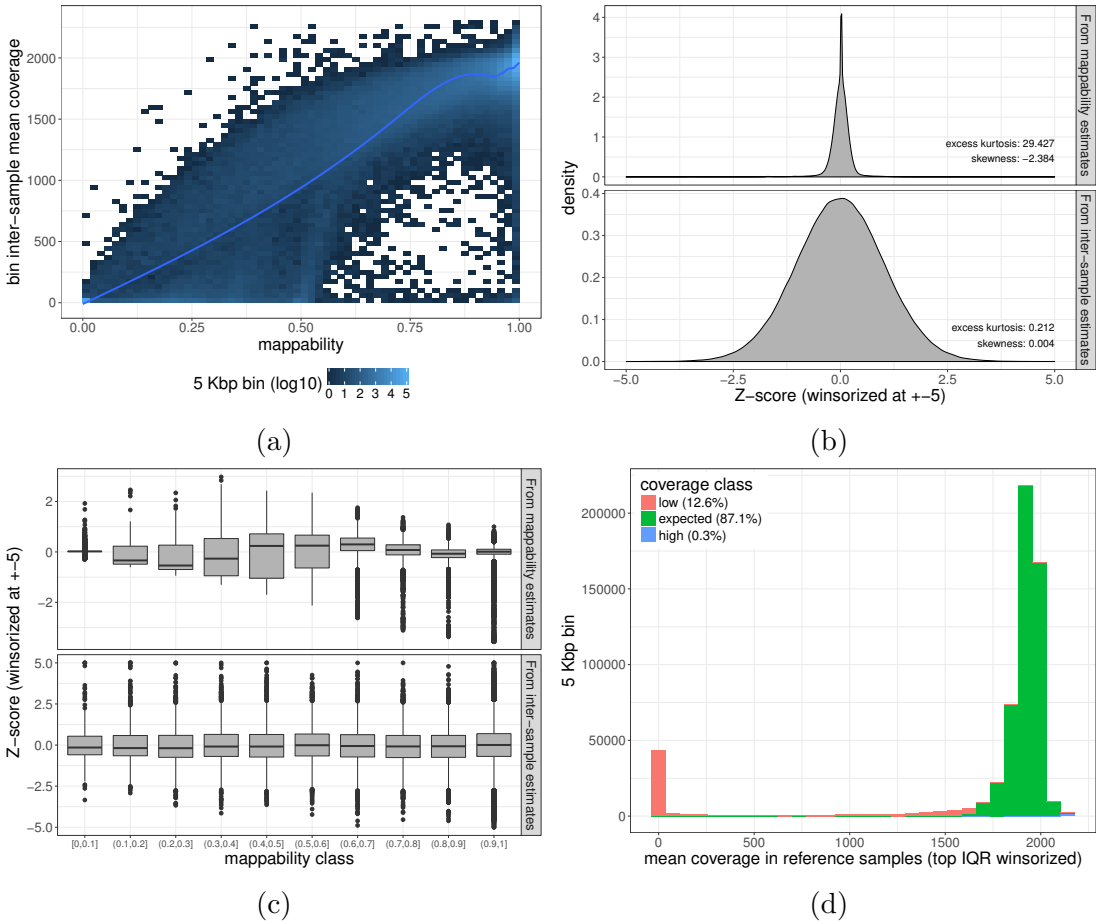


Figure 4.1: **Mappability and population-based RD estimates.** a) Inter-sample mean RD and average mappability in 5 Kbp bins. Regions with the same mappability estimate can have different RD levels. b) Z-score distribution. In *mappability*, Z-scores were computed from the mappability-predicted RD and global standard deviation; In *population estimates* from the inter-sample mean and standard deviation. c) Z-score distribution across the mappability spectrum. d) Average RD in the Twin study. The right-tail of the histogram was winsorized using the IQR and the different coverage classes are shown with colors.

ferent datasets, we noticed cohort-specific profiles in term of RD level and variance even though RD had been quantile normalized (Fig. S4.1c and S4.1d), suggesting that population-based estimates will be better at capturing subtle cohort-specific variation.

These results suggest that a population-based strategy such as PopSV[2] could be extended to investigate CNVs in regions of low-mappability. To define low-mappability regions in the population, we used the average RD in the reference samples track produced by PopSV. In the Twin study for example, 12.6% of the covered 5 Kbp bins were labeled as low-coverage (Fig. 4.1d), more than half of which were regions with extremely low coverage (lower than 100 reads on average). Slightly fewer regions were labeled as low-coverage in the other cohorts (Fig. S4.2). As expected, low-coverage regions were depleted in gene content with only 15.3% of the 5 Kbp bins in these regions overlapping a protein-coding gene versus 48.8% for other regions. Nonetheless, 4,044 protein-coding genes overlapped a low-coverage region. Finally, 23.2% of the low-mappability regions overlapped segmental duplications and 69.1% were located at less than 1 Mbp from a centromere, telomere or assembly gap, versus respectively 2.9% and 8.8% for other regions.

Replication rates in regions of low-mappability

We previously demonstrated that CNV detection with PopSV was overall more sensitive than FREEC[40], CNVnator[41], cn.MOPS[42] and LUMPY[67] methods[2]. In the following, we focused on the performance of PopSV in low-mappability regions. We first investigated the general concordance of the CNV calls with the pedigree in the Twin study. Using calls in extremely low-mappability regions (average RD below 100 reads) only, we clustered the individuals and compared the result to the known pedigree. We found that PopSV showed better concordance, as assessed by the Rand index (Fig. S4.3), compared to the other methods. Indeed, the clustering dendogram from PopSV calls, even in these challenging regions, captured almost perfectly the family relationships (Fig. 4.2a). We then investigated if the call replication rate was

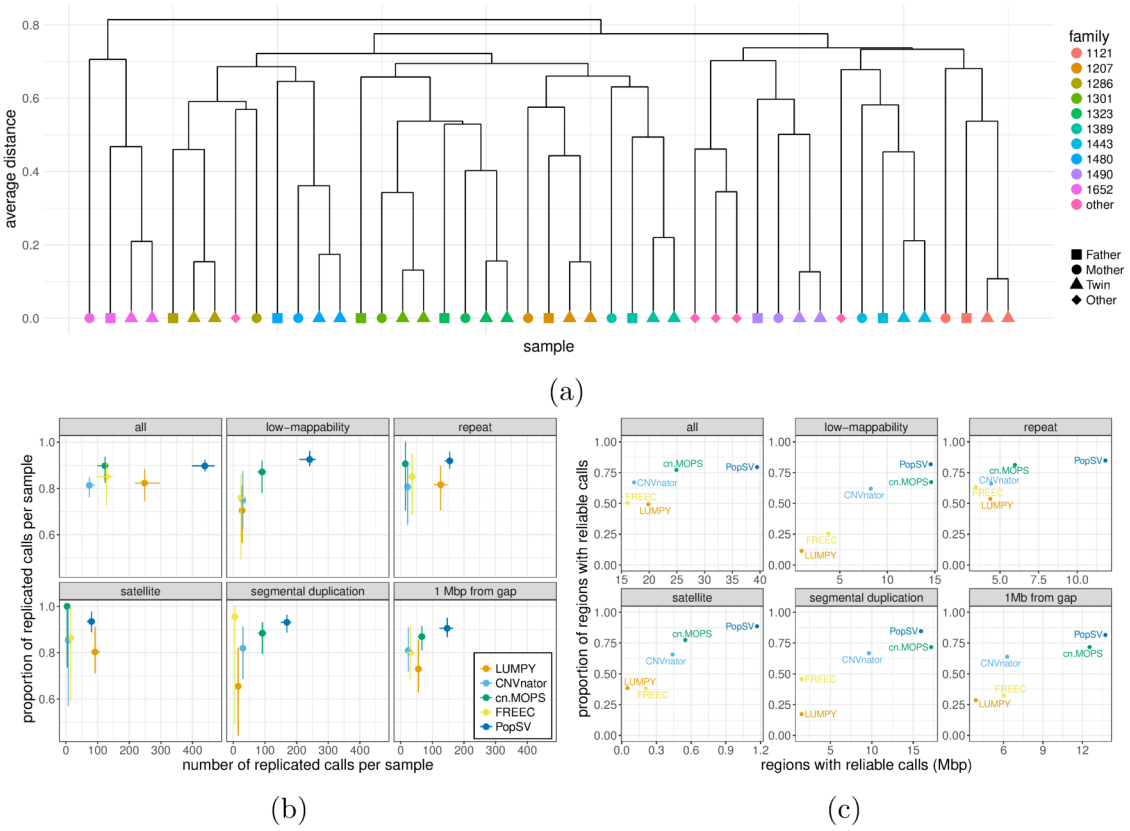


Figure 4.2: **PopSV's performance in low-mappability regions.** a) Cluster using PopSV calls in extremely low coverage regions (below 100 reads). b) Proportion and number of calls replicated in the monozygotic twin. The point shows the median value per sample, the error bars the 95% confidence interval. c) Proportion and number of regions with reliable calls, computed from call replication in twins.

stable across different mappability profiles. Using calls present in less than 50% of the population to avoid systematic bias, the overall replication rate in the other twin was found to be 89.7%. Focusing on calls in low-coverage regions, we found a comparable replication rate of 92.5%. The replication rate remained constant in regions with different repeat profiles (Fig. 4.2b) such as regions overlapping segmental duplication, annotated repeats, or close to centromeres, telomeres and assembly gaps. In contrast, the other methods showed a reduced replication and higher variance in repeat-rich regions. The superior replication rate was complemented by a larger number of calls: PopSV called between 2.7 and 9.9 times more replicated CNVs per sample in low-coverage regions compared to the other methods. We observed the same results in the cancer dataset when comparing the agreement between germline events in normal/tumor pairs. PopSV had between 1.8 and 17.8 times more calls in low-mappability regions compared to the other methods and a stable replication rate across repeat profiles (Fig. S4.4). We next wanted to assess the performance in each region of the genome, rather than overall rates per sample, and used the replication in twins to identify regions with reliable calls. Again we observed that PopSV was as reliable overall as in regions with different repeat profiles (Fig. 4.2c). This analysis also showed that PopSV provides reliable calls in a larger fraction of the genome compared to other methods. The strongest gain was observed for regions overlapping satellites or overlapping almost completely annotated repeats, with around twice as many regions reliably called by PopSV. cn.MOPS showed the second best performance, especially in regions overlapping segmental duplications or close to centromeres, telomeres and assembly gap.

Validation of CNVs in regions of low-mappability

Using Real-Time PCR validation across 151 regions, we previously demonstrated that the replication estimates from the Twin dataset are consistent with experimental validation[2]. We had tested variants of different types, sizes and frequencies and validated 90.7% of the calls, similar to our twin-based replication estimates. Here

we tested additional deletions in individuals from the Twin study using PCR validation. We first validated randomly selected deletions and found a validation rate close to the overall replication rate, with 18 out of 20 deletions (90%) successfully validated (Table S4.1). In a second validation batch, we focused on rare deletions in low-mappability regions, of which 11 out of the 17 (65%) were successfully validated (Table S4.2). We noticed that the majority of the non-validated deletions were predicted to be smaller than 100 bp and most likely due to a problem during the breakpoint fine-tuning. If we consider only deletions larger than 100 bp, the validation rate in regions of low-mappability increased to 83% (10/12) once again close to PopSV's replication rates in the Twin dataset.

Regions with extreme repeat content remained difficult to target and validate using PCR approaches. To further interrogate the performance of PopSV in those regions, we turned to whole-genome data from long-read sequencing technology. Publicly available assemblies for CEPH12878 samples confirmed several deletions called by PopSV in low-mappability regions. Out of the 14 homozygous deletions that could be assessed, 13 were confirmed in a contig, 12 of which were observed in both assemblies[59,215]. Only one region seemed to be a false positive, an assembled contig supporting the reference sequence in one assembly. Eleven regions could not be assessed because the flanks in the reference genome didn't map to any assembled contigs or their MUMmer plots neither supported a deletion nor the reference sequence. In summary, we confirmed 92.8% of the homozygous deletions in low-mappability regions that could be compared with the assemblies. Deletions can be confirmed by direct comparison of the variant region and, if homozygous, should be present in the assembly. In contrast, heterozygous deletions could be missing from an assembly if only the reference allele was assembled. We confirmed 27 out of the 44 heterozygous deletions in low-mappability regions that could be assessed (Table S4.3). As expected, only one allele was supported for many regions: 16 regions with only the deleted allele observed and 17 regions with only the reference allele observed. Both deleted and reference alleles were observed for 11 variants. Although

only 61.3% of the heterozygous deletion were confirmed, many variants might have been missed because of assembly preference to one allele, as suggested by the similar number of regions with only one supported allele. Using variants identified by Pendleton et al.[59] and by assembling raw PacBio reads, we found support for 3 additional homozygous deletions and 15 heterozygous deletions that had remained inconclusive in the assembly comparison. Most of the regions that couldn't be confirmed were located close to assembly gaps in the reference genome (Fig. S4.5). This observation highlighted that even with long-read sequencing data, it is not straightforward to clearly assess some genomic regions close to assembly gaps.

Global patterns of CNVs across the human genome

Having demonstrated the robustness of PopSV in low-mappability regions, we wanted to characterize the global patterns of CNVs across the human genome. We were especially interested in looking at calls in regions of low-mappability which represents between 9-12% of the human genome (Fig. 4.1d and S4.2). We started with an analysis of the twins and the normal samples in the renal cancer dataset, both of which have an average sequencing depth around 40X. PopSV was used to call CNV using 500 bp and 5 Kbp bins, which were then merged to create a final set of variants. On average per genome, 7.4 Mbp of the reference genome had abnormal read coverage, 4 Mbp showing an excess of reads indicating duplications and 3.4 Mbp showing a lack of reads indicating deletions (Table 4.1). In both datasets, the average variant size was around 3.7 Kbp and 70% of the variants found were smaller than 3 Kbp. We compared our numbers to equivalent CNVs detected in the most recent human SV catalog from the 1000 Genomes Project (1000GP), where 6.1 Mbp was found to be copy-number variable on average in each genome (Table S4.4). In those calls, we notice that no variants except for a few deletions were identified in regions of extremely low-mappability regions. Similarly, small duplications (< 3 Kbp) were absent from that catalog. In contrast, the set of variants identified by PopSV included variants in extremely low-mappability regions as well as small deletions and

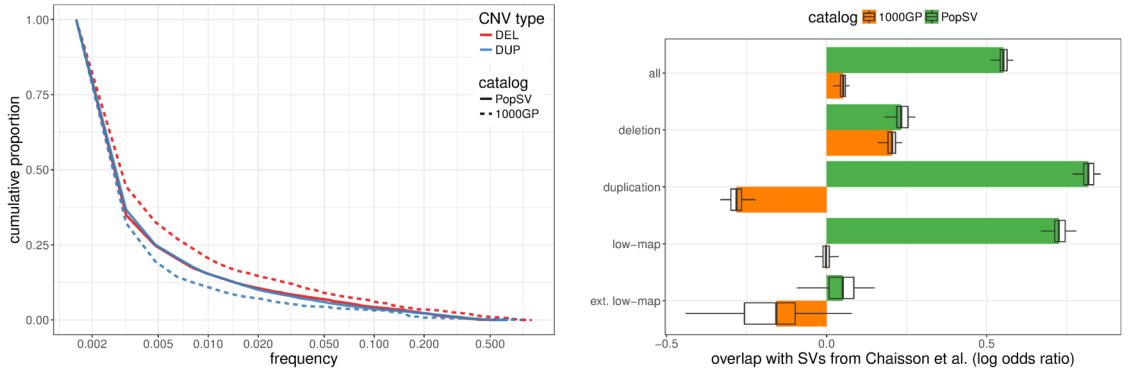
Set	Depth	Samples	Variants			Avg Size (Kbp)	Variants <3 Kbp		Affected genome (Mbp)			
			Total	Per sample			Proportion	Per sample	Total	Per sample		
				WG	*ELC*					*min*	*mean*	*max*
Twin study	42x	45	20,222	1,637.27	243.24	4.21	0.65	1,056.84	62.22	5.30	6.89	9.03
deletion			10,661	727.04	13.20	4.53	0.58	423.80	33.97	2.79	3.30	3.85
duplication			10,396	910.22	230.04	3.94	0.70	633.04	34.20	2.50	3.59	5.29
CageKid normals	40x	95	56,256	2,132.81	336.46	3.58	0.71	1,521.16	134.77	5.53	7.63	10.24
deletion			25,367	805.08	12.74	4.30	0.63	508.56	70.65	2.65	3.46	7.26
duplication			32,356	1,327.73	323.73	3.14	0.76	1,012.60	76.28	2.31	4.17	6.70
GoNL	13x	500	27,945	549.52	81.97	8.71	0.46	250.24	226.50	3.05	4.79	8.16
deletion			13,818	262.41	1.45	8.50	0.42	110.16	106.83	1.30	2.23	3.96
duplication			15,291	287.10	80.52	8.91	0.49	140.08	139.21	1.45	2.56	5.72

Table 4.1: **CNVs in the Twins, CageKid normals and GoNL datasets.** WG: whole genome; ELC: extremely low-coverage regions. The *Total* number of variants is the total number after collapsing recurrent variants. *Affected genome* represents the amount of the reference genome that overlaps at least one CNV.

duplications (Table 4.1), explaining in part the $\sim 20\%$ increase in affected genome. While the study from the 1000GP[5] explored a wider range of SVs, our catalog is likely more representative of the distribution of CNVs in a normal genome since a larger portion of the genome could be analyzed. Small duplications and events in low-mappability regions were also under-represented in more recent CNV surveys that used higher sequencing depth or joint-calling of CNVs[33,36,107] (Table S4.4), confirming the uniqueness of the PopSV catalog.

Next, we applied PopSV to the 500 unrelated samples from the GoNL cohort (Table 4.1). Due to a lower sequencing depth ($\sim$13X), we used bins of size 2 Kbp and 5Kbp, explaining the lower number of variants found in these samples. Nevertheless, a large sample size helps better characterize the frequency patterns and provides a more comprehensive map of rare CNVs. In total, across these three cohorts, 325.6 Mbp were found to be affected by a CNV with more duplications (50,856) detected than deletions (44,110). This contrasts with the CNVs reported by the 1000GP[5] that were heavily skewed towards deletions (Table S4.4), likely due to the conservative ensemble approached used to detect CNVs. The frequency distribution of deletions and duplications found using PopSV were also much more balanced compared with the ones from the 1000GP[5] (Fig. 4.3a).

We also compared our CNV catalog with an orthogonal set of calls from Chaisson et al.[58] that were obtained using long-read sequencing. Although these calls came from a different genome, we expect both catalogs to share a number of common variants. We found a significant overlap between the two catalogs, overall and separately for deletions, duplications, low-mappability regions and extremely low-mappability regions (Fig 4.3b). In all categories, the overlap was stronger for PopSV's catalog compared to the 1000GP CNV catalog. We noted that the enrichment for the 1000GP catalog disappeared for duplications and low-mappability regions but was even stronger for PopSV's catalog. Like PopSV, the long-read sequencing study[58] also found a better balance between deletions and duplications. Similar observations were made using another set of calls from long-read sequencing of the CEPH12878



(a) (b)

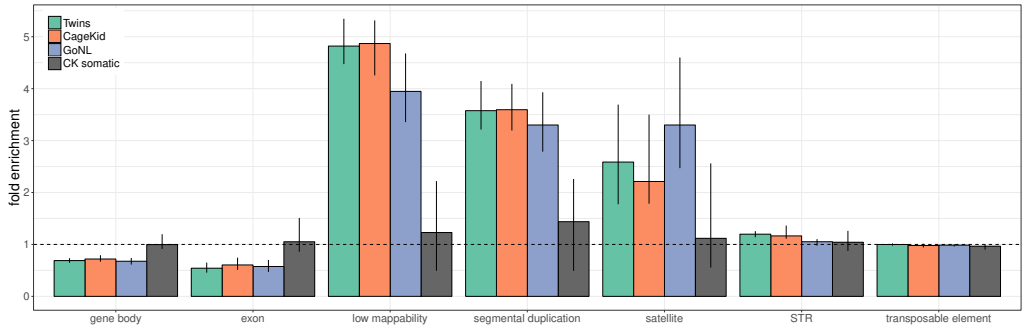
Figure 4.3: **Comparison with CNV catalogs from the 1000 Genomes Project[5] (1000GP) and a long-read sequencing study[58].** a) The x-axis represents the proportion of individuals with a CNV overlapping a region. The y-axis represents the cumulative proportion of the affected genome. b) Overlap with the SV catalog from Chaisson et al.[58]. In each cohort (color), the proportion of collapsed calls overlapping calls from Chaisson et al.[58] or control regions with similar size distribution was modeled using a logistic regression. Boxplots show variation across 50 sampling of control regions. *low-map*: calls in low-mappability regions; *ext. low-map*: calls in extremely low-mappability regions.

sample[59] (Fig. S4.6).

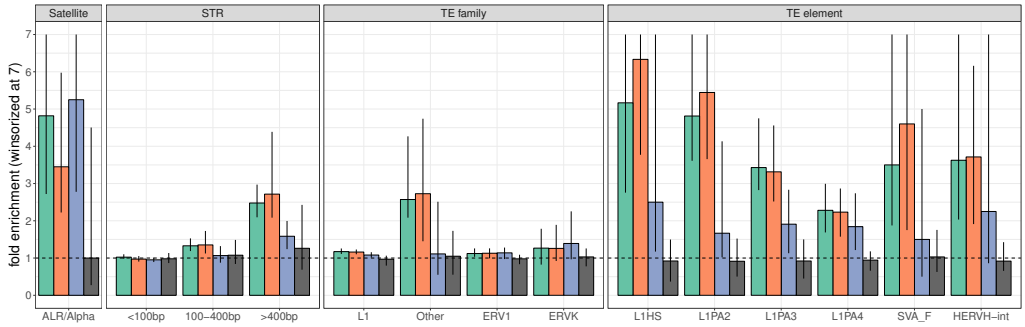
CNVs are enriched near centromeres and telomeres and in regions of low-mappability

Large CNVs have been shown to be enriched near centromeres, telomeres and assembly gaps (CTGs)[219]. 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. S4.7). In an average genome, we found that 33.5% of the CNVs calls were within 1 Mbp of a CTG, while we would have expected only 11.2% by chance. To verify that these observations were not simply a consequence of the methodology used, we also looked at the somatic CNVs (sCNVs) that we could detect in the renal cell carcinoma dataset. For this purpose, we extracted the variants found by PopSV in the tumor sample of an individual but missing from its paired normal sample. Reassuringly, and in contrast to germline CNVs, sCNVs were not preferentially found near CTGs (Fig. S4.7), with 11.1% of the sCNVs within 1 Mbp of a CTG.

After correcting for the distance to CTGs, we also observed a 4.7 fold-enrichment of variants in regions of low mappability (Fig. 4.4a). Segmental duplications (SD),



(a)



(b)

Figure 4.4: **CNVs in normal genomes.** a) Enrichment of CNVs in different genomic classes (x-axis) across different cohorts (colors) and controlling for the distance to centromere/telomere/gap. Bars show the median fold enrichment compared to control regions. The error bar represents 90% of the samples in the cohort. b) Enrichment of CNVs in repeat families (x-axis) controlling for the overlap with segmental duplication and distance to centromere/telomere/gap. The error bars were winsorized at 7 for clarity. *STR: Short Tandem Repeat*; *TE: Transposable Element*.

DNA satellites and Short Tandem Repeats (STRs) were also significantly enriched with fold-enrichment of 3.6, 2.6 and 1.2, respectively. The over-representation of CNVs in SDs has been described before[10] and in a recent study[192], half of the CNV base pairs were shown to overlap a SD. To investigate the contribution of low-mappability regions beyond SDs, we used matched control regions and included segmental duplication overlap in the logistic regression model. Even after controlling for this known enrichment, we found that CNVs overlapped low-coverage regions more than twice as much as expected (Fig. S4.8a). This two-fold enrichment is independent of the SD association and consistently observed in the 3 cohorts of

normal genomes. In contrast to germline CNVs, sCNVs were once again found to be more uniformly distributed (Fig. 4.4a and S4.8a). These results suggest that the enrichments of germline CNVs near CTGs and in regions of low-mappability are unlikely to be the result of a methodological artifact.

Various repeat families are more prone to harbor CNVs

We wanted to further characterize the distribution of germline CNVs in relation to different repeat classes and families. By comparing CNVs to the same control regions with matched overlap with SD and distance to CTGs we can look for patterns that are specific to repeat sub-families without the risk of being biased by the global enrichments (Fig. 4.4b). Using this approach, we found that CNVs were still significantly enriched in satellites repeats and in short tandem repeats (P-value $< 10^{-4}$, Fig. S4.8a), with fold-enrichments of 2.3 and 1.2 respectively.

Although it is known that DNA satellites and simple repeats are more unstable[220], 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. S4.8b). In particular, Alpha satellites have the highest and most significant enrichment (P-value $< 10^{-5}$), with more than 3 times more CNVs than in the control regions (Fig. 4.4b). We noted that satellites tend to span completely CNVs (Fig. S4.9), suggesting that satellites are likely directly involved in the CNV formation. Short and long tandem repeats can be highly polymorphic[211,212]. Constrained by read length, recent studies[221,222] focused on variation of STRs smaller than 100 bp. In our analysis we found that CNVs were significantly enriched in the largest annotated STRs (>100 bp or >400 bp, Fig. 4.4b). STR can be grouped by motif and we further tested the largest and most frequent families (Fig. S4.8c). Except for the weak enrichment in AT (TA) repeats, the STR enrichment appeared mostly independent of the repeat motif. Here the repeats tend to overlap just a fraction of the variant, but a clear subset of the

variants are fully covered by these tandem repeats (Fig. S4.9). Finally, although transposable elements (TEs) as a whole did not show enrichment (Fig. 4.4a), the "Other" repeat class, which contains SVA repeats, was found to be significantly enriched in the two higher depth datasets (Fig. 4.4b). Moreover, looking at TEs at the level of individual repeat families, we found a number of them to be significantly enriched including SVA F or L1Hs. Notably, HERV-H, an older ERV sub-family, was also in the list of enriched TEs. This sub-family has been shown to be expressed and important in human embryonic stem cells[223,224]. Alu elements contributed to the formation of human segmental duplications[225] and are often found around SV breakpoints[226] but this TE family was not enriched in CNVs in our data. On the other hand, several families of L1 repeats older than the still active L1HS family were also enriched (e.g. L1PA2 to L1PA4) and often implicated in what appears to be non-allelic homologous recombination (see examples in Fig. S4.10). Reassuringly, the somatic CNVs once again did not show any of these enrichments (Fig. 4.4b).

Impact of CNVs in regions of low-mappability

Compared to the latest 1000GP catalog[5], we identified 3,455 novel regions with CNVs in more than 1% of the population. 81.3% of these regions were located in low-mappability regions while 18.4% were located in extremely low-mappability regions. These novel CNV regions were missing from the 1000GP catalog and also mostly absent in other recent CNV surveys; only 7.9-15.1% of the novel regions overlapped with a CNV in three recent CNV catalogs[33,36,107] (Fig. S4.11). Among the regions with a CNV in the CEPH12878 sample, we identified a deletion in the second intron of the *TRIM16* gene that was found by both Pendleton et al.[59] and PopSV. Across the 640 individuals analyzed by PopSV, 12% carried the variant. Thanks to the long-read data, the exact breakpoints had been pinpointed in Pendleton et al.[59] and it was in fact a SVA-F transposable element located within the 6 Kbp intron in the reference genome but absent from the assembled sequence. SVA-F is one of the youngest repeat family in the human genome and their high similarity remains

a challenge for CNV analysis. Furthermore, the variant is located within a segmental duplication with 98.5% similarity and absent from public catalogs such as the 1000GP or GoNL. Another deletion supported by both public assemblies and local reassembly of the PacBio read was located 12 Kbp downstream of *TMPRSS11E*. 6.6% of the individuals carried the variant in the PopSV catalog. The assembled sequence helped pinpoint the breakpoints to an annotated L1PA2 in the reference genome. The variant was also located in a segmental duplication and absent from public catalogs such as the 1000GP or GoNL. Finally, a deletion affecting 8 different exons from the *CR1* gene was found by both Pendleton et al.[59] and PopSV in CEPH12878. *CR1* has been associated with Alzheimer disease[227] and is located within embedded segmental duplications with high similarity. The deletion was present in 3% of the population analyzed with PopSV but is absent from public CNV catalogs.

Overall, 7,206 protein-coding genes were found to have an exon overlapping a variant in at least one of the 640 normal genomes studied (Table 4.2). If we included the promoter regions (10 Kbp upstream of the transcription start site), at least 11,341 protein-coding genes were potentially affected by at least one CNV in the population. Focusing on regions of low-mappability, we found 4,285 different CNVs that were completely included in regions annotated as STR. These STR-CNVs overlapped the coding sequence of 45 protein-coding genes, and 286 genes when including the promoter region (Table 4.2). In contrast, for CNVs included in satellite regions, only 21 genes had an exon or the promoter region overlapping one of the 1,822 Satellite-CNVs. Finally, we focused on CNVs that were novel compared to the 1000GP[5] and in low-mappability regions. Even there, 347 genes were found to have an exon overlapping such CNVs and this number increased to 560 when including the promoter regions. Out of these 347 genes, 29 were previously associated to a mendelian disorder or phenotype in the OMIM database (Online Mendelian Inheritance in Man; http://omim.org/, Table S4.5).

Set	CNVs	Genes with CNVs			OMIM genes with CNVs		
		Exon	+ Promoter	+ Intron	Exon	+ Promoter	+ Intron
All CNVs							
All	91,735	7,206	11,341	13,259	1,241	1,857	2,196
Low coverage	32,707	848	1,491	2,648	95	160	371
Extremely low coverage	9,348	304	401	442	11	14	25
TE	20,491	164	1,747	3,998	29	233	664
STR	4,285	45	286	748	5	39	129
Satellite	1,822	2	21	33	0	0	0
Novel CNVs							
All	17,046	418	680	1,102	38	59	135
Low coverage	15,263	347	560	894	29	47	111
Extremely low coverage	6,591	189	263	285	5	6	8
TE	3,896	17	192	504	1	12	66
STR	1,806	14	81	230	0	9	41
Satellite	890	1	4	5	0	0	0

Table 4.2: **Impact of CNVs on protein-coding genes.** The *CNVs* number represents the number of different CNVs, after collapsing CNVs with more than 50% reciprocal overlap. Repeat CNV: more than 90% of the CNV is annotated as repeat. Genes are protein-coding genes and the promoter region is defined as the 10 Kbp region upstream of the transcription start site. *Novel CNVs* are located within regions annotated as novel compared to the 1000 Genome Project catalog.

4.5 Discussion

Despite the strong interest in CNVs because of their role in diseases, detecting them accurately has remained a challenge, especially in regions of low-mappability. This is mostly due to technical variation in RD that cannot be fully modeled by mappability estimates. Using a recently developed CNV-calling approach that relies on a set of reference samples to estimate the expected RD[2], we show that it is possible to accurately detect CNVs across the genome, even in repeat-rich regions. Indeed, using monozygotic twins and normal/tumor pairs, we were able to demonstrate that the performance of PopSV was stable and in most cases superior to other methods across different types of low-mappability regions. Although experimental validation can be challenging in these regions, we were able to confirm a number of deletions using PCR validation as well as variants in some of the most difficult regions by taking advantage of public datasets from long-read sequencing studies.

Notably, using PopSV on 140 normal genomes with high sequencing depth (∼40X) and 500 additional samples with medium coverage (∼13X), we found that regions of low mappability, which only represent ∼10% of the genome, were around 5 times more likely to harbor CNVs. The fact that this enrichment was observed for germline

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

Fine-tuning the location of breakpoints is often possible by reanalyzing the local read coverage or using orthogonal methods such as split-read or local assembly. In repeat-rich regions however, these methods generally do not perform well. Long read sequencing is currently the only experimental method that actually results in unambiguous SV calls with nearly quasi-base-pair resolution in low-mappability regions. Indeed, recent studies using long-read sequencing[58,59] found many novel SVs and highlighted variation involving complex repetitive DNA. The increased resolution and ability to span repeated regions expanded existing SV catalogs but only a handful of genomes have been sequenced in this way so far due to the higher cost of this technology. Although breakpoint and allele characterization is limited with short reads, we were able to detect the presence of such CNVs across a large population of normal genomes. Compared to previous studies, our CNV catalog strongly overlaps with the variants found by long-read sequencing studies in low-mappability regions. With hundreds of genomes at our disposal we identified frequent CNVs in repeat-rich regions that had escaped previous population-scale

surveys. In the CEPH12878 sample, we independently identified low-mappability variants and showed that some novel deletions were recurrent in our cohort. For example, an exonic deletion in the *CR1* gene absent from public CNV catalogs was identified by the long-read sequencing and found in ∼3% of the samples tested by PopSV. *CR1* has been associated with Alzheimer Disease[227] thus this exonic deletion in a low-mappability region might be relevant for association studies. Using our full CNV catalog, we identified 3,455 novel regions that were not present in 1000G public SV database[5] but found in more than 1% of our 640 genomes. These regions overlapped exons of 418 protein-coding genes, 38 of which were associated with a disease phenotype in the OMIM database. The amount of genes hit by CNVs in novel or low-mappability regions and the enrichment of CNVs in repeat-rich regions suggest that they be included in genome-wide surveys. As other types of variant are likely enriched in repeat-rich regions, we anticipate that population-based methods, such as PopSV, will facilitate the identification not only of CNVs but also of other types of SVs in both normal and cancer genomes.

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

4.6 Data and Code Availability

The PopSV R package and documentation are available at http://jmonlong.github.io/PopSV/. The scripts and instructions to reproduce the graphs and numbers in

this study have been deposited at http://github.com/jmonlong/reppopsv/ and archived in https://doi.org/10.5281/zenodo.1241137.

4.7 Accessions numbers

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

4.8 Acknowledgments

We are grateful to the team of the Québec Study of Newborn twins who provided the twin dataset and the Cagekid consortium who provided the renal cancer dataset. This study also made use of data generated by the Genome of the Netherlands Project. A full list of the investigators is available from www.nlgenome.nl. Funding for the project was provided by the Netherlands Organization for Scientific Research under award number 184021007, dated 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.

4.9 Funding

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.

Chapter 5

Discussion of Results and Implications

Population-based approaches and whole-genome sequencing

The three chapters highlighted the usefulness of population-based approach to detect genomic variation. The superior sensitivity made it possible to detect somatic variants in tumor samples, small and non-coding variants in epilepsy patients and variants in low-mappability regions in healthy individuals. The main results and implications concerning ccRCC, epilepsy and repeat-rich regions have been described in their respective chapters. However, all these results relied on a first step of variant detection that used the same strategy: analyzing a genome in the context of many to minimize the effect of technical noise. This population-based approach can benefit variant calling as long as the genomes in the population are sequenced with similar protocol and machine, and the raw data is pre-processed with the same pipeline. If this is the case, other factors that don't affect technical variation, for example gender or ethnicity, should not affect the variant calling. The main result of this work, at least in term of methodology, is the demonstrated power of population-based approach across multiple applications.

All three chapters used WGS data across dozens or hundreds of genomes. At the beginning of this work, WGS data across that many samples was not widespread. Large-scale sequencing projects used to be rare and international efforts involving numerous centers[57,228]. Yet, the cost of sequencing has been steadily decreasing and new machines currently advertise sub-1000$ costs for whole-genome sequencing. As a result, sequencing projects involving hundreds of genomes are more and more common. Motivated by the future opportunities for personalized medicine, hundreds of genomes are being sequenced to characterize the population in different countries[107,229,230]. For disease studies as well, whole genome sequencing of hundreds of participants has been performed or is currently under way. Large-scale project that focus on single disease include the MSSNG collaboration* that aims at sequencing 10,000 families affected by autism and whose first results were published recently[231] and the Alzheimer's Disease Sequencing Project[232] which is sequencing more than a thousand patients with Alzheimer disease. WGS is also at the core of multi-disease projects that are currently in progress. The Genomics England initiative† will sequence 100,000 genomes from patients with rare diseases and cancers while the TOPMed consortium‡ plans on sequencing the genome of 120,000 individuals to study the contribution of genetics to heart, lung, blood and sleep disorders. Often, this data is first processed at the individual level and later pooled to derive population information or association metrics. Our data highlights the benefit of pooling individuals earlier in the analysis workflow. Instead of pooling results after variant calling, we found that variant calling could be significantly improved thanks to population information. I believe that population-based approaches similar to PopSV could have a large impact on the variant discovery across these large-scale projects. It is regrettable to have information across hundreds of other experiments and to keep it aside for variant calling. Of note, promising avenues are currently being explored to reverse this trend, for example by integrating population information

*https://www.mss.ng
†https://www.genomicsengland.co.uk
‡https://www.nhlbiwgs.org/

even before variant calling. The new field of genome graphs aims at constructing a reference genome that includes known variation in order to directly improve read mapping, variant calling and population representation.

In addition to the population-based approach, our results also contribute to the case for whole-genome sequencing when designing a large scale genomic study. Although more costly than genotyping arrays or exome sequencing, WGS can detect a larger range of variants: rare and frequent, coding and non-coding, SNVs and CNVs or other SVs. WGS data produced with a primary objective in mind can often be re-used to study different aspects of genomic variation. For example, the data used in Chapter 2 was originally produced to describe somatic SNVs and broad CNV patterns in ccRCC[143]. In Chapter 2, we re-analyzed this dataset to further investigate the arm-level aberrations across the cohort and more precisely estimate the amount of somatic LOY in tumors from male patients. These results were replicated with a PCR-based approach in a different cohort of tumors and represent the foundation of the functional importance of somatic LOY advocated in our study. In chapter 3 and 4, the combination of WGS and PopSV was also instrumental in discovering novel scientific results. Novel coding variants in epilepsy or low-mappability regions were identified. A CNV profile that had never been associated with epilepsy, rare non-coding CNVs, was for the first time seen to be strongly enriched close to known epilepsy genes. Thousands of low-mappability regions that frequently experience CNV were identified, including some located near or within protein-coding genes.

Somatic loss of Y and gender imbalance in cancer

The gender imbalance in the renal cancer incidence is not completely understood. Our results suggest that somatic LOY occurs frequently in male tumors and have a functional impact. Hence, loss of chromosome Y could explain part of the higher incidence in males. Other cancers, such as liver and bladder cancers, show gender imbalance toward males and might also be partly explained by LOY[131]. Several

studies have found that more than 30% of tumors from these cancers experienced LOY[138,139]. Our study and others suggest that LOY could be a driver mechanism of tumor progression by disregulating tumor suppressors, such as *KDM5D* and *KDM6C*. Of note, *KDM5D* and *KDM6C* are both expressed in the liver and the bladder. Because of the multiple similarities with our study of ccRCC, it is sensible to speculate that the same mechanism, that is downregulation of *KDM5D* and *KDM6C* through LOY, is involved in the tumor progression of cancers of the liver and bladder. A population-based approach such as the one we used might identify more subtle LOY in sequencing datasets which could help estimate the rate and impact of this type of variation.

Although incidental, we detected LOY in the blood samples as well, both in the WGS and the PCR-based replication. As described in the literature, somatic LOY in blood was associated with older age in our cohort. Two studies suggest that the presence of LOY in blood samples can be associated with higher incidence rate of non-hematological cancers or Alzheimer disease[168,233]. Unfortunately, due to the absence of matched healthy controls we couldn't directly test the association with ccRCC. We found no significant associations between LOY in blood and the tumor stage or grade. Further investigations in the future will require more samples, including matched controls and balanced numbers of different tumor grades.

Small and non-coding CNVs in epilepsy and neurological disorders

Our results suggest that WGS will be necessary to fully characterize the genetic factors associated with epilepsy. First, WGS is more suitable for CNV detection compared to other sequencing approaches (e.g. exome sequencing). Despite its successes in detecting rare SNVs associated with genetic disorders, exome sequencing is less suited for CNV and SV detection. Indeed, the step that captures the regions of interest adds another layer of technical bias. Not only is the read coverage affected

by the capture efficiency in each region, the fragmented representation of the genome also hinders the sensitive detection of CNVs. Considering the importance of CNVs in epilepsy as shown by our study and others, WGS will be key to efficiently detect even exonic variants associated with the disease. Second, we show for the first time an association between non-coding CNVs and epilepsy. Thanks to WGS, we were able to interrogate the presence of both small and non-coding variants. In contrast, previous array-based study were limited to large CNVs which tended to overlap exonic regions. We found a clear enrichment of rare non-coding CNVs in patients close to genes that were previously associated with epilepsy. Even more convincing, the enrichment increased the closer to the exon and was boosted for deletions and variants overlapping regulatory regions. Similar results were found in individuals with autism where non-coding *de novo* CNVs and SNVs were enriched up to 100 Kbp from autism-associated genes[234].

These conclusions are relevant for epilepsy but also for other neurological disorders. Large CNVs have also been associated with autism[117], mental retardation[112] and schizophrenia[115,235]. In each disease, the same pattern emerged: probands tend to have rare large CNVs and/or in hotspot regions flanked by segmental duplications. Often, the same CNV hotspot regions have been associated with several neurological disorders as is the case for 1q21.1, 15q13.3 and 16p11.2[112]. Based on these commonalities and our results, I expect that small and non-coding CNVs play a role in other neurological disorders as well. In general, the study of the genetics of neurological disorder could greatly benefit from WGS approaches similar to the one we used in chapter 3.

The enhanced resolution of WGS and our population-based approach suggest that some types of variants had been missed before. The inclusion of such genomic variants might help characterize the exact syndrome or grouping patients with similar etiology. Drug resistance is an important problem in idiopathic epilepsy and much remains to be done to understand its causes. The study presented in chapter 3 contributes to the identification of genes that might be associated with epilepsy

and advocates for the inclusion of small and non-coding CNVs in the genomic profile of epilepsy patients. It is my hope that either these candidate genes or the inclusion of those variants will one day help to devise new drugs or to assist clinician when choosing a treatment.

Including low-mappability regions in genome-wide studies

More and more evidence supports the importance of repeated regions in disease and human phenotypes. After being considered junk DNA for a long period, the role of many types of repeats is becoming clear. From the highly variable STRs and their association with gene expression changes[79], to the contribution of TEs in regulation networks[12] or the satellite instability in cancer[95], repeats are gathering more attention. Still, repeats are under-studied because of the technical challenges and reluctant mentality shift. Our results show that WGS and a population-based approach is sufficient to detect CNV in many repeat-rich regions. Moreover, we found that these regions are more likely to harbor CNVs compared to the rest of the genome. Many low-mappability regions are also located close to protein-coding genes, some of which have been linked to Mendelian diseases before. By masking repeats, genome-wide association studies have discarded a small part of the genome but a large fraction of the genomic variation. I believe that approaches similar to the one presented in this work, as well as future technologies that will make repeat integration easier, could lead to a better genomic characterization of diseases and human phenotypes.

Chapter 6

Conclusions and Future Directions

The population-based approaches presented here have been successful at detecting somatic and challenging germline changes in copy number but more remains to be done. Although we were able to detect the presence of variation, more effort will be necessary to fully characterize the alternate alleles. The methods could also be extended to other sequencing technologies, particularly targeted sequencing such as whole-exome sequencing. Similar approaches could also be extended to other types of SVs, for example translocations and inversions. Finally, recent technological advances shine a promising light on the future of SV characterization. Genome editing might help investigate the functional impact of non-coding CNVs while new sequencing technologies will likely be used in concert to integrate SVs and low-mappability in genomics studies.

From variant detection to base-pair resolution

The population-based approach presented in this work is powerful to detect variation but cannot fully characterize the variants. Deriving the exact sequence of the mutated allele or breakpoints remains challenging in repeat-rich regions or for complex SVs. Indeed, large sample sizes and population-oriented analysis are mostly able to flag the presence of an abnormal signal in these regions. To unlock a base-pair level resolution for the breakpoints or the variant sequence in general, a different type

of data will be necessary rather than more of the same data. In WGS early years, genomes were sequenced at low depth and pooling individuals helped increase read support necessary to estimate the variant sequences. A typical WGS genome is now often 30x deep or more and pooling reads across different individuals will have only marginal benefit on the base-pair resolution of the SVs. At this depth, the problematic variants are either in repeat-rich regions or more complex than the canonical SV types. More of the same short reads wouldn't bring much new information and the ambiguity would likely remain. In the end, the short read size is the main limitation. Hence, long read sequencing will likely have a large impact on characterizing many of the SVs that remain challenging. Several studies already showed the power of long reads for SV detection[58,59]. Unfortunately the higher cost of these technologies limits their use. The cost and usage might change but it is unlikely that a very large number of genome will be sequenced with these technologies. Still, this technology could be used to validate and confirm SVs identified in large scale short-read projects. By carefully choosing a few genomes to sequence, the catalog of complex and repeat-rich SVs could be greatly improved. For example, sequencing genomes carrying candidate pathogenic variants could validate and provide more insights into the variant impact or potential mechanism of action. Selecting a few genomes with different complex SVs and low-mappability variants could also maximize the gain from each long-read sequencing experiment. In summary, long-read sequencing of a few carefully selected genomes would nicely complement deep short-read sequencing across large cohorts. Of note, a more cost effective approach would be to capture the regions containing the SVs detected from the large-scale short-read surveys. The challenge here is to efficiently and accurately capture large and potentially repeat-rich regions. If this type of capture is possible, Sanger sequencing could also be used to characterize the variants at the base-pair level. Having a validated set of variants at the base-pair level and covering low-mappability regions will also provide an extremely useful gold-standard to assess the sensitivity and specificity of CNV methods. In the absence of such a gold-standard, we had to rely on indirect assess-

ment, such as the replication in twins or the comparison with just a few long-read assemblies.

Extension to targeted sequencing

Targeted sequencing involve a step of capture in which the desired genomic regions are selectively amplified. Thanks to the capture, only a set of regions, for example coding regions, are sequenced. However, the capture efficiency varies depending on the design and regions, introducing additional technical variation in the read coverage. CNV detection from the read coverage is challenging as a result. Existing methods turned to population-based approach, similar to our WGS method PopSV, in order to normalize read coverage using other experiments that used similar capture[236,237]. PopSV could be easily extended to such application as it doesn't rely on coverage uniformity but rather considers each region separately in the context of reference samples. We actually ran PopSV successfully on several whole-exome sequencing projects with minimal changes. Briefly, one additional step in the pipeline was necessary: the removal of regions that were not covered by the sequencing experiment. Of note, the quality control step was even more important here in order to ensure that the reference and tested samples originated from the same sequencing protocol. Indeed, capture protocols are continuously evolving with new exome capture kits available every year. Although we believe that the normalization used by PopSV is superior to several of existing methods, the regions with abnormal coverage are merged using a simplistic approach. For WGS, it is natural to merge consecutive regions that share a CNV signal as they are located one next to the other. In targeted sequencing, there are often stretches of non-covered regions that separate captured regions. By integrating this information a better segmentation can be performed as shown by several methods[238,239]. In the context of an extension to targeted sequencing, we could improve PopSV's segmentation step with such algorithms in order to have both a more advanced normalization and an appropriate segmentation.

Extension to balanced structural variation

PopSV could also be extended to other types of SVs, as the name originally intended. Currently, abnormal coverage of properly mapped reads identifies regions with a change in copy number. In the future, PopSV could operate on the coverage of reads with discordant paired-end mapping. For example, reads whose pair couldn't be aligned or aligned at an abnormal distance or orientation. An abnormal excess of discordant reads would be a sign of SV. With PopSV's population-based approach, discordant mapping due to repeats will be corrected for and only regions with a real excess of discordant mapping should be called. This type of test could be run across genomic regions, as for the CNV analysis, or between pairs genomic regions. The objective of the paired-regions approach is to increase the signal-to-noise ratio between the SV-caused discordant reads and the background level. Indeed, the background level of discordant reads in a region might be high enough to drown the signal from the additional reads created by the SV. When focusing on discordant read pairs with one read mapped to a region and the other in another distant region, the amount of background discordant mapping is much lower. The excess of discordant read pair linking the boundaries of a translocation or an inversion might be more easily detected. Both of these approaches are currently being tested. The original scan, i.e. testing one region at a time, seems to be under-powered and return much fewer calls than the CNV scan. Alternatively, this extension of PopSV would also be useful to annotate calls from other methods, for example those that identify clusters of discordant or split reads. These methods tend to suffer from a high false-positive rate most likely due to repeat confusion.

Investigating the functional impact of non-coding CNVs

We identified dozens of rare non-coding CNVs located close to known epilepsy genes and absent from our controls and public CNV databases. The most promising

candidates are located in regions that were previously associated with changes in the gene expression and are predicted to host regulatory regions. While the frequency and location of these non-coding CNVs are encouraging, it is not clear which variant really has an impact on the gene and eventually the disease phenotype. Indeed, we observed a clear enrichment which means that some CNVs are associated with the disease. In order to narrow down the list of candidate regions we could either increase the sample size or experimentally test the variant's impact. Both strategies have their own set of challenges.

To achieve a higher sample size, more patients are necessary and might require national or international collaboration. Because epilepsy is a diverse disease, the new patients should be matched as much as possible with our cohort in order to maximize the chances of observing recurrent variants. Finally, the number of probands that need to be sequenced to identify single genes or non-coding regions might be unrealistically high, again because of the diverse phenotype and complex gene network involved in the disease mechanism.

Investigating deeper our current candidates might be more feasible thanks to recent advances in cell reprogramming and DNA editing. It is not feasible to study live human brains and unpractical to collect post-mortem brains in correct conditions for functional assays. Cell reprogramming provides a better alternative. By reprogramming cells from a carrier into relevant cell types, e.g. neurons, the effect of a variant on the gene function could be tested in the laboratory. If patients' samples are unavailable or to better control for the genetic background, DNA editing like CRISPR-Cas9 could recreate the non-coding deletion in cells. DNA editing before cell differentiation would provide a culture of cells that could then be differentiated into different cell types to extensively study the effect of the variants. While a true "epilepsy" phenotype is almost impossible to test in a cell culture, we could investigate the effect of the variants on gene expression or other molecular phenotypes. Although this approach requires time, resources and expertise, it is more geared toward validating our non-coding candidates and more likely to give

conclusive results.

SV and WGS in the near future

Thanks to better sequencing technologies and methods, systematic *de novo* assembly of genomes might eventually supplant the re-sequencing approach that is dominant today. The optimal output would be phased sequence for each genome that is sequenced and assembled. SV detection would then come down to comparing the assembled genomes. Recent advances in long-read, linked-read, or conformation capture sequencing greatly improve the quality of the *de novo* assemblies and contains long-range information useful for phasing. However, for SV detection, long-read sequencing will be necessary to reach a base-pair resolution across the full genome, as mentioned above. Because of its cost, only a limited number of genomes will be sequenced with long-read sequencing in the near future. Furthermore, hundreds of thousands of genomes have been sequenced or are being sequenced with short-read technology. Until affordable and comprehensive *de novo* assembly is available, the field will likely move to hybrid strategies for SV analysis. But how could we integrate SV information from different approaches and across hundreds to thousands of genomes? The recent development of genome graphs provides a promising solution[240]. Genome graphs represent the genome and genomic variation in a population using a graph structure. One of its current form corresponds to the current reference genome augmented with SNVs and indels from the 1000 Genomes Project[241]. Genome graphs are by nature flexible so that we could imagine further improving their breadth with high-quality *de novo* assemblies or SV catalogs from long-read sequencing datasets. Using a genome graph populated with the high-resolution variants, SV could be more efficiently genotyped in short-reads datasets across large populations. Genome graphs also provide an ideal structure to represent complex haplotypes, such as those involving SVs. As for other types of variants, haplotype information will be invaluable to assist variant calling and to predict the functional impact of SVs. Short-read datasets could benefit from the long-range information

present in the high-resolution datasets by integrating complex haplotype information in the genome graphs. With coordination and data sharing, and with genome graphs as a new reference system, there is hope that we can rapidly reach the point where phased high-resolution SVs can be genotyped accurately from short-read datasets spanning hundreds of thousands of genomes.

Bibliography

[1] M. Arseneault, J. Monlong, N. S. Vasudev, R. S. Laskar, M. Safisamghabadi *et al.* Loss of chromosome Y leads to down regulation of KDM5D and KDM6C epigenetic modifiers in clear cell renal cell carcinoma. *Scientific Reports*, 7: 44876, mar 2017. doi:10.1038/srep44876.

[2] J. Monlong, S. L. Girard, C. Meloche, M. Cadieux-Dion, D. M. Andrade *et al.* Global characterization of copy number variants in epilepsy patients from whole genome sequencing. *PLoS Genetics*, 14(4):e1007285, 2018. doi:10.1371/journal.pgen.1007285.

[3] J. Monlong, P. Cossette, C. Meloche, G. Rouleau, S. L. Girard *et al.* Human copy number variants are enriched in regions of low mappability. *Nucleic Acids Research*, page gky538, 2018. doi:10.1093/nar/gky538.

[4] I. M. Hall and A. R. Quinlan. Detection and Interpretation of Genomic Structural Variation in Mammals. In *Methods in molecular biology (Clifton, N.J.)*, volume 838, pages 225–248. Springer Science, jan 2012. doi:10.1007/978-1-61779-507-7_11.

[5] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, oct 2015. doi:10.1038/nature15394.

[6] R. L. Collins, H. Brand, C. E. Redin, C. Hanscom, C. Antolik *et al.* Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome biology*, 18(1):36, mar 2017. doi:10.1186/s13059-017-1158-6.

[7] A. R. Quinlan and I. M. Hall. Characterizing complex structural variation in germline and somatic genomes. *Trends in genetics : TIG*, 28(1):43–53, jan 2012. doi:10.1016/j.tig.2011.10.002.

[8] H. Brand, R. L. Collins, C. Hanscom, J. A. Rosenfeld, V. Pillalamarri *et al.* Paired-Duplication Signatures Mark Cryptic Inversions and Other Complex Structural Variation. *American journal of human genetics*, 97(1):170–6, jul 2015. doi:10.1016/j.ajhg.2015.05.012.

[9] A. Malhotra, M. Lindberg, G. G. Faust, M. L. Leibowitz, R. a. Clark *et al.* Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome research*, 23(5):762–76, may 2013. doi:10.1101/gr.143677.112.

[10] A. J. Sharp, Z. Cheng, and E. E. Eichler. Structural Variation of the Human Genome. *Annual Review of Genomics and Human Genetics*, 7(1):407–442, jan 2006. doi:10.1146/annurev.genom.7.080505.115618.

[11] R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature*, 470 (7332):59–65, feb 2011. doi:10.1038/nature09708.

[12] G. Bourque. Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current opinion in genetics & development*, 19(6): 607–12, dec 2009. doi:10.1016/j.gde.2009.10.013.

[13] P. M. Kim, H. Y. K. Lam, a. E. Urban, J. O. Korbel, J. Affourtit *et al.* Analysis of copy number variants and segmental duplication in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Research*, 18:1865–1874, 2008. doi:10.1101/gr.081422.108.

[14] W. Gu, F. Zhang, and J. R. Lupski. Mechanisms for human genomic rearrangements. *PathoGenetics*, 1(1):4, nov 2008. doi:10.1186/1755-8417-1-4.

[15] C. Martinez-A and K. H. M. van Wely. Centromere fission, not telomere erosion, triggers chromosomal instability in human carcinomas. *Carcinogenesis*, 32(6):796–803, jun 2011. doi:10.1093/carcin/bgr069.

[16] E. G. Bochukova, N. Huang, J. Keogh, E. Henning, C. Purmann *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, 463(7281):666–670, 2010. doi:10.1038/nature08689.

[17] J. L. Stone, M. C. O'Donovan, H. Gurling, G. K. Kirov, D. H. R. Blackwood *et al.* Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 455(7210):237–241, sep 2008. doi:10.1038/nature07239.

[18] H. Stefansson, A. Meyer-Lindenberg, S. Steinberg, B. Magnusdottir, K. Morgen *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature*, 505(7483):361–6, 2014. doi:10.1038/nature12818.

[19] H. C. Mefford, S. C. Yendle, C. Hsu, J. Cook, E. Geraghty *et al.* Rare copy number variants are an important cause of epileptic encephalopathies. *Annals of Neurology*, 70(6):974–985, dec 2011. doi:10.1002/ana.22645.

[20] S. a. McCarroll, A. Huett, P. Kuballa, S. D. Chilewski, A. Landry *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nature genetics*, 40(9):1107–1112, 2008. doi:10.1038/ng.215.

[21] R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905, mar 2010. doi:10.1038/nature08822.

[22] 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. doi:10.1038/nature08979.

[23] 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. doi:10.5808/GI.2014.12.3.98.

[24] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen *et al.* Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, apr 2010. doi:10.1038/nature08516.

[25] H. V. Firth, S. M. Richards, a. P. Bevan, S. Clayton, M. Corpas *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, 84(4):524–533, 2009. doi:10.1016/j.ajhg.2009.03.010.

[26] M. Spielmann and E. Klopocki. CNVs of noncoding cis-regulatory elements in human disease. *Current Opinion in Genetics & Development*, 23(3):249–256, jun 2013. doi:10.1016/j.gde.2013.02.013.

[27] J. R. Lupski, R. M. de Oca-Luna, S. Slaugenhaupt, L. Pentao, V. Guzzetta *et al.* DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*, 66(2):219–232, 1991. doi:10.1016/0092-8674(91)90613-4.

[28] S. Liang, X. Wei, C. Gong, J. Wei, Z. Chen *et al.* Significant association between asthma risk and the GSTM1 and GSTT1 deletion polymorphisms: an updated meta-analysis of case-control studies. *Respirology (Carlton, Vic.)*, 18(5):774–83, jul 2013. doi:10.1111/resp.12097.

[29] K. Fellermann, D. E. Stange, E. Schaeffeler, H. Schmalzl, J. Wehkamp *et al.* A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *American journal of human genetics*, 79(3):439–48, sep 2006. doi:10.1086/505915.

[30] E. J. Hollox, U. Huffmeier, P. L. J. M. Zeeuwen, R. Palla, J. Lascorz *et al.* Psoriasis is associated with increased beta-defensin genomic copy number. *Nature genetics*, 40(1):23–5, jan 2008. doi:10.1038/ng.2007.48.

[31] E. Gonzalez, H. Kulkarni, H. Bolivar, A. Mangano, R. Sanchez *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science (New York, N.Y.)*, 307(5714):1434–40, mar 2005. doi:10.1126/science.1101160.

[32] C. McKinney, M. E. Merriman, P. T. Chapman, P. J. Gow, A. A. Harrison *et al.* Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. *Annals of the rheumatic diseases*, 67(3):409–13, mar 2008. doi:10.1136/ard.2007.075028.

[33] R. E. Handsaker, V. Van Doren, J. R. Berman, G. Genovese, S. Kashin *et al.* Large multiallelic copy number variations in humans. *Nature Genetics*, 47(3):296–303, jan 2015. doi:10.1038/ng.3200.

[34] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science (New York, N.Y.)*, 315(5813):848–53, feb 2007. doi:10.1126/science.1136678.

[35] K. M. Lower, J. R. Hughes, M. De Gobbi, S. Henderson, V. Viprakasit *et al.* Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21771–6, dec 2009. doi:10.1073/pnas.0909331106.

[36] C. Chiang, A. J. Scott, J. R. Davis, E. K. Tsang, X. Li *et al.* The impact of structural variation on human gene expression. *Nature genetics*, 49(5):692–699, may 2017. doi:10.1038/ng.3834.

[37] D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, may 2015. doi:10.1016/j.cell.2015.04.004.

[38] J. Weischenfeldt, T. Dubash, A. P. Drainas, B. R. Mardin, Y. Chen *et al.* Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nature Genetics*, 49(1):65–74, nov 2016. doi:10.1038/ng.3722.

[39] C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. *Nature reviews. Genetics*, 12(5):363–76, may 2011. doi:10.1038/nrg2958.

[40] V. Boeva, A. Zinovyev, K. Bleakley, J. P. Vert, I. Janoueix-Lerosey *et al.* Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, 27(2):268–269, jan 2011. doi:10.1093/bioinformatics/btq635.

[41] A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21(6):974–984, jun 2011. doi:10.1101/gr.114876.110.

[42] G. Klambauer, K. Schwarzbauer, A. Mayr, D. A. Clevert, A. Mitterecker *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–e69, may 2012. doi:10.1093/nar/gks003.

[43] K. Chen, J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods*, 6(9):677–81, sep 2009. doi:10.1038/nmeth.1363.

[44] M. R. Lindberg, I. M. Hall, and A. R. Quinlan. Population-based structural variation discovery with Hydra-Multi. *Bioinformatics (Oxford, England)*, pages 4–6, dec 2014. doi:10.1093/bioinformatics/btu771.

[45] 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, nov 2009. doi:10.1093/bioinformatics/btp394.

[46] A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics*, 46(8):912–918, jul 2014. doi:10.1038/ng.3036.

[47] R. E. Handsaker, J. M. Korn, J. Nemesh, and S. A. McCarroll. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature genetics*, 43(3):269–76, mar 2011. doi:10.1038/ng.768.

[48] T. J. Treangen and S. L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46, nov 2011. doi:10.1038/nrg3117.

[49] 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*, 28(21):2711–2718, nov 2012. doi:10.1093/bioinformatics/bts535.

[50] Y. Benjamini and T. P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):e72–e72, may 2012. doi:10.1093/nar/gks001.

[51] A. Koren, R. E. Handsaker, N. Kamitaki, R. Karlić, S. Ghosh *et al.* Genetic variation in human DNA replication timing. *Cell*, 159(5):1015–1026, nov 2014. doi:10.1016/j.cell.2014.10.025.

[52] 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–e103, aug 2011. doi:10.1093/nar/gkr425.

[53] D. Aird, M. G. Ross, W.-S. Chen, M. Danielsson, T. Fennell *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biology*, 12(2):R18, 2011. doi:10.1186/gb-2011-12-2-r18.

[54] I. Kozarewa, Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature methods*, 6(4):291–5, apr 2009. doi:10.1038/nmeth.1311.

[55] A. Koren, P. Polak, J. Nemesh, J. J. Michaelson, J. Sebat *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *American Journal of Human Genetics*, 91(6):1033–1040, 2012. doi:10.1016/j.ajhg.2012.10.018.

[56] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7, dec 1977.

[57] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, feb 2001. doi:10.1038/35057062.

[58] M. J. P. Chaisson, J. Huddleston, M. Y. Dennis, P. H. Sudmant, M. Malig *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608–11, jan 2015. doi:10.1038/nature13907.

[59] M. Pendleton, R. Sebra, A. W. C. Pang, A. Ummat, O. Franzen *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods*, 12(8):780–6, aug 2015. doi:10.1038/nmeth.3454.

[60] J. Huddleston, M. J. P. Chaisson, K. M. Steinberg, W. Warren, K. Hoekzema *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research*, 27(5):677–685, may 2017. doi:10.1101/gr.214007.116.

[61] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, (December 2017), jan 2018. doi:10.1038/nbt.4060.

[62] P. Van Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye *et al.* Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(39):16910–5, sep 2010. doi:10.1073/pnas.1009843107.

[63] M. Zhu, A. C. Need, Y. Han, D. Ge, J. M. Maia *et al.* Using ERDS to infer copy-number variants in high-coverage genomes. *American Journal of Human Genetics*, 91(3):408–421, 2012. doi:10.1016/j.ajhg.2012.07.004.

[64] V. Boeva, T. Popova, K. Bleakley, P. Chiche, J. Cappo *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(3):423–5, feb 2012. doi:10.1093/bioinformatics/btr670.

[65] F. Favero, T. Joshi, A. M. Marquard, N. J. Birkbak, M. Krzystanek *et al.* Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology*, 26(1):64–70, 2015. doi:10.1093/annonc/mdu479.

[66] G. G. Faust and I. M. Hall. YAHA: Fast and flexible long-read alignment with optimal breakpoint detection. *Bioinformatics*, 28(19):2417–2424, oct 2012. doi:10.1093/bioinformatics/bts456.

[67] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):R84, jun 2014. doi:10.1186/gb-2014-15-6-r84.

[68] T. Rausch, D. T. W. Jones, M. Zapatka, A. M. Stütz, T. Zichner *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell*, 148(1-2):59–71, jan 2012. doi:10.1016/j.cell.2011.12.013.

[69] K. Wong, T. M. Keane, J. Stalker, and D. J. Adams. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome biology*, 11(12):R128, jan 2010. doi:10.1186/gb-2010-11-12-r128.

[70] K. Chen, L. Chen, X. Fan, J. Wallis, L. Ding *et al.* TIGRA: A targeted iterative graph routing assembler for breakpoint assembly. *Genome research*, pages 310–317, jan 2014. doi:10.1101/gr.162883.113.

[71] J. Zhuang and Z. Weng. Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes. *Nucleic Acids Research*, 43(17):8146–8156, 2015. doi:10.1093/nar/gkv831.

[72] Z. Chong, J. Ruan, M. Gao, W. Zhou, T. Chen *et al.* novoBreak: local assembly for breakpoint detection in cancer genomes. *Nature methods*, 14(1): 65–67, jan 2017. doi:10.1038/nmeth.4084.

[73] M. Nattestad and M. C. Schatz. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*, 32(19):3021–3023, oct 2016. doi:10.1093/bioinformatics/btw369.

[74] P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics*, 40(6):722–729, jun 2008. doi:10.1038/ng.128.

[75] G. Ha, A. Roth, J. Khattra, J. Ho, D. Yap *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome research*, pages 1881–1893, jul 2014. doi:10.1101/gr.180281.114.

[76] A. Smit, R. Hubley, and P. Green. RepeatMasker Open-4.0, 2015. URL `http://www.repeatmasker.org`.

[77] J. A. Bailey and E. E. Eichler. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews Genetics*, 7(7):552–564, 2006. doi:10.1038/nrg1895.

[78] R. E. Mills, E. A. Bennett, R. C. Iskow, and S. E. Devine. Which transposable elements are active in the human genome? *Trends in Genetics*, 23(4):183–191, 2007. doi:10.1016/j.tig.2007.02.006.

[79] M. Gymrek, T. Willems, A. Guilmatre, H. Zeng, B. Markus *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature Genetics*, 48(1):22–29, jan 2016. doi:10.1038/ng.3461.

[80] J. Quilez, A. Guilmatre, P. Garg, G. Highnam, M. Gymrek *et al.* Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Research*, 44(8):gkw219, 2016. doi:10.1093/nar/gkw219.

[81] A. J. Hannan. Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews Genetics*, 2018. doi:10.1038/nrg.2017.115.

[82] F. Hormozdiari, I. Hajirasouliha, P. Dao, F. Hach, D. Yorukoglu *et al.* Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, 26(12):i350–i357, 2010.

[83] D. He, F. Hormozdiari, N. Furlotte, and E. Eskin. Efficient algorithms for tandem copy number variation reconstruction in repeat-rich regions. *Bioinformatics (Oxford, England)*, 27(11):1513–20, jun 2011. doi:10.1093/bioinformatics/btr169.

[84] C. Alkan, J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics*, 41(10):1061–1067, 2009. doi:10.1038/ng.437.

[85] P. H. Sudmant, J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig *et al.* Diversity of Human Copy Number Variation and Multicopy Genes. *Science*, 330 (6004):641–646, oct 2010. doi:10.1126/science.1197005.

[86] S. Ivakhno, T. Royce, A. J. Cox, D. J. Evers, R. K. Cheetham *et al.* CNAseg–a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics (Oxford, England)*, 26(24): 3051–8, dec 2010. doi:10.1093/bioinformatics/btq587.

[87] M. E. MacDonald, C. M. Ambrose, M. P. Duyao, R. H. Myers, C. Lin *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. doi:10.1016/0092-8674(93)90585-E.

[88] S. E. Andrew, Y. P. Goldberg, B. Kremer, H. Telenius, J. Theilmann *et al.* The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nature genetics*, 4(4):398–403, aug 1993. doi:10.1038/ng0893-398.

[89] A. J. Verkerk, M. Pieretti, J. S. Sutcliffe, Y. H. Fu, D. P. Kuhl *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, 65(5): 905–14, may 1991.

[90] J. Rich, V. V. Ogryzko, and I. V. Pirozhkova. Satellite DNA and related diseases. *Biopolymers and Cell*, 30(4):249–259, jul 2014. doi:10.7124/bc.00089E.

[91] E. Lee, R. Iskow, L. Yang, O. Gokcumen, P. Haseley *et al.* Landscape of somatic retrotransposition in human cancers. *Science (New York, N.Y.)*, 337 (6097):967–71, aug 2012. doi:10.1126/science.1222077.

[92] J. M. C. Tubio, Y. Li, Y. S. Ju, I. Martincorena, S. L. Cooke *et al.* Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science*, 345(6196):1251343–1251343, jul 2014. doi:10.1126/science.1251343.

[93] K. H. Burns. Transposable elements in cancer. *Nature Reviews Cancer*, 17(7): 415–424, 2017. doi:10.1038/nrc.2017.35.

[94] Y. Miki, I. Nishisho, A. Horii, Y. Miyoshi, J. Utsunomiya *et al.* Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer research*, 52(3):643–5, feb 1992.

[95] T.-M. Kim, P. W. Laird, and P. J. Park. The Landscape of Microsatellite Instability in Colorectal and Endometrial Cancer Genomes. *Cell*, 155(4):858–868, nov 2013. doi:10.1016/j.cell.2013.10.015.

[96] A. de la Chapelle and H. Hampel. Clinical relevance of microsatellite instability in colorectal cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(20):3380–7, jul 2010. doi:10.1200/JCO.2009.27.0652.

[97] A. Fungtammasan, E. Walsh, F. Chiaromonte, K. A. Eckert, and K. D. Makova. A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? *Genome research*, 22 (6):993–1005, jun 2012. doi:10.1101/gr.134395.111.

[98] S. G. Durkin and T. W. Glover. Chromosome fragile sites. *Annual review of genetics*, 41(1):169–92, dec 2007. doi:10.1146/annurev.genet.41.042007.165900.

[99] M. A. Lehrman, W. J. Schneider, T. C. Südhof, M. S. Brown, J. L. Goldstein *et al.* Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science (New York, N.Y.)*, 227 (4683):140–6, jan 1985.

[100] C. Sun, H. Skaletsky, S. Rozen, J. Gromoll, E. Nieschlag *et al.* Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. *Human molecular genetics*, 9 (15):2291–6, sep 2000.

[101] V. P. Belancio, P. L. Deininger, and A. M. Roy-Engel. LINE dancing in the human genome: transposable elements and disease. *Genome medicine*, 1(10): 97, oct 2009. doi:10.1186/gm97.

[102] L. Li, S. McVety, R. Younan, P. Liang, D. Du Sart *et al.* Distinct patterns of germ-line deletions in MLH1 and MSH2: the implication of Alu repetitive element in the genetic etiology of Lynch syndrome (HNPCC). *Human mutation*, 27(4):388, apr 2006. doi:10.1002/humu.9417.

[103] M. Miné, J.-M. Chen, M. Brivet, I. Desguerre, D. Marchant *et al.* A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element. *Human mutation*, 28(2):137–42, feb 2007. doi:10.1002/humu.20449.

[104] A. W. Pang, J. R. MacDonald, D. Pinto, J. Wei, M. a. Rafiq *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome biology*, 11(5):R52, jan 2010. doi:10.1186/gb-2010-11-5-r52.

[105] M. Zarrei, J. R. Macdonald, D. Merico, and S. W. Scherer. A copy number variation map of the human genome. *Nature Publishing Group*, 16(3):172–183, 2015. doi:10.1038/nrg3871.

[106] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry *et al.* Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, nov 2006. doi:10.1038/nature05329.

[107] L. C. Francioli, A. Menelaou, S. L. Pulit, F. van Dijk, P. F. Palamara *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 46(8):818–825, jun 2014. doi:10.1038/ng.3021.

[108] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe *et al.* Detection of large-scale variation in the human genome. *Nature genetics*, 36 (9):949–51, sep 2004. doi:10.1038/ng1416.

[109] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young *et al.* Large-scale copy number polymorphism in the human genome. *Science (New York, N.Y.)*, 305 (5683):525–8, jul 2004. doi:10.1126/science.1098918.

[110] K. K. Wong, R. J. DeLeeuw, N. S. Dosanjh, L. R. Kimm, Z. Cheng *et al.* A Comprehensive Analysis of Common Copy-Number Variations in the Human Genome. *The American Journal of Human Genetics*, 80(1):91–104, 2007. doi:10.1086/510560.

[111] A. Itsara, G. M. Cooper, C. Baker, S. Girirajan, J. Li *et al.* Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease. *The American Journal of Human Genetics*, 84(2):148–161, feb 2009. doi:10.1016/j.ajhg.2008.12.014.

[112] H. C. Mefford and E. E. Eichler. Duplication hotspots, rare genomic disorders, and common disease. *Current opinion in genetics & development*, 19(3):196–204, jun 2009. doi:10.1016/j.gde.2009.04.003.

[113] X. She, G. Liu, M. Ventura, S. Zhao, D. Misceo *et al.* A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome research*, 16(5):576–83, may 2006. doi:10.1101/gr.4949406.

[114] D. A. Koolen, L. E. L. M. Vissers, R. Pfundt, N. de Leeuw, S. J. L. Knight *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nature genetics*, 38(9):999–1001, sep 2006. doi:10.1038/ng1853.

[115] T. Walsh, J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science (New York, N.Y.)*, 320(5875):539–43, apr 2008. doi:10.1126/science.1155174.

[116] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin *et al.* Strong association of de novo copy number mutations with autism. *Science (New York, N.Y.)*, 316(5823):445–9, apr 2007. doi:10.1126/science.1138659.

[117] D. Pinto, A. T. Pagnamenta, L. Klei, R. Anney, D. Merico *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304):368–72, jul 2010. doi:10.1038/nature09146.

[118] X. Ran, J. Li, Q. Shao, H. Chen, Z. Lin *et al.* EpilepsyGene: A genetic resource for genes and mutations related to epilepsy. *Nucleic Acids Research*, 43(D1):D893–D899, jan 2015. doi:10.1093/nar/gku943.

[119] H. C. Mefford, H. Muhle, P. Ostertag, S. von Spiczak, K. Buysse *et al.* Genome-Wide Copy Number Variation in Epilepsy: Novel Susceptibility Loci in Idiopathic Generalized and Focal Epilepsies. *PLoS Genetics*, 6(5):e1000962, may 2010. doi:10.1371/journal.pgen.1000962.

[120] I. Helbig, M. E. M. Swinkels, E. Aten, A. Caliebe, R. van 't Slot *et al.* Structural genomic variation in childhood epilepsies with complex phenotypes. *European Journal of Human Genetics*, 22(7):896–901, jul 2014. doi:10.1038/ejhg.2013.262.

[121] E. L. Heinzen, R. A. Radtke, T. J. Urban, G. L. Cavalleri, C. Depondt *et al.* Rare deletions at 16p13.11 predispose to a diverse spectrum of sporadic epilepsy syndromes. *American journal of human genetics*, 86(5):707–18, may 2010. doi:10.1016/j.ajhg.2010.03.018.

[122] P. Striano, A. Coppola, R. Paravidino, M. Malacarne, S. Gimelli *et al.* Clinical significance of rare copy number variations in epilepsy: a case-control survey using microarray-based comparative genomic hybridization. *Archives of neurology*, 69(3):322–30, mar 2012. doi:10.1001/archneurol.2011.1999.

[123] H. Olson, Y. Shen, J. Avallone, B. R. Sheidley, R. Pinsky *et al.* Copy number variation plays an important role in clinical epilepsy. *Annals of neurology*, 75 (6):943–58, jun 2014. doi:10.1002/ana.24178.

[124] H. Mefford. Copy number variant analysis from exome data in 349 patients with epileptic encephalopathy. *Annals of Neurology*, 78(2):323–328, aug 2015. doi:10.1002/ana.24457.

[125] L. Addis, R. E. Rosch, A. Valentin, A. Makoff, R. Robinson *et al.* Analysis of rare copy number variation in absence epilepsies. *Neurology Genetics*, 2(2): e56, apr 2016. doi:10.1212/NXG.0000000000000056.

[126] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature genetics*, 45 (10):1134–1140, sep 2013. doi:10.1038/ng.2760.

[127] M. Baudis. Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC cancer*, 7:226, dec 2007. doi:10.1186/1471-2407-7-226.

[128] T.-m. Kim, R. Xi, L. J. Luquette, R. W. Park, M. D. Johnson *et al.* Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome research*, 23(2):217–27, feb 2013. doi:10.1101/gr.140301.112.

[129] J. Jen, H. Kim, S. Piantadosi, Z. F. Liu, R. C. Levitt *et al.* Allelic loss of chromosome 18q and prognosis in colorectal cancer. *The New England journal of medicine*, 331(4):213–21, jul 1994. doi:10.1056/NEJM199407283310401.

[130] D. M. Roy, L. A. Walsh, A. Desrichard, J. T. Huse, W. Wu *et al.* Integrated Genomics for Pinpointing Survival Loci within Arm-Level Somatic Copy Number Alterations. *Cancer Cell*, 29(5):737–750, 2016. doi:10.1016/j.ccell.2016.03.025.

[131] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer Statistics, 2017. *CA: a cancer journal for clinicians*, 67(1):7–30, jan 2017. doi:10.3322/caac.21387.

[132] T. J. O'Grady, M. A. Gates, and F. P. Boscoe. Thyroid cancer incidence attributable to overdiagnosis in the United States 1981-2011. *International journal of cancer*, 137(11):2664–73, dec 2015. doi:10.1002/ijc.29634.

[133] C. R. Scoggins, M. I. Ross, D. S. Reintgen, R. D. Noyes, J. S. Goydos *et al.* Gender-related differences in outcome for melanoma patients. *Annals of surgery*, 243(5):693–8; discussion 698–700, may 2006. doi:10.1097/01.sla.0000216771.81362.6b.

[134] S. Wirén, C. Häggström, H. Ulmer, J. Manjer, T. Bjørge *et al.* Pooled cohort study on height and risk of cancer and cancer death. *Cancer causes & control : CCC*, 25(2):151–9, feb 2014. doi:10.1007/s10552-013-0317-7.

[135] R. B. Walter, T. M. Brasky, S. A. Buckley, J. D. Potter, and E. White. Height as an explanatory factor for sex differences in human cancer. *Journal of the National Cancer Institute*, 105(12):860–8, jun 2013. doi:10.1093/jnci/djt102.

[136] O. B. Eden, G. Harrison, S. Richards, J. S. Lilleyman, C. C. Bailey *et al.* Long-term follow-up of the United Kingdom Medical Research Council protocols for childhood acute lymphoblastic leukaemia, 1980-1997. Medical Research Council Childhood Leukaemia Working Party. *Leukemia*, 14(12):2307–20, dec 2000.

[137] J. J. König, W. Teubel, J. C. Romijn, F. H. Schröder, and A. Hagemeijer. Gain and loss of chromosomes 1, 7, 8, 10, 18, and Y in 46 prostate cancers. *Human pathology*, 27(7):720–7, jul 1996.

[138] G. Sauter, H. Moch, U. Wagner, H. Novotna, T. C. Gasser *et al.* Y chromosome loss detected by FISH in bladder cancer. *Cancer genetics and cytogenetics*, 82 (2):163–9, jul 1995.

[139] S.-J. Park, S.-Y. Jeong, and H. J. Kim. Y chromosome loss and other genomic alterations in hepatocellular carcinoma cell lines analyzed by CGH and CGH array. *Cancer genetics and cytogenetics*, 166(1):56–64, apr 2006. doi:10.1016/j.cancergencyto.2005.08.022.

[140] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*, 136(5):E359–86, mar 2015. doi:10.1002/ijc.29210.

[141] W. H. Chow, G. Gridley, J. F. Fraumeni, and B. Järvholm. Obesity, hypertension, and the risk of kidney cancer in men. *The New England journal of medicine*, 343(18):1305–11, nov 2000. doi:10.1056/NEJM200011023431804.

[142] Y. Sato, T. Yoshizato, Y. Shiraishi, S. Maekawa, Y. Okuno *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nature Genetics*, 45(8): 860–867, 2013. doi:10.1038/ng.2699.

[143] G. Scelo, Y. Riazalhosseini, L. Greger, L. Letourneau, M. Gonzàlez-Porta *et al.* Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nature Communications*, 5(May):5135, oct 2014. doi:10.1038/ncomms6135.

[144] F. Latif, K. Tory, J. Gnarra, M. Yao, F. Duh *et al.* Identification of the von Hippel-Lindau disease tumor suppressor gene. *Science*, 260(5112):1317–1320, may 1993. doi:10.1126/science.8493574.

[145] R. E. Banks, P. Tirukonda, C. Taylor, N. Hornigold, D. Astuti *et al.* Genetic and epigenetic analysis of von Hippel-Lindau (VHL) gene alterations and relationship with clinical variables in sporadic renal cancer. *Cancer research*, 66 (4):2000–11, feb 2006. doi:10.1158/0008-5472.CAN-05-3074.

[146] I. J. Frew and H. Moch. A clearer view of the molecular complexity of clear cell renal cell carcinoma. *Annual review of pathology*, 10:263–89, 2015. doi:10.1146/annurev-pathol-012414-040306.

[147] P. R. Benusiglio, S. Couvé, B. Gilbert-Dussardier, S. Deveaux, H. Le Jeune *et al.* A germline mutation in PBRM1 predisposes to renal cell carcinoma. *Journal of Medical Genetics*, 52(6):426–430, jun 2015. doi:10.1136/jmedgenet-2014-102912.

[148] S. Carvalho, A. C. Vítor, S. C. Sridhara, F. B. Martins, A. C. Raposo *et al.* SETD2 is required for DNA double-strand break repair and activation of the p53-mediated checkpoint. *eLife*, 3:e02482, may 2014. doi:10.7554/eLife.02482.

[149] T. Popova, L. Hebert, V. Jacquemin, S. Gad, V. Caux-Moncoutier *et al.* Germline BAP1 mutations predispose to renal cell carcinomas. *American Journal of Human Genetics*, 92(6):974–980, jun 2013. doi:10.1016/j.ajhg.2013.04.012.

[150] T. Ito, J. Pei, E. Dulaimi, C. Menges, P. H. Abbosh *et al.* Genomic Copy Number Alterations in Renal Cell Carcinoma with Sarcomatoid Features. *The Journal of urology*, 195(4 Pt 1):852–8, apr 2016. doi:10.1016/j.juro.2015.10.180.

[151] G. L. Dalgliesh, K. Furge, C. Greenman, L. Chen, G. Bignell *et al.* Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature*, 463(7279):360–3, jan 2010. doi:10.1038/nature08672.

[152] I. Varela, P. Tarpey, K. Raine, D. Huang, C. K. Ong *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*, 469(7331):539–42, jan 2011. doi:10.1038/nature09639.

[153] C. J. Creighton, M. Morgan, P. H. Gunaratne, D. A. Wheeler, R. A. Gibbs *et al.* Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49, jun 2013. doi:10.1038/nature12222.

[154] A. L. Harris. Hypoxia — a Key Regulatory Factor in Tumour Growth. *Nature Reviews Cancer*, 2(1):38–47, jan 2002. doi:10.1038/nrc704.

[155] U. Capitanio and F. Montorsi. Renal cancer. *Lancet (London, England)*, 387 (10021):894–906, feb 2016. doi:10.1016/S0140-6736(15)00046-X.

[156] R. V. Pierre and H. C. Hoagland. Age-associated aneuploidy: loss of Y chromosome from human bone marrow cells with aging. *Cancer*, 30(4):889–94, oct 1972.

[157] W. Zhou, M. J. Machiela, N. D. Freedman, N. Rothman, N. Malats *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nature Genetics*, 48(5):563–568, apr 2016. doi:10.1038/ng.3545.

[158] M. B. Wozniak, F. Le Calvez-Kelm, B. Abedi-Ardekani, G. Byrnes, G. Durand *et al.* Integrative Genome-Wide Gene Expression Profiling of Clear Cell Renal Cell Carcinoma in Czech Republic and in the United States. *PLoS ONE*, 8 (3):e57886, 2013. doi:10.1371/journal.pone.0057886.

[159] H. Y. Wong, G. M. Wang, S. Croessmann, D. J. Zabransky, D. Chu *et al.* TMSB4Y is a candidate tumor suppressor on the Y chromosome and is deleted in male breast cancer. *Oncotarget*, 6(42):44927–40, dec 2015. doi:10.18632/oncotarget.6743.

[160] G. Perinchery, M. Sasaki, A. Angan, V. Kumar, P. Carroll *et al.* Deletion of Y-chromosome specific genes in human prostate cancer. *The Journal of urology*, 163(4):1339–42, apr 2000. doi:10.1016/S0022-5347(05)67774-9.

[161] M. G. Lee, J. Norman, A. Shilatifard, and R. Shiekhattar. Physical and functional association of a trimethyl H3K4 demethylase and Ring6a/MBLR, a polycomb-like protein. *Cell*, 128(5):877–87, mar 2007. doi:10.1016/j.cell.2007.02.004.

[162] Z. Jangravi, M. S. Tabar, M. Mirzaei, P. Parsamatin, H. Vakilian *et al.* Two Splice Variants of Y Chromosome-Located Lysine-Specific Demethylase 5D Have Distinct Function in Prostate Cancer Cell Line (DU-145). *Journal of proteome research*, 14(9):3492–502, sep 2015. doi:10.1021/acs.jproteome.5b00333.

[163] B. Rondinelli, D. Rosano, E. Antonini, M. Frenquelli, L. Montanini *et al.* Histone demethylase JARID1C inactivation triggers genomic instability in sporadic renal cancer. *Journal of Clinical Investigation*, 125(12):4625–4637, dec 2015. doi:10.1172/JCI81040.

[164] A. I. Agulnik, M. J. Mitchell, M. G. Mattei, G. Borsani, P. A. Avner *et al.* A novel X gene with a widely transcribed Y-linked homologue escapes X-inactivation in mouse and human. *Human Molecular Genetics*, 3(6):879–884, jun 1994. doi:10.1093/hmg/3.6.879.

[165] L. J. Walport, R. J. Hopkinson, M. Vollmar, S. K. Madden, C. Gileadi *et al.* Human UTY(KDM6C) Is a Male-specific N ϵ -Methyl Lysyl Demethylase. *Journal of Biological Chemistry*, 289(26):18302–18313, jun 2014. doi:10.1074/jbc.M114.555052.

[166] G. van Haaften, G. L. Dalgliesh, H. Davies, L. Chen, G. Bignell *et al.* Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nature Genetics*, 41(5):521–523, may 2009. doi:10.1038/ng.349.

[167] S. du Manoir, M. R. Speicher, S. Joos, E. Schröck, S. Popp *et al.* Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. *Human Genetics*, 90(6):590–610, feb 1993. doi:10.1007/BF00202476.

[168] L. a. Forsberg, C. Rasi, N. Malmqvist, H. Davies, S. Pasupulati *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nature Genetics*, 46(6):624–628, jun 2014. doi:10.1038/ng.2966.

[169] A. Kakarougkas, A. Ismail, A. L. Chambers, E. Riballo, A. D. Herbert *et al.* Requirement for PBAF in Transcriptional Repression and Repair at DNA Breaks in Actively Transcribed Regions of Chromatin. *Molecular Cell*, 55(5): 723–732, sep 2014. doi:10.1016/j.molcel.2014.06.028.

[170] J. M. Simon, K. E. Hacker, D. Singh, A. R. Brannon, J. S. Parker *et al.* Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects. *Genome Research*, 24(2):241–250, feb 2014. doi:10.1101/gr.158253.113.

[171] S. X. Pfister, S. Ahrabi, L. P. Zalmas, S. Sarkar, F. Aymard *et al.* SETD2-Dependent Histone H3K36 Trimethylation Is Required for Homologous Recombination Repair and Genome Stability. *Cell Reports*, 7(6):2006–2018, jun 2014. doi:10.1016/j.celrep.2014.05.026.

[172] Y. Riazalhosseini and M. Lathrop. Precision medicine from the renal cancer genome. *Nature Reviews Nephrology*, 12(11):655–666, nov 2016. doi:10.1038/nrneph.2016.133.

[173] J. Monlong, C. Meloche, G. Rouleau, P. Cossette, S. L. Girard *et al.* Human copy number variants are enriched in regions of low-mappability. *bioRxiv*, 2016. doi:10.1101/034165.

[174] M. Zhao, Q. Wang, Q. Wang, P. Jia, and Z. Zhao. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, 14(Suppl 11):S1, 2013. doi:10.1186/1471-2105-14-S11-S1.

[175] M. Pirooznia, F. Goes, and P. P. Zandi. Whole-genome CNV analysis: Advances in computational approaches. *Frontiers in Genetics*, 6(MAR):1–9, 2015. doi:10.3389/fgene.2015.00138.

[176] G. Glusman, A. Severson, V. Dhankani, M. Robinson, T. Farrah *et al.* Identification of copy number variants in whole-genome data using reference coverage profiles. *Frontiers in Genetics*, 5(FEB):1–13, 2015. doi:10.3389/fgene.2015.00045.

[177] E. L. van Dijk, Y. Jaszczyszyn, and C. Thermes. Library preparation methods for next-generation sequencing: tone down the bias. *Experimental cell research*, 322(1):12–20, mar 2014. doi:10.1016/j.yexcr.2014.01.008.

[178] S. F. Berkovic, R. A. Howell, D. A. Hay, and J. L. Hopper. Epilepsies in twins: Genetics of the major epilepsy syndromes. *Annals of Neurology*, 43(4): 435–445, apr 1998. doi:10.1002/ana.410430405.

[179] F. Zara, A. Bianchi, G. Avanzini, S. Di Donato, B. Castellotti *et al.* Mapping of genes predisposing to idiopathic generalized epilepsy. *Human Molecular Genetics*, 4(7):1201–7, jul 1995.

[180] D. Kasperavičiüte, C. B. Catarino, E. L. Heinzen, C. Depondt, G. L. Cavalleri *et al.* Common genetic variation and susceptibility to partial epilepsies: A genome-wide association study. *Brain*, 133(7):2136–2147, jul 2010. doi:10.1093/brain/awq130.

[181] Y. Guo, L. W. Baum, P. C. Sham, V. Wong, P. W. Ng *et al.* Two-stage genome-wide association study identifies variants in CAMSAP1L1 as susceptibility loci for epilepsy in Chinese. *Human Molecular Genetics*, 21(5):1184–1189, mar 2012. doi:10.1093/hmg/ddr550.

[182] International League Against Epilepsy Consortium on Complex Epilepsies. Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *The Lancet Neurology*, 13(9):893–903, sep 2014. doi:10.1016/S1474-4422(14)70171-1.

[183] A. Escayg, B. T. MacDonald, M. H. Meisler, S. Baulac, G. Huberfeld *et al.* Mutations of SCN1A, encoding a neuronal sodium channel, in two families with GEFS+2. *Nature genetics*, 24(4):343–5, apr 2000. doi:10.1038/74159.

[184] L. Claes, B. Ceulemans, D. Audenaert, K. Smets, A. Löfgren *et al.* De novo SCN1A mutations are a major cause of severe myoclonic epilepsy of infancy. *Human mutation*, 21(6):615–21, jun 2003. doi:10.1002/humu.10217.

[185] I. Helbig, H. C. Mefford, A. J. Sharp, M. Guipponi, M. Fichera *et al.* 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nature genetics*, 41(2):160–2, feb 2009. doi:10.1038/ng.292.

[186] C. G. F. de Kovel, H. Trucks, I. Helbig, H. C. Mefford, C. Baker *et al.* Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. *Brain : a journal of neurology*, 133(Pt 1):23–32, jan 2010. doi:10.1093/brain/awp262.

[187] C. Biervert. A Potassium Channel Mutation in Neonatal Human Epilepsy. *Science*, 279(5349):403–406, jan 1998. doi:10.1126/science.279.5349.403.

[188] D. Lal, A.-K. Ruppert, H. Trucks, H. Schulz, C. G. de Kovel *et al.* Burden Analysis of Rare Microdeletions Suggests a Strong Impact of Neurodevelopmental Genes in Genetic Generalised Epilepsies. *PLOS Genetics*, 11(5): e1005226, may 2015. doi:10.1371/journal.pgen.1005226.

[189] M. Boivin, M. Brendgen, G. Dionne, L. Dubois, D. Pérusse *et al.* The Quebec Newborn Twin Study Into Adolescence: 15 Years Later. *Twin Research and Human Genetics*, 16(01):64–69, feb 2013. doi:10.1017/thg.2012.129.

[190] I. Scheinin, D. Sie, H. Bengtsson, M. A. van de Wiel, A. B. Olshen *et al.* DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome research*, 24(12):2022–32, dec 2014. doi:10.1101/gr.175141.114.

[191] R. Xi, A. G. Hadjipanayis, L. J. Luquette, T.-M. Kim, E. Lee *et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proceedings of the National Academy of Sciences*, 108 (46):E1128–E1136, nov 2011. doi:10.1073/pnas.1110574108.

[192] P. H. Sudmant, S. Mallick, B. J. Nelson, F. Hormozdiari, N. Krumm *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science*, 349(6253):aab3761–aab3761, sep 2015. doi:10.1126/science.aab3761.

[193] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536 (7616):285–291, aug 2016. doi:10.1038/nature19057.

[194] A. V. Delgado-Escueta, B. P. Koeleman, J. N. Bailey, M. T. Medina, and R. M. Durón. The quest for Juvenile Myoclonic Epilepsy genes. *Epilepsy & Behavior*, 28:S52–S57, jul 2013. doi:10.1016/j.yebeh.2012.06.033.

[195] K. G. Ardlie, D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, may 2015. doi:10.1126/science.1262110.

[196] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195, sep 2012. doi:10.1126/science.1222794.

[197] L. M. Dibbens, B. de Vries, S. Donatello, S. E. Heron, B. L. Hodgson *et al.* Mutations in DEPDC5 cause familial focal epilepsy with variable foci. *Nature Genetics*, 45(5):546–551, may 2013. doi:10.1038/ng.2599.

[198] S. Ishida, F. Picard, G. Rudolf, E. Noé, G. Achaz *et al.* Mutations of DEPDC5 cause autosomal dominant focal epilepsies. *Nature Genetics*, 45(5):552–555, may 2013. doi:10.1038/ng.2601.

[199] E. C. Galizia, C. T. Myers, C. Leu, C. G. F. de Kovel, T. Afrikanova *et al.* CHD2 variants are a risk factor for photosensitivity in epilepsy. *Brain*, 138 (5):1198–1208, may 2015. doi:10.1093/brain/awv052.

[200] J. Elia, X. Gai, H. M. Xie, J. C. Perin, E. Geiger *et al.* Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Molecular Psychiatry*, 15(6):637–646, jun 2010. doi:10.1038/mp.2009.57.

[201] N. Choucair, C. Mignon-Ravix, P. Cacciagli, J. Abou Ghoch, A. Fawaz *et al.* Evidence that homozygous PTPRD gene microdeletion causes trigonocephaly, hearing loss, and intellectual disability. *Molecular Cytogenetics*, 8(1):39, dec 2015. doi:10.1186/s13039-015-0149-0.

[202] D. Speed, C. Hoggart, S. Petrovski, I. Tachmazidou, A. Coffey *et al.* A genome-wide association study and biological pathway analysis of epilepsy prognosis in a prospective cohort of newly treated epilepsy. *Human molecular genetics*, 23(1):247–58, jan 2014. doi:10.1093/hmg/ddt403.

[203] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, mar 2010. doi:10.1093/bioinformatics/btp698.

[204] V. Seshan and A. Olshen. DNAcopy: DNA copy number data analysis. *R package version 1.50.1.*, 2017.

[205] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–92, mar 2013. doi:10.1093/bib/bbs017.

[206] T. Derrien, J. Estellé, S. Marco Sola, D. G. Knowles, E. Raineri *et al.* Fast computation and applications of genome mappability. *PloS one*, 7(1):e30377, jan 2012. doi:10.1371/journal.pone.0030377.

[207] 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, sep 2009. doi:10.1101/gr.092981.109.

[208] W. P. Kloosterman, L. C. Francioli, F. Hormozdiari, T. Marschall, J. Y. Hehir-Kwa *et al.* Characteristics of de novo structural changes in the human genome. *Genome Research*, 25(6):792–801, jun 2015. doi:10.1101/gr.185041.114.

[209] S. M. Mirkin. Expandable DNA repeats and human disease. *Nature*, 447 (7147):932–940, 2007. doi:10.1038/nature05977.

[210] C. M. B. Carvalho and J. R. Lupski. Mechanisms underlying structural variant formation in genomic disorders. *Nature reviews. Genetics*, 17(4):224–38, apr 2016. doi:10.1038/nrg.2015.25.

[211] M. Gymrek, D. Golan, S. Rosset, and Y. Erlich. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research*, 22(6):1154–1162, jun 2012. doi:10.1101/gr.135780.111.

[212] P. E. Warburton, D. Hasson, F. Guillem, C. Lescale, X. Jin *et al.* Analysis of the largest tandemly repeated DNA families in the human genome. *BMC genomics*, 9:533, 2008. doi:10.1186/1471-2164-9-533.

[213] J. O. Korbel, A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, N.Y.)*, 318(5849):420–6, oct 2007. doi:10.1126/science.1149504.

[214] H. H. Kazazian and J. V. Moran. Mobile DNA in Health and Disease. *New England Journal of Medicine*, 377(4):361–370, 2017. doi:10.1056/NEJMra1510092.

[215] Y. Mostovoy, M. Levy-Sakin, J. Lam, E. T. Lam, A. R. Hastie *et al.* A hybrid approach for de novo human genome sequence assembly and phasing. *Nature Methods*, 13(7):587–590, may 2016. doi:10.1038/nmeth.3865.

[216] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.* BLAST+: architecture and applications. *BMC bioinformatics*, 10:421, dec 2009. doi:10.1186/1471-2105-10-421.

[217] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway *et al.* Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12, 2004. doi:10.1186/gb-2004-5-2-r12.

[218] K. R. Rosenbloom, J. Armstrong, G. P. Barber, J. Casper, H. Clawson *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Research*, 43(D1):D670–D681, 2015. doi:10.1093/nar/gku1177.

[219] D.-Q. Nguyen, C. Webber, and C. P. Ponting. Bias of selection on human copy-number variants. *PLoS genetics*, 2(2):e20, feb 2006. doi:10.1371/journal.pgen.0020020.

[220] 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. doi:10.1002/mc.20499.

[221] T. F. Willems, M. Gymrek, G. Highnam, D. Mittelman, and Y. Erlich. The landscape of human STR variation. *Genome Research*, pages 1894–1904, aug 2014. doi:10.1101/gr.177774.114.

[222] A. Fungtammasan, G. Ananda, S. E. Hile, M. S.-w. Su, C. Sun *et al.* Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Research*, 25(5):736–749, may 2015. doi:10.1101/gr.185892.114.

[223] D. Kelley and J. Rinn. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome biology*, 13(11):R107, nov 2012. doi:10.1186/gb-2012-13-11-r107.

[224] X. Lu, F. Sachs, L. Ramsay, P.-É. Jacques, J. Göke *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. doi:10.1038/nsmb.2799.

[225] J. A. Bailey, G. Liu, and E. E. Eichler. An Alu transposition model for the origin and expansion of human segmental duplications. *American journal of human genetics*, 73(4):823–34, oct 2003. doi:10.1086/378594.

[226] J. M. Kidd, T. Graves, T. L. Newman, R. Fulton, H. S. Hayden *et al.* A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, 143(5):837–847, 2010. doi:10.1016/j.cell.2010.10.027.

[227] J.-C. Lambert, S. Heath, G. Even, D. Campion, K. Sleegers *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nature genetics*, 41(10):1094–9, oct 2009. doi:10.1038/ng.439.

[228] R. M. Durbin, D. L. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley *et al.* A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, oct 2010. doi:10.1038/nature09534.

[229] L.-P. Wong, R. T.-H. Ong, W.-T. Poh, X. Liu, P. Chen *et al.* Deep whole-genome sequencing of 100 southeast Asian Malays. *American journal of human genetics*, 92(1):52–66, jan 2013. doi:10.1016/j.ajhg.2012.12.005.

[230] D. F. Gudbjartsson, H. Helgason, S. A. Gudjonsson, F. Zink, A. Oddson *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, 47(5):435–444, 2015. doi:10.1038/ng.3247.

[231] R. K. Yuen, D. Merico, M. Bookman, J. L. Howe, B. Thiruvahindrapuram *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature Neuroscience*, 20(4):602–611, 2017. doi:10.1038/nn.4524.

[232] G. W. Beecham, J. Bis, E. Martin, S.-H. Choi, A. L. DeStefano *et al.* The Alzheimer's Disease Sequencing Project: Study design and sample selection. *Neurology Genetics*, 3(5):e194, 2017. doi:10.1212/NXG.0000000000000194.

[233] J. P. Dumanski, J. C. Lambert, C. Rasi, V. Giedraitis, H. Davies *et al.* Mosaic Loss of Chromosome y in Blood Is Associated with Alzheimer Disease. *American Journal of Human Genetics*, 98(6):1208–1219, 2016. doi:10.1016/j.ajhg.2016.05.014.

[234] T. N. Turner, F. Hormozdiari, M. H. Duyzend, S. A. McClymont, P. W. Hook *et al.* Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *American Journal of Human Genetics*, 98(1):58–74, 2016. doi:10.1016/j.ajhg.2015.11.023.

[235] C. Marshall, D. Howrigan, D. Merico, B. Thiruvahindrapuram, W. Wu *et al.* A contribution of novel CNVs to schizophrenia from a genome-wide study of 41,321 subjects. *bioRxiv*, page 040493, 2016. doi:10.1101/040493.

[236] 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, jun 2013. doi:10.1093/bioinformatics/btt151.

[237] E. Talevich, A. H. Shain, T. Botton, and B. C. Bastian. CNVkit: Copy number detection and visualization for targeted sequencing using off-target reads. *bioRxiv*, pages 1–27, oct 2014. doi:10.1101/010876.

[238] M. Fromer, J. L. Moran, K. Chambert, E. Banks, S. E. Bergen *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *American journal of human genetics*, 91(4):597–607, oct 2012. doi:10.1016/j.ajhg.2012.08.005.

[239] A. Magi, L. Tattini, I. Cifola, R. D'Aurizio, M. Benelli *et al.* EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome biology*, 14(10):R120, jan 2013. doi:10.1186/gb-2013-14-10-r120.

[240] B. Paten, A. M. Novak, J. M. Eizenga, and E. Garrison. Genome graphs and the evolution of genome inference. *Genome research*, 27(5):665–676, may 2017. doi:10.1101/gr.214155.116.

[241] E. Garrison, A. Novak, G. Hickey, J. Eizenga, E. Dawson *et al.* Sequence variation aware references and read mapping with vg : the variation graph toolkit. *bioRxiv*, pages 1–27, 2017. doi:10.1101/234856.

[242] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27(5):722–736, may 2017. doi:10.1101/gr.215087.116.

[243] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7:539, oct 2011. doi:10.1038/msb.2011.75.

Appendices

Appendix A

The contribution of the thesis author to other published works during the thesis period is listed below.

Contribution in another project from the Canadian Epilepsy Network (CENet):

- F. F. Hamdan, C. T. Myers, P. Cossette, P. Lemay, D. Spiegelman, A. D. Laporte, C. Nassif, O. Diallo, **J. Monlong**, M. Cadieux-Dion, S. Dobrzeniecka, C. Meloche, K. Retterer, M. T. Cho, J. A. Rosenfeld, W. Bi, C. Massicotte, M. Miguet, L. Brunga, B. M. Regan, K. Mo, C. Tam, A. Schneider, G. Hollingsworth, D. R. FitzPatrick, A. Donaldson, N. Canham, E. Blair, B. Kerr, A. E. Fry, R. H. Thomas, J. Shelagh, J. A. Hurst, H. Brittain, M. Blyth, R. R. Lebel, E. H. Gerkes, L. Davis-Keppen, Q. Stein, W. K. Chung, S. J. Dorison, P. J. Benke, E. Fassi, N. Corsten-Janssen, E.-J. Kamsteeg, F. T. Mau-Them, A.-L. Bruel, A. Verloes, K. Õunap, M. H. Wojcik, D. V. Albert, S. Venkateswaran, T. Ware, D. Jones, Y.-C. Liu, S. S. Mohammad, P. Bizargity, C. A. Bacino, V. Leuzzi, S. Martinelli, B. Dallapiccola, M. Tartaglia, L. Blumkin, K. J. Wierenga, G. Purcarin, J. J. O'Byrne, S. Stockler, A. Lehman, B. Keren, M.-C. Nougues, C. Mignot, S. Auvin, C. Nava, S. M. Hiatt, M. Bebin, Y. Shao, F. Scaglia, S. R. Lalani, R. E. Frye, I. T. Jarjour, S. Jacques, R.-M. Boucher, E. Riou, M. Srour, L. Carmant, A. Lortie, P. Major, P. Diadori, F. Dubeau, G. D'Anjou, G. Bourque, S. F. Berkovic, L. G. Sadleir, P. M. Campeau, Z. Kibar, R. G. Lafrenière, S. L. Girard, S. Mercimek-Mahmutoglu, C. Boelman, G. A. Rouleau, I. E. Scheffer, H. C. Mefford, D. M. Andrade, E. Rossignol, B. A. Minassian, and J. L. Michaud. High Rate of Recurrent De Novo Mutations in Developmental and Epileptic Encephalopathies. *The American Journal of Human Genetics*, 101(5):664–685, nov 2017. doi:10.1016/j.ajhg.2017.09.008.

Contributions to projects not directly relevant for the thesis topic:

- M. Mele, P. G. Ferreira, F. Reverter, D. S. DeLuca, **J. Monlong**, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segre, S. Djebali, A. Niarchou, GTEx Consortium, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, and R. Guigo. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, may 2015. doi:10.1126/science.aaa0355.

- D. D. Pervouchine, S. Djebali, A. Breschi, C. A. Davis, P. P. Barja, A. Dobin, A. Tanzer, J. Lagarde, C. Zaleski, L.-H. See, M. Fastuca, J. Drenkow, H. Wang, G. Bussotti, B. Pei, S. Balasubramanian, **J. Monlong**, A. Harmanci, M. Gerstein, M. A. Beer, C. Notredame, R. Guigó, and T. R. Gingeras. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nature communications*, 6:5903, jan 2015. doi:10.1038/ncomms6903.

- **J. Monlong**, M. Calvo, P. G. Ferreira, and R. Guigó. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nature Communications*, 5(May):4698, aug 2014. doi:10.1038/ncomms5698.

- P. G. Ferreira, P. Jares, D. Rico, G. Gomez-Lopez, A. Martinez-Trillos, N. Villamor, S. Ecker, A. Gonzalez-Perez, D. G. Knowles, **J. Monlong**, R. Johnson, V. Quesada, S. Djebali, P. Papasaikas, M. Lopez-Guerra, D. Colomer, C. Royo, M. Cazorla, M. Pinyol, G. Clot, M. Aymerich, M. Rozman, M. Kulis, D. Tamborero, A. Gouin, J. Blanc, M. Gut, I. Gut, X. S. Puente, D. G. Pisano, J. I. Martin-Subero, N. Lopez-Bigas, A. Lopez-Guillermo, A. Valencia, C. Lopez-Otin, E. Campo, and R. Guigo. Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Research*, 24(2):212–226, feb 2014. doi:10.1101/gr.152132.112.

- T. Lappalainen, M. Sammeth*, M. R. Friedländer*, P. a. C. 't Hoen*, **J. Monlong***, M. a. Rivas*, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, Á. Carracedo, S. E. Antonarakis, R. Häsler, A.-C. Syvänen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, and E. T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, sep 2013. doi:10.1038/nature12531.

Appendix B

Supplementary material for chapter 2 and its corresponding manuscript: **Loss of chromosome Y leads to down regulation of *KDM5D* and *KDM6C* epigenetic modifiers in clear cell renal cell carcinoma**.

Supplementary Tables

Characteristics	Discovery set				Validation set		Total	
	Females		Males		(males only)		(males)	
	N	%	N	%	N	%	N	%
Tumor grade								
1	1	*2.4*	1	*1.9*	3	*6.3*	4	*4.0*
2	28	*68.3*	26	*50.0*	23	*47.9*	49	*49.0*
3	5	*12.2*	13	*25.0*	17	*35.4*	30	*30.0*
4	7	*17.1*	12	*23.1*	5	*10.4*	17	*17.0*
Tumor stage								
I	27	*65.9*	24	*46.2*	29	*60.4*	53	*53.0*
II	5	*12.2*	4	*7.7*	1	*2.1*	5	*5.0*
III	5	*12.2*	15	*28.8*	6	*12.5*	21	*21.0*
IV	4	*9.8*	9	*17.3*	10	*20.8*	19	*19.0*
NA	0	*0.0*	0	*0.0*	2	*4.2*	2	*2.0*
Country of residence								
Czech Republic	8	*19.5*	20	*38.5*	23	*47.9*	43	*43.0*
Romania	5	*12.2*	8	*15.4*	0	*0.0*	8	*8.0*
Russia	15	*36.6*	8	*15.4*	18	*37.5*	26	*26.0*
UK	13	*31.7*	16	*30.8*	7	*14.6*	23	*23.0*
Age at surgery: median (range)	62 (39 - 83)		60 (40 - 79)		56 (38 - 74)		58 (38 - 79)	
Total	**41**		**52**		**48**		**100**	

Table S2.1: **Characteristics of patients included in the study.**

Gene	Chromosome	Fold-change (Differential expression)
KDM5D	Y	-1.3
USP9Y	Y	-1.4
ZFY	Y	-1.1
UTY/KDM6C	Y	-1.2
NLGN4Y	Y	-0.8
DDX3Y	Y	-1.8
EIF1AY	Y	-1.8
TMSB4Y	Y	-0.9
RPS4Y1	Y	-2.6

Table S2.2: **Genes differentially expressed between tumors with and without somatic LOY.** Fold-change of differential expression between the two tumor sets.

Supplementary Figures

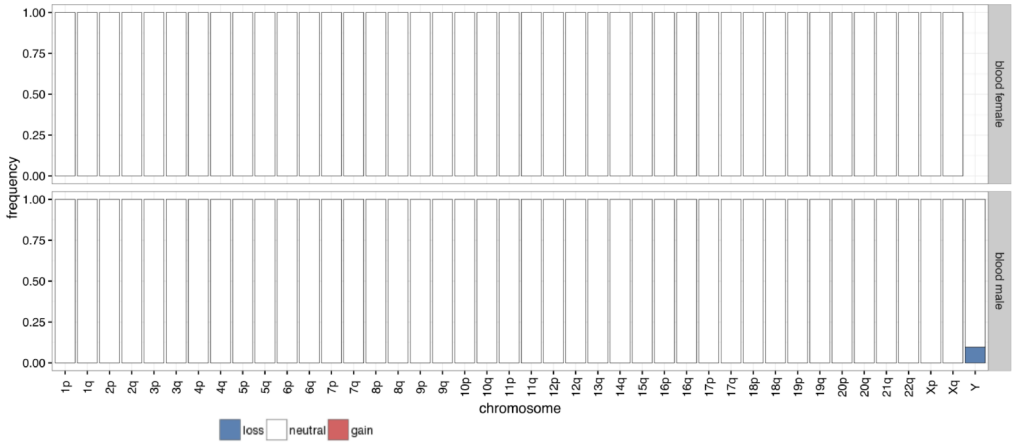


Figure S2.1: **Copy number analysis in peripheral blood.** Bar graphs show the frequency of copy number variations across the genome in peripheral blood. Frequencies are presented in samples from female and male cases separately.

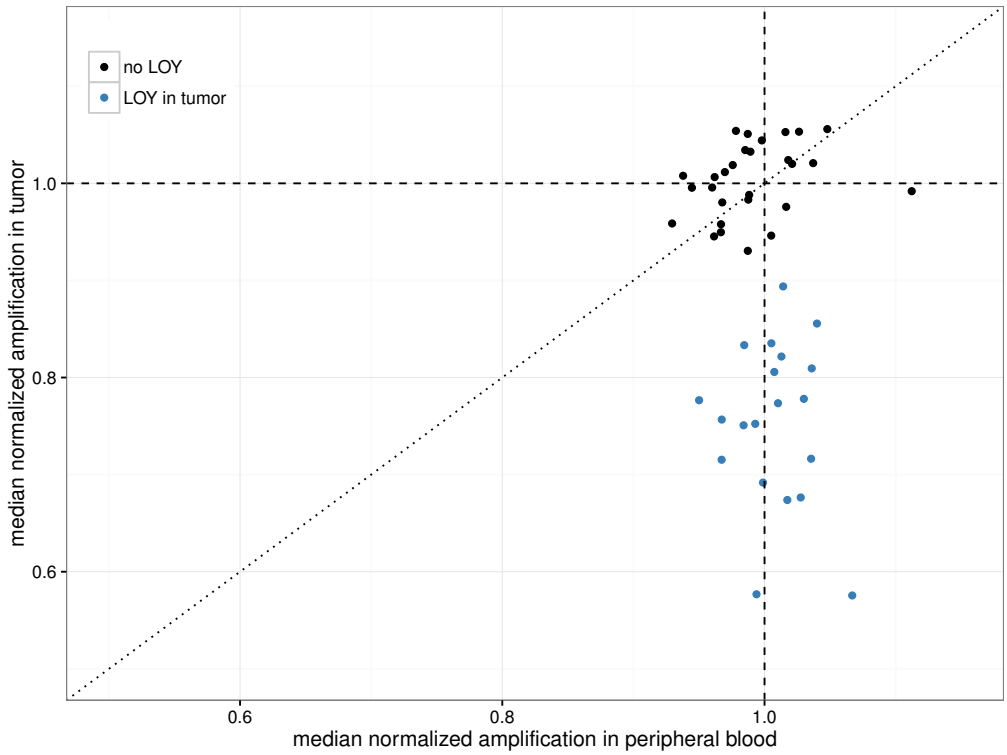


Figure S2.2: **Validation using PCR amplification.** Status of chromosome Y in tumors and patient-matched normal samples is shown for individual male subjects of the validation set. The PCR amplification values are normalized and summarized by their median in each sample. Individuals affected by somatic LOY are shown in blue

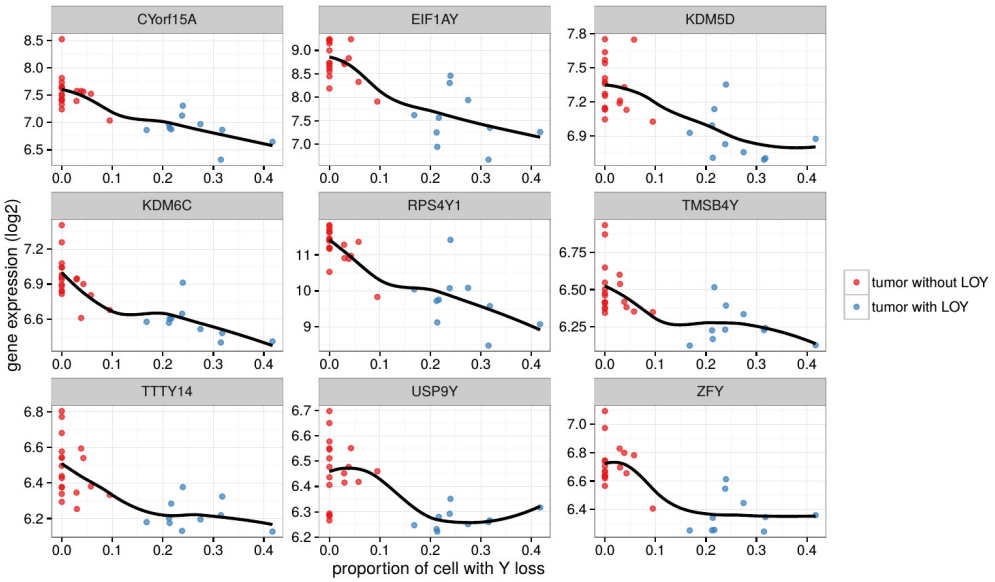


Figure S2.3: **Somatic LOY Y-linked genes down-regulation from array-based expression experiments.** The proportion of cells with Y loss was estimated by the PCR amplification values.

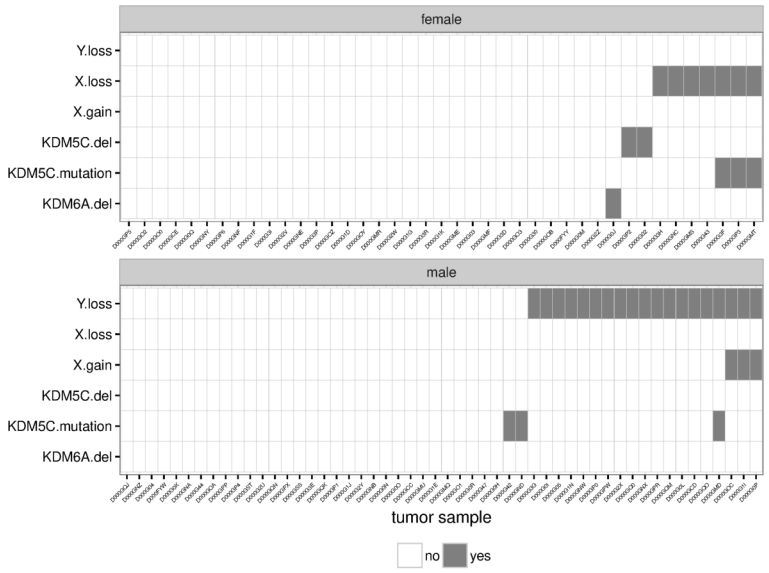


Figure S2.4: **Copy number aberrations of sex chromosomes and genomic status of X-linked epigenetic modifying genes in tumors of female and male patients.** Nearly half of the female tumors harboring somatic LOX are also affected by mutations of *KDM5C*

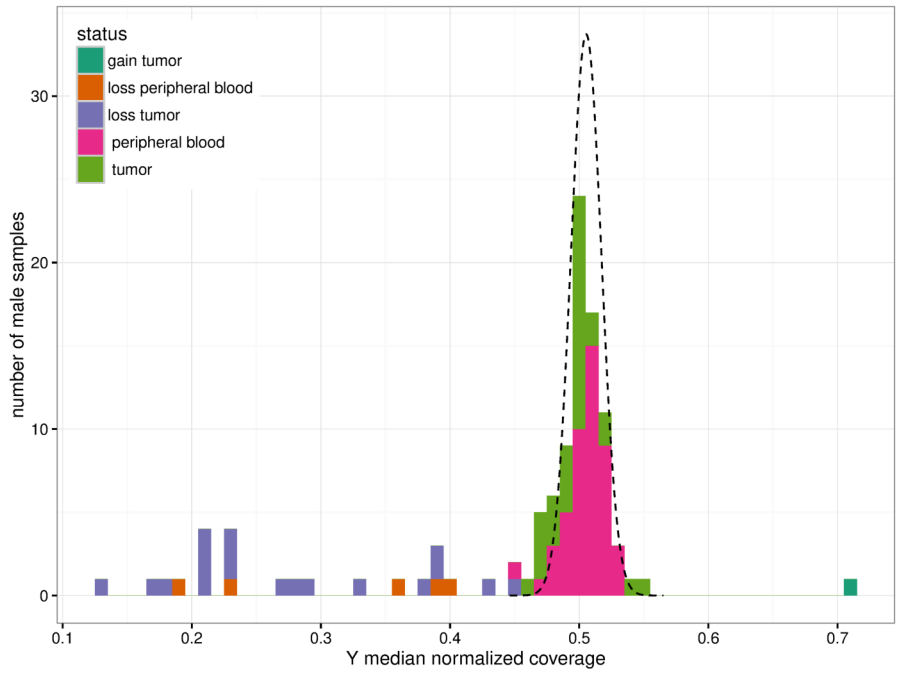


Figure S2.5: **Detecting loss of Y.** The main Gaussian, fitted on the median normalized coverage, is used to detect significant loss/gain. Each male sample, blood and tumor, is colored according to its loss/gain status.

Appendix C

Supplementary material for chapter 3 and its corresponding manuscript: **Global characterization of copy number variants in epilepsy patients from whole genome sequencing**.

Supplementary Tables

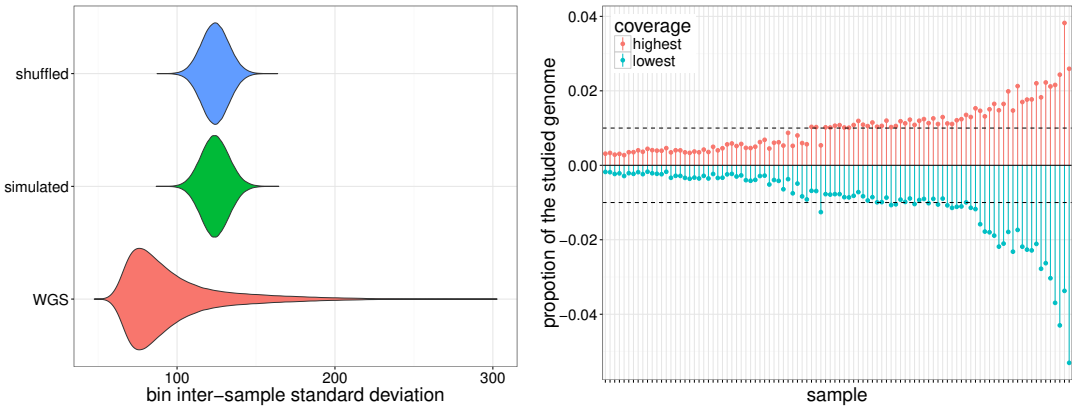
Table S3.1: **PopSV calls validated by RT-PCR.** The Excel file contains the location of each region, the CNV type, the number of carriers in the CENet cohorts, the maximum proportion of carriers in the CNV databases, Taqman probe ID and validation status. https://doi.org/10.1371/journal.pgen.1007285.s003

Patient	Epilepsy type	Syndrome	Copy number	Chr.	CNV start	CNV end	Exon disrupted	Taqman probe
CNET0188	Focal	Mesial temporal lobe sclerosis	1	2	141335001	141365000	*LRP1B*	Hs02078420_cn
CNET0084	Focal	Temporal lobe	1	4	120205001	120280000	*USP53;FABP2;C4orf3*	Hs04813260_cn
CNET0143	Generalized	Childhood absence epilepsy	1	5	65055001	65465000	*NLN;ERBB2IP;SREK1*	Hs03552554_cn
CNET0151	Generalized	Eyelid myoclonia epilepsy with absence	1	9	8600001	8770000	*PTPRD*	Hs06875003_cn
CNET0041	Generalized	Idiopathic generalized epilepsies	1	11	62625001	62645000	*SLC3A2*	Hs03777991_cn
CNET0066	Generalized	Idiopathic generalized epilepsies	1	13	67325001	67575000	*PCDH9*	Hs06378870_cn
CNET0025	Generalized	Early onset absence epilepsy (onset ¡4, absence with or without GTCs)	1	15	60735001	60805000	*RORA;NARG2*	Hs05369880_cn
CNET0195	Focal	Occipital lobe epilepsy	1	22	34095001	34200000	*LARGE*	Hs05575584_cn
CNET0005	Generalized	febrile sz, child onset GTCs	1	22	41960001	42050000	*PMM1;DESI1;CSDC2;XRCC6*	Hs05580065_cn

Table S3.2: Other pathogenic profiles.

Table S3.3: **Clinical features of epileptic patients.** The Excel file contains the type of epilepsy, age of onset, sex, family history, pharmaco-resistance and potential intellectual disabilities. https://doi.org/10.1371/journal.pgen.1007285.s002

Supplementary Figures



(a) (b)

Figure S3.1: **Variation and bias in whole-genome sequencing experiments in the epilepsy cohort.** a) Distribution of the bin inter-sample standard deviation coverage (red) and null distribution (blue: bins shuffled, green: simulated normal distribution). b) 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).

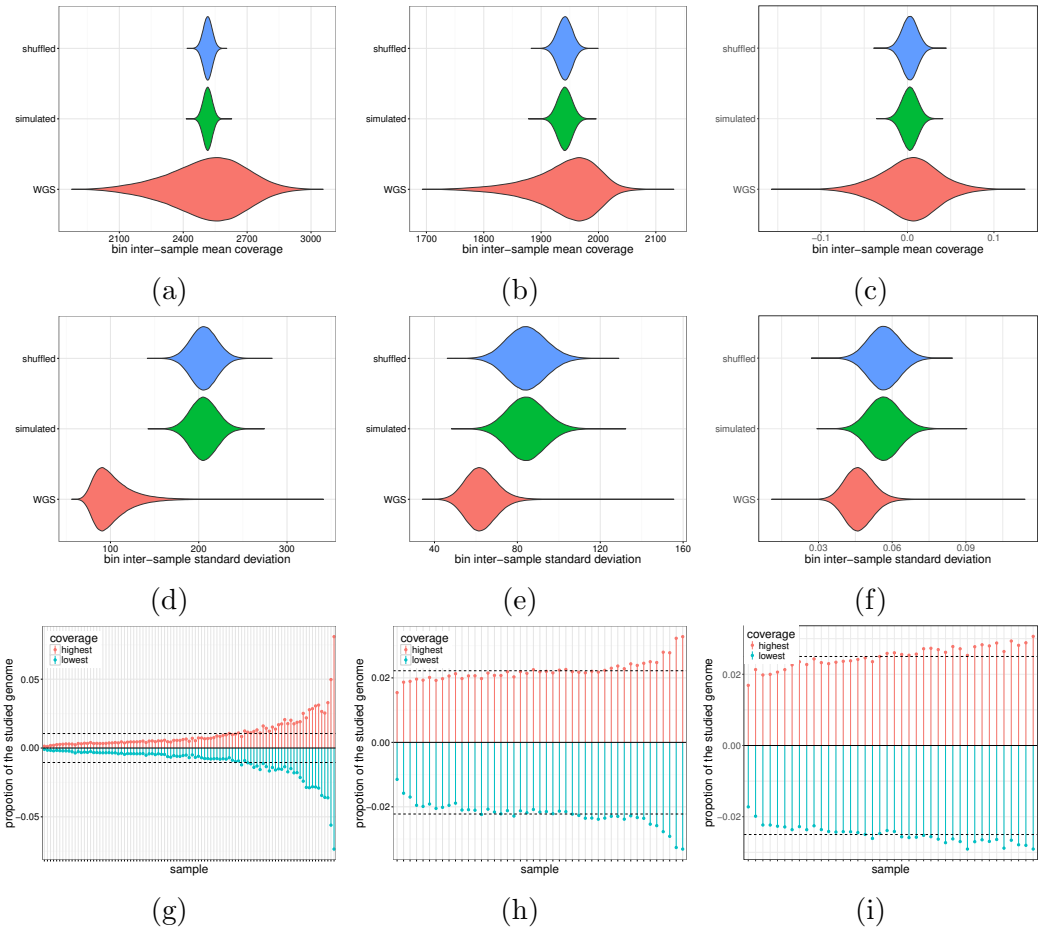
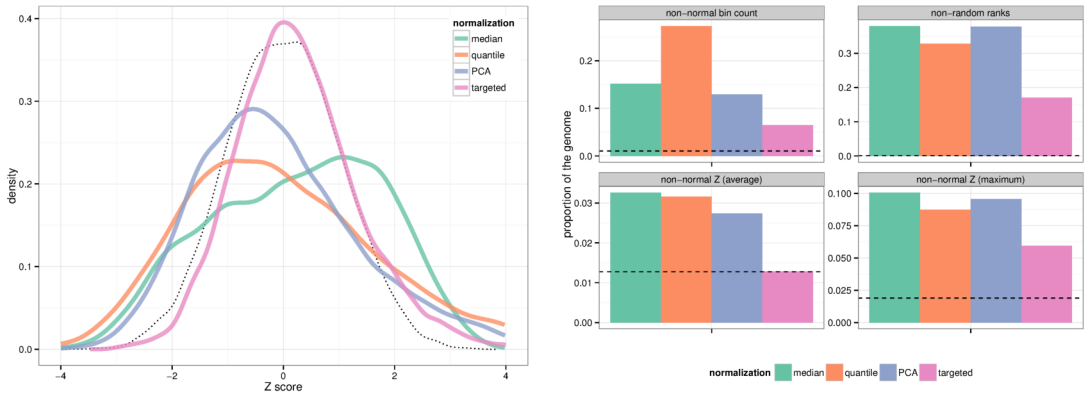


Figure S3.2: **Variation and bias in whole-genome sequencing experiments in the normals from CageKid (a,d,g), the twin dataset (b,e,h) and the twin dataset after using QDNAseq[190] correction (c,f,i).** a-c) Distribution of the bin inter-sample standard deviation coverage (red) and null distribution (blue: bins shuffled, green: simulated normal distribution). d-f) Same for the bin inter-sample standard deviation coverage. g-i) 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)

Figure S3.3: **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. Fewer bins have non-normal bin counts (top-left), the sample ranks are more random suggesting 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.

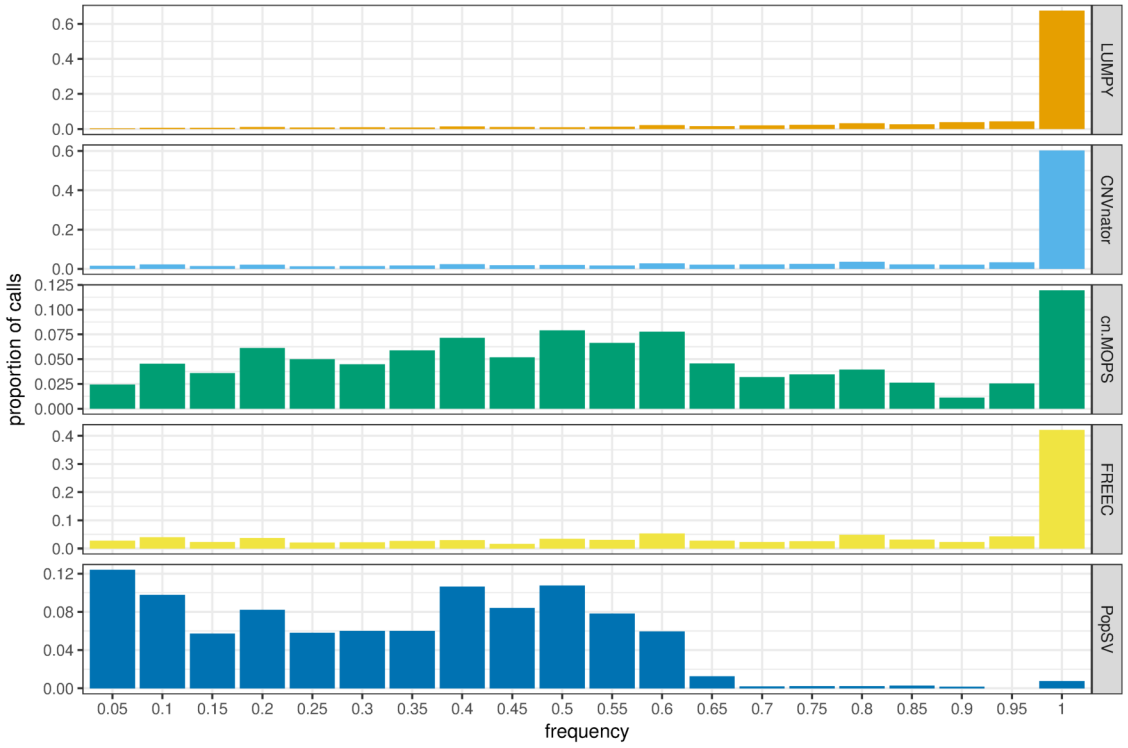


Figure S3.4: **Frequency of calls in an average sample from the Twin study.** The bars show the proportion of calls in an average samples (y-axis), grouped by the frequency of the call in the dataset (x-axis), for different methods.

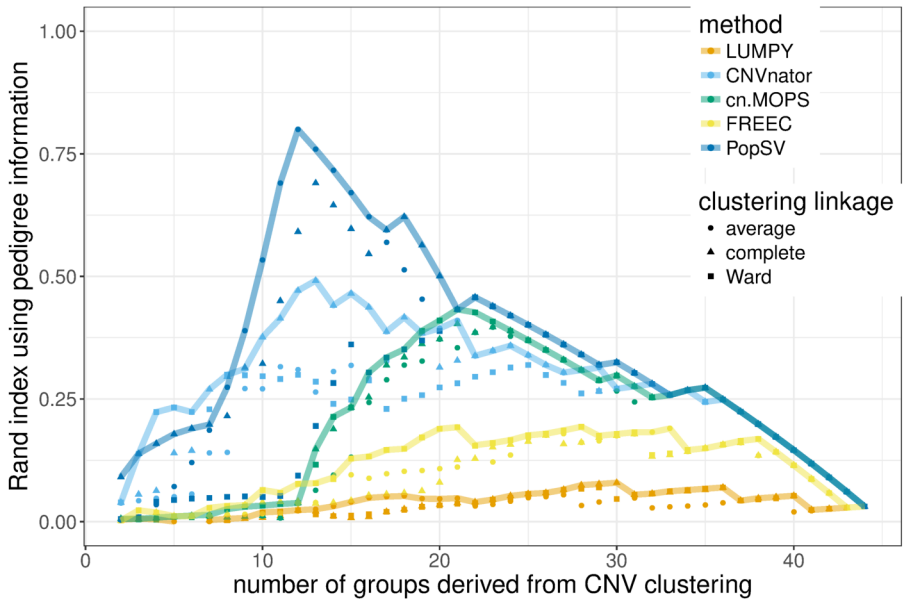


Figure S3.5: **CNV clustering and twin pedigree.** The hierarchical cluster tree from the CNV calls is cut at different levels (*x-axis*), cluster groups are compared to the known pedigree using the Rand index (*y-axis*). Different clustering linkage criteria (*point style*) are used and the one showing the best Rand index is highlighted by the line.

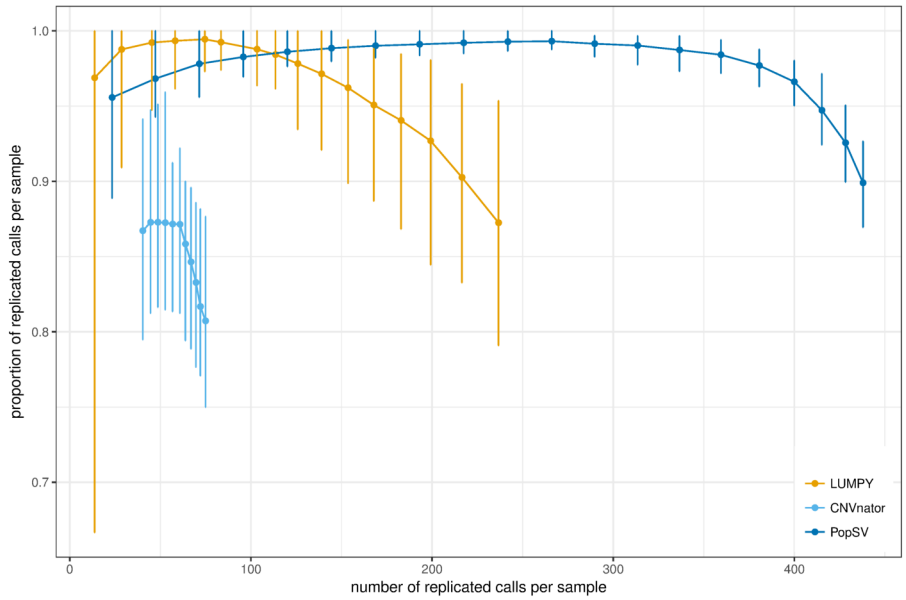


Figure S3.6: **Replication in twins for different significance thresholds.** Each point represents the number of replicated calls per sample (average across samples) and the proportion of replicated calls per sample. The vertical error bar shows the variation of the replication rate across the samples. The points and lines were computed by filtering calls at different significance levels (q-value for PopSV, number of supporting reads for LUMPY and eval1/eval2 for CNVnator, see Supplementary Information).

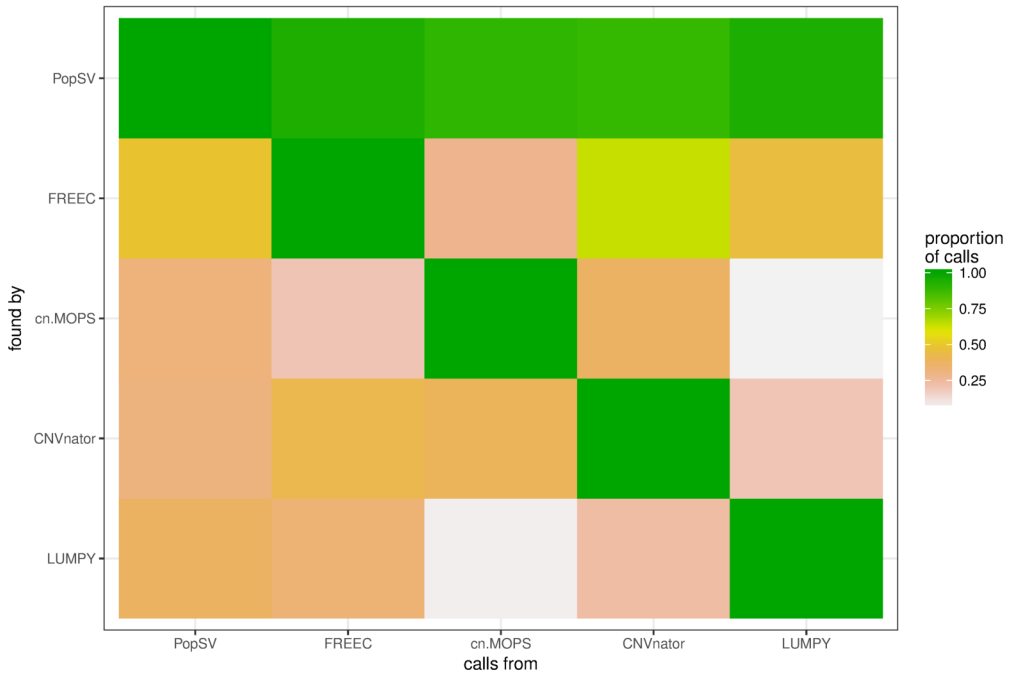


Figure S3.7: **Calls found by several methods.** Focusing on calls found by at least two methods, the heatmap shows the proportion of calls from one method (x-axis) that were also found by another (y-axis) on average per sample.

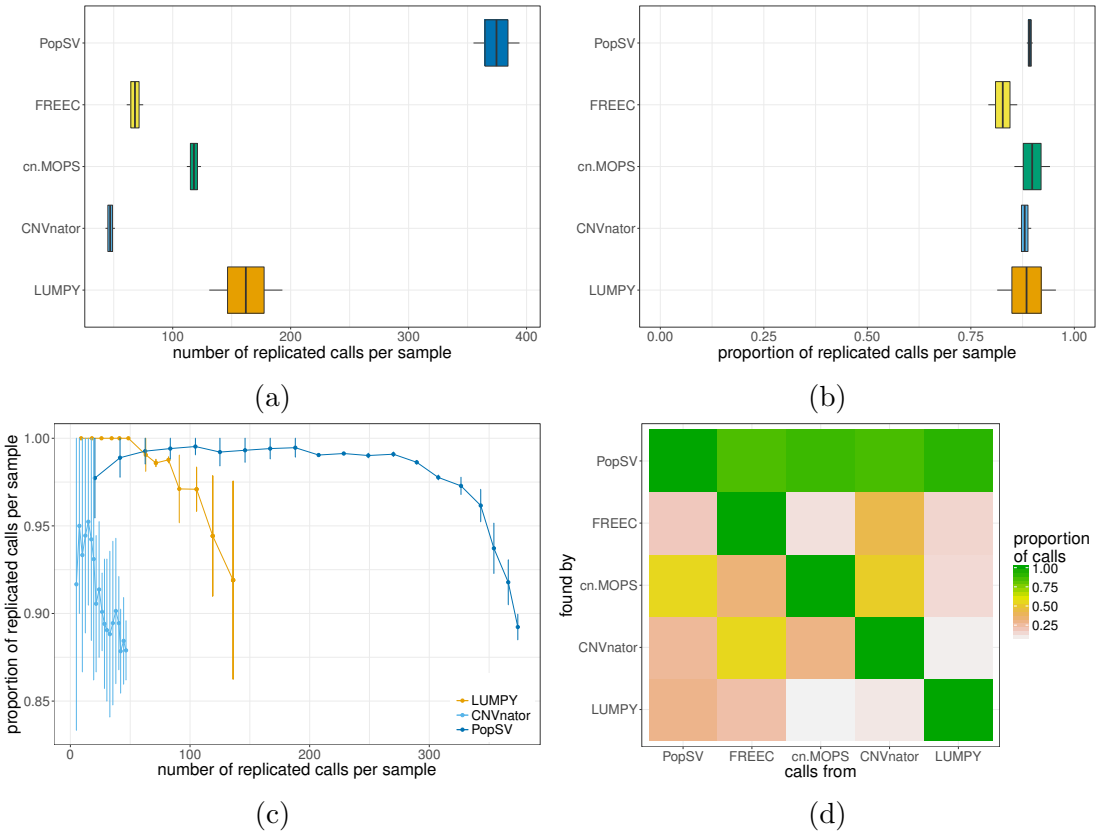


Figure S3.8: **Benchmark across paired normal/tumor in CageKid.** Number (a) and proportion (b) of germline calls replicated in the paired tumor in CageKid. c) Number and proportion of replicated calls when filtering calls at different significance levels. d) Focusing on calls found by at least two methods, the color shows the proportion of calls from one method (x-axis) that were also found by another (y-axis) on average per sample.

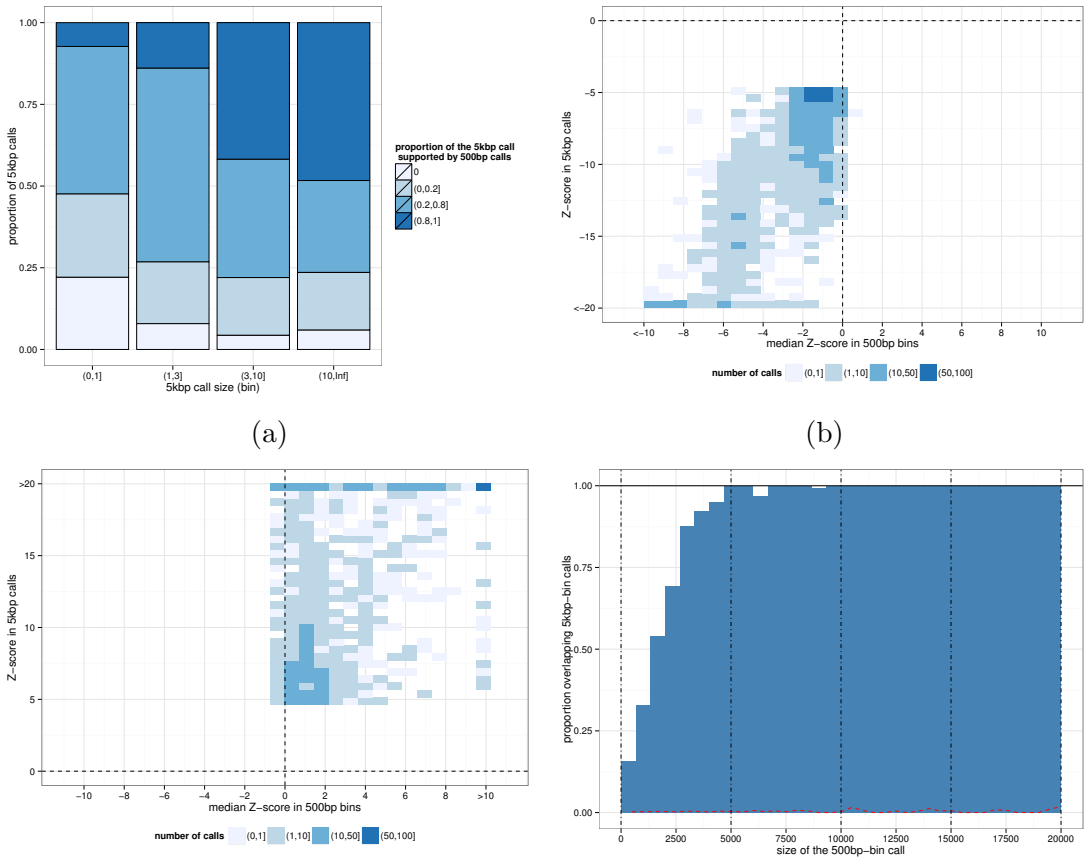


Figure S3.9: **Comparison of PopSV results using different bin sizes.** a) 5 Kbp calls of different sizes (x-axis) are split according to the proportion of the call supported by 500 bp calls. The Z-score of 500 bp bins in 5 Kbp calls is consistent with the call for deletion b) and duplication c) signal. 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. d) Proportion of 500 bp calls of different sizes (x-axis) overlapping a 5 Kbp call.

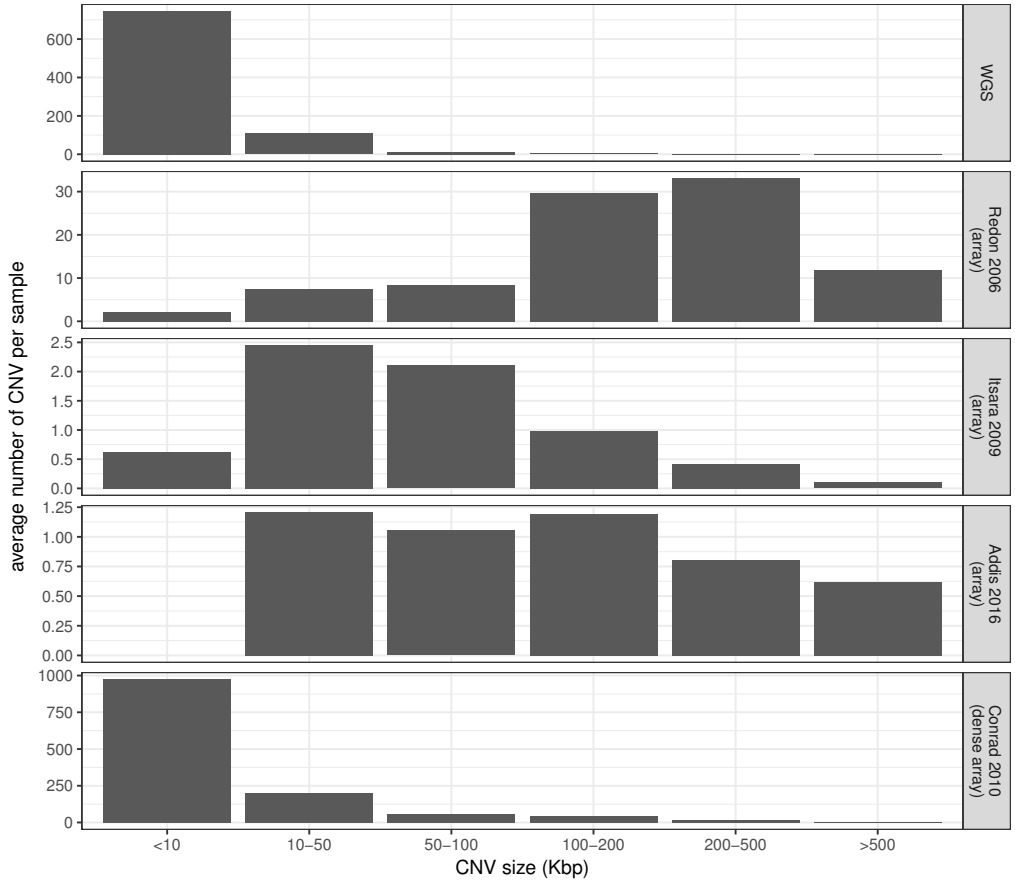


Figure S3.10: **CNV size in our cohort and four array-based studies.** The bars show the average number of CNVs called in a sample in the different cohorts. *Redon 2006*[106] and *Itsara 2009*[111] are population studies using technology similar to previous epilepsy studies. *Addis 2016*[125] is a recent study of large CNVs in absence epilepsy. *Conrad 2010*[24] is a population study that used multiple arrays to increase its resolution.

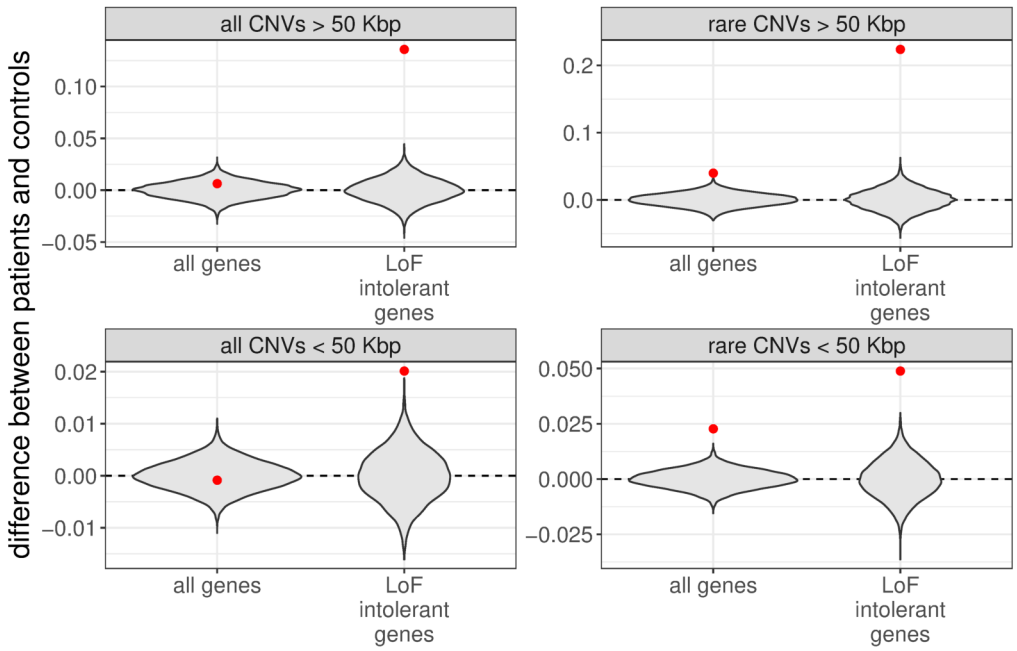


Figure S3.11: **Exonic enrichment significance.** The grey violin plot represents the difference in fold-enrichment between patients and controls across 10,000 permutations where the patient/control labels had been shuffled. The red point represents the observed difference between patients and controls (Fig. 3.2c).

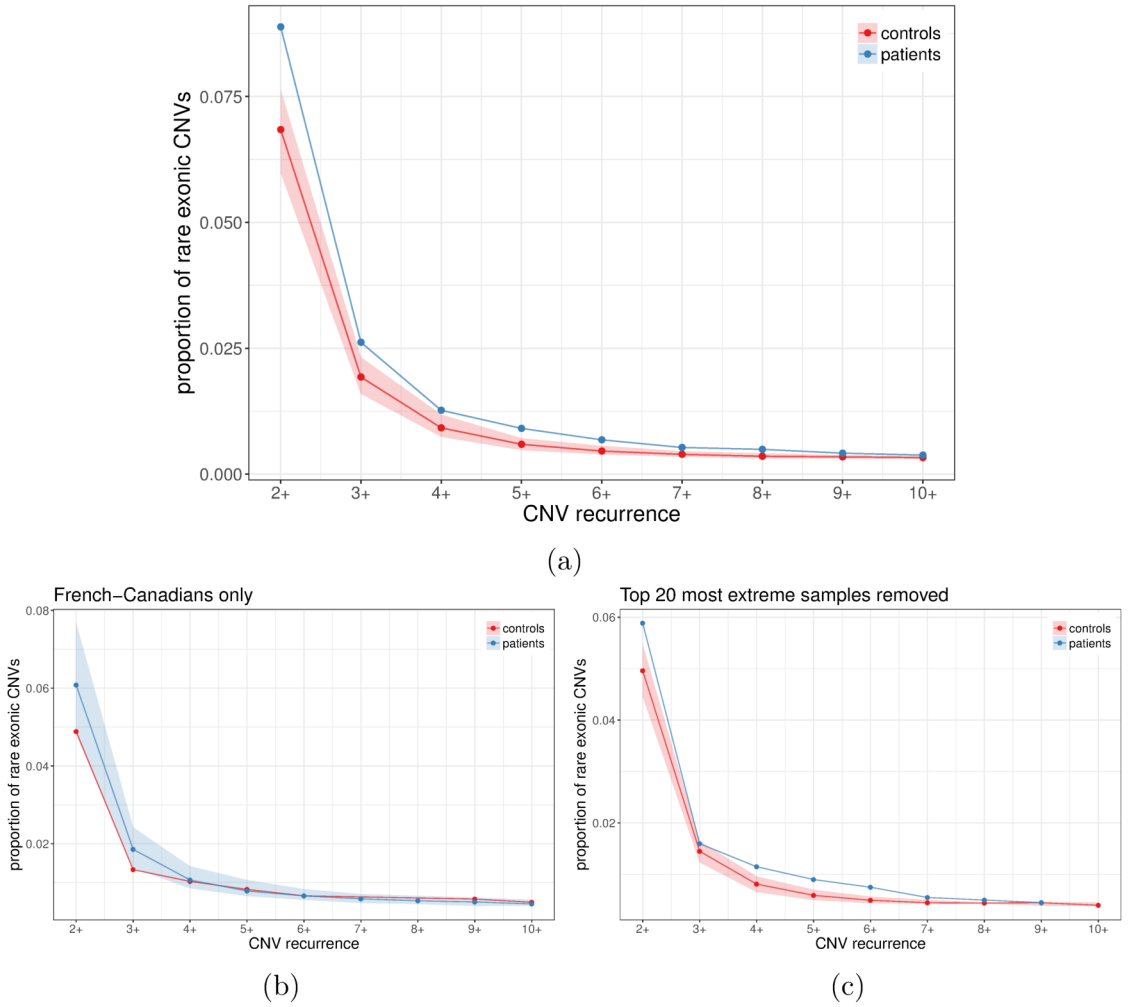
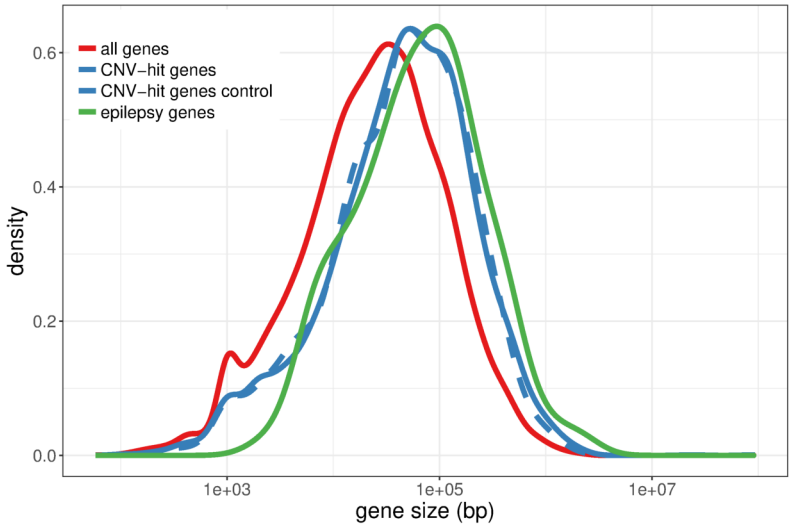
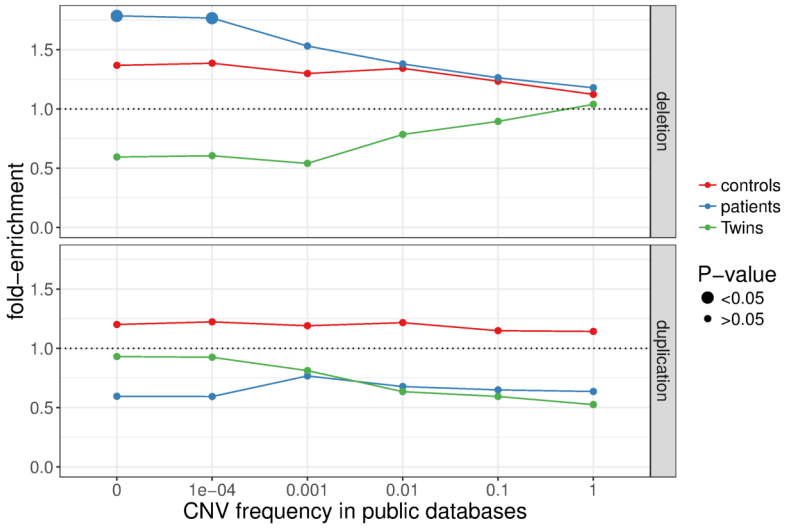


Figure S3.12: **Rare exonic CNVs are less private in the epilepsy cohort.** Proportion of rare exonic CNVs (y-axis) seen in X or more individuals (x-axis). The ribbon shows the 5%-95% confidence interval. In b), only French-Canadians individuals were analyzed and we down-sampled the epilepsy cohort to match the sample size of the French-Canadians controls. In c), the top 20 samples with the most non-private rare exonic SVs were removed.

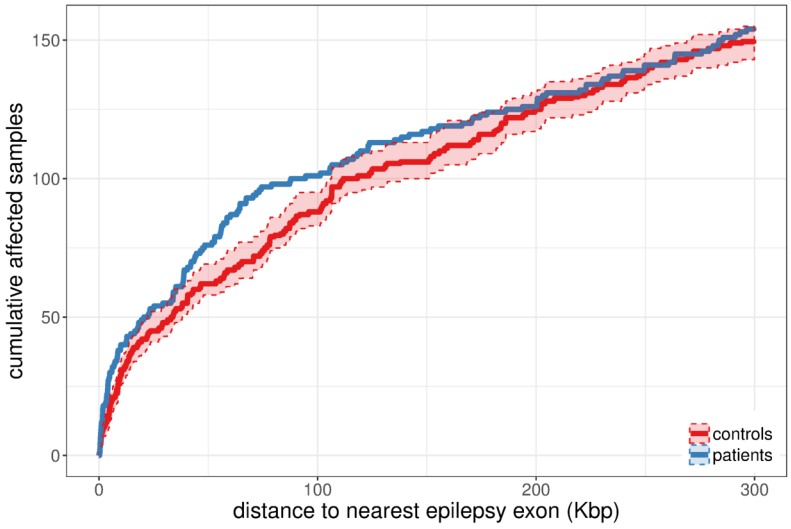


(a)

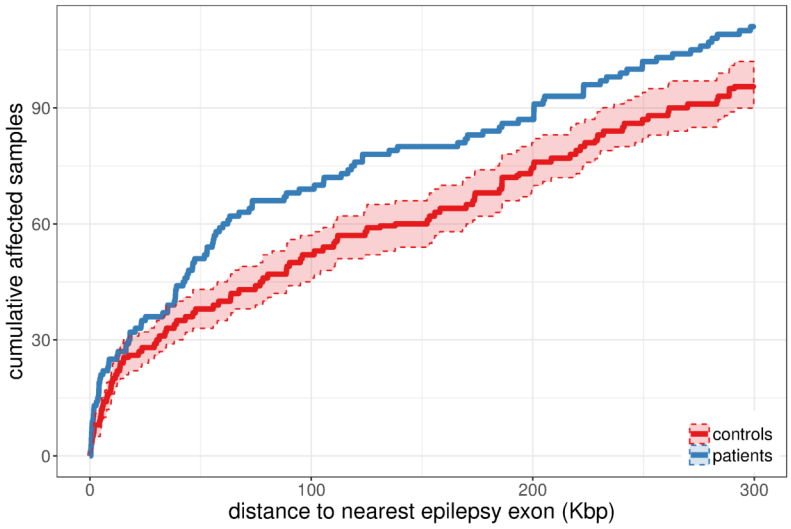


(b)

Figure S3.13: **Enrichment in epilepsy genes.** a) Epilepsy genes (red) are genes known to be associated with epilepsy. The control genes (dotted blue) are random genes selected so that the size distribution is similar to the sizes of genes hit by CNVs (plain blue). b) In three different datasets (color), genes hit by rare deletion (top) or duplications (bottom) at different frequency thresholds (x-axis) were tested for enrichment in epilepsy genes (y-axis, point-size).

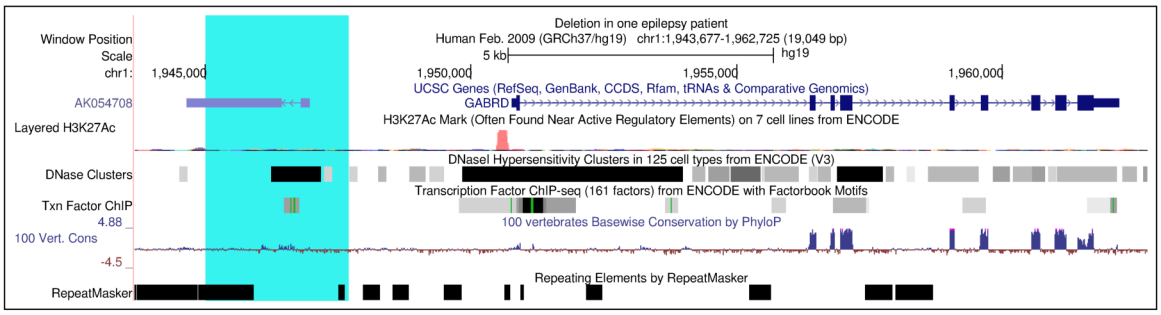


(a) Rare CNVs

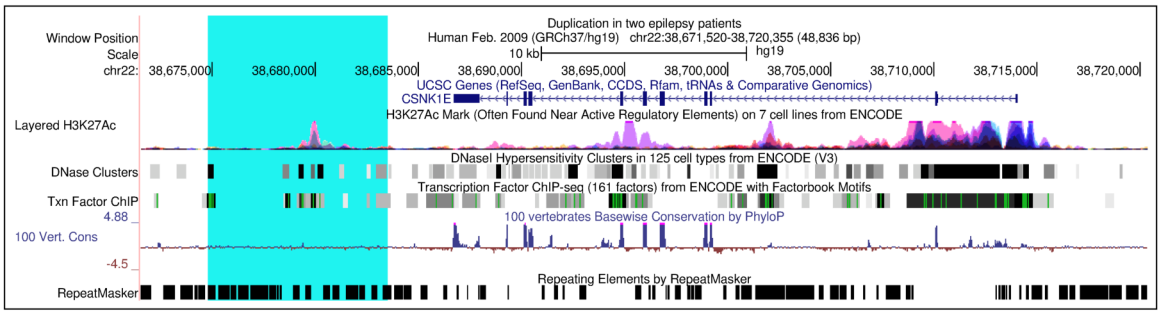


(b) Rare deletions

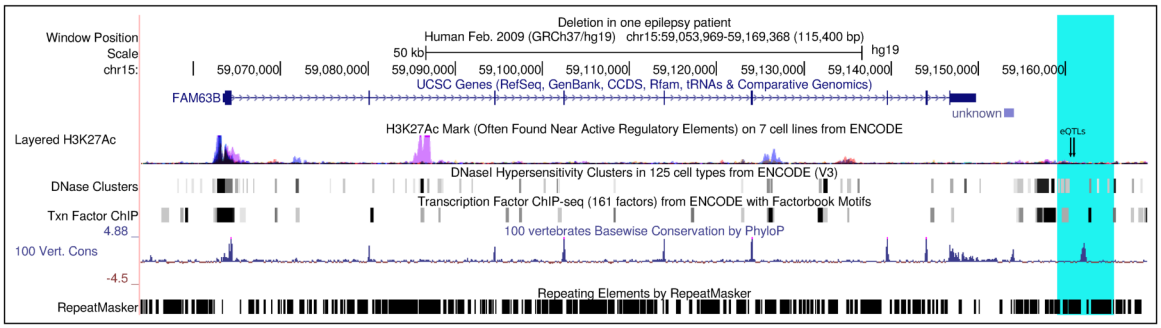
Figure S3.14: **Rare non-coding CNVs near epilepsy genes.** The graphs show the cumulative number of individuals (y-axis) with a rare non-coding variants located at X Kbp or less (x-axis) from the exonic sequence of a known epilepsy gene. The controls were down-sampled to the sample size of the epilepsy cohort. The ribbon shows the 5%/95% confidence interval. In a), deletions and duplications were considered; in b), only deletions were used.



(a)



(b)



(c)

Figure S3.15: **Non-coding CNVs with putative pathogenicity.** a) 2.7 Kbp deletion in an epilepsy patient, never seen in controls or CNV databases. Three other epilepsy patients have a rare non-coding deletions located at less than 200 Kbp from the *GABRD* gene. b) 8.8 Kbp duplication in two epilepsy patients, never seen in controls or CNV databases and overlapping a regulatory region associated with *CSNK1E*. c) 6.5 Kbp deletion of an ultra-conserved regions downstream of *FAM63B*. Two expression QTLs for this gene are highlighted with arrows.

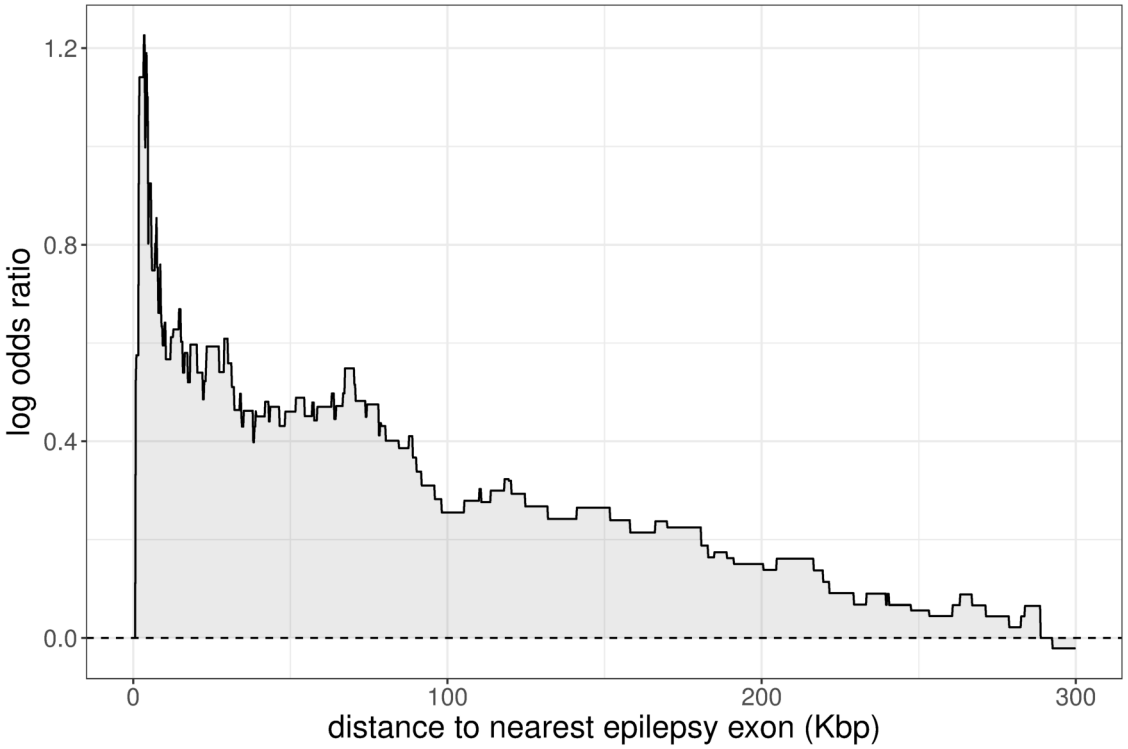


Figure S3.16: **The enrichment in rare non-coding CNVs overlapping functional regions increases close to epilepsy genes.** The graph shows the log odds ratio of having a rare non-coding CNV located at X Kbp or less (x-axis) from the exonic sequence of a known epilepsy gene. The y-axis shows the log odds ratio between epilepsy patients and controls. The controls were down-sampled to the sample size of the epilepsy cohort. We used CNVs overlapping regions functionally associated with the epilepsy gene (eQTL or promoter-associated DNase site).

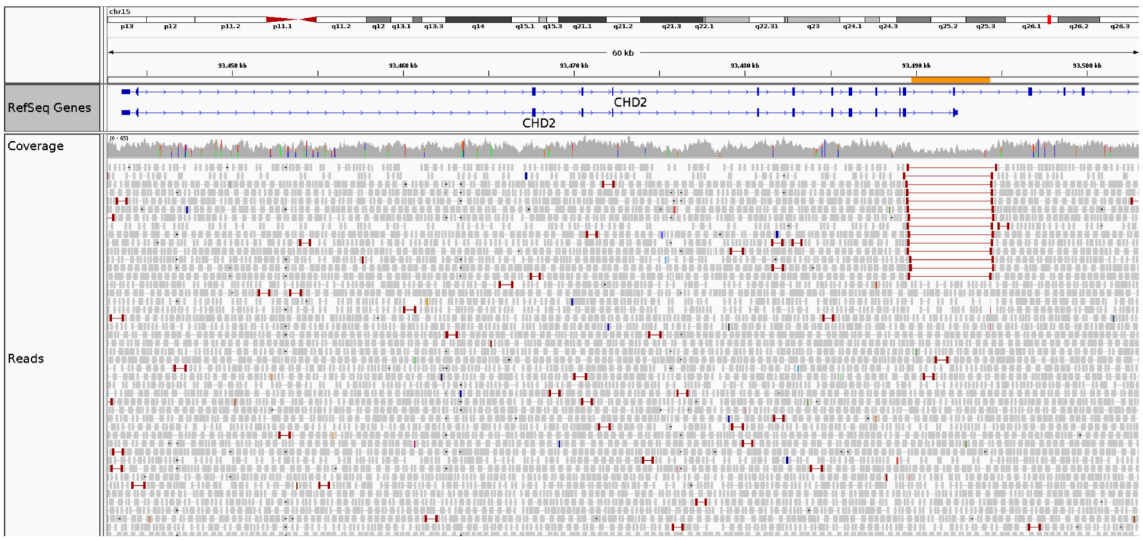
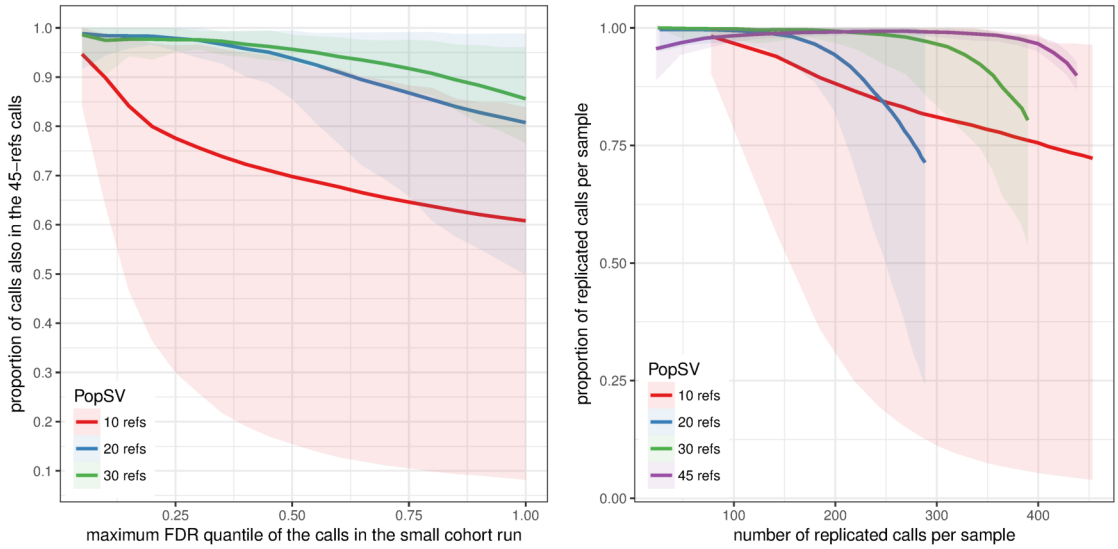


Figure S3.17: **Small deletion of exon 13 in *CHD2*.** Abnormal mapping of the read pairs highlighted in red support the deletion detected by PopSV using the read coverage. The deletion region is highlighted in orange.



(a) (b)

Figure S3.18: **Reference cohort size and CNV detection quality.** PopSV was run on the Twins study using 10, 20, 30 or 45 samples as reference (color). In a), the y-axis shows how many calls from the down-sampled run were found in the original 45-refs run. The x-axis represents the FDR threshold (lower threshold being more stringent). b) Replication in twins. For different cohort sizes and FDR thresholds, the number (x-axis) and proportion (y-axis) of calls replicated in the other twin is shown. In both graphs, the lines represents the median per sample and the ribbon the minimum/maximum values.

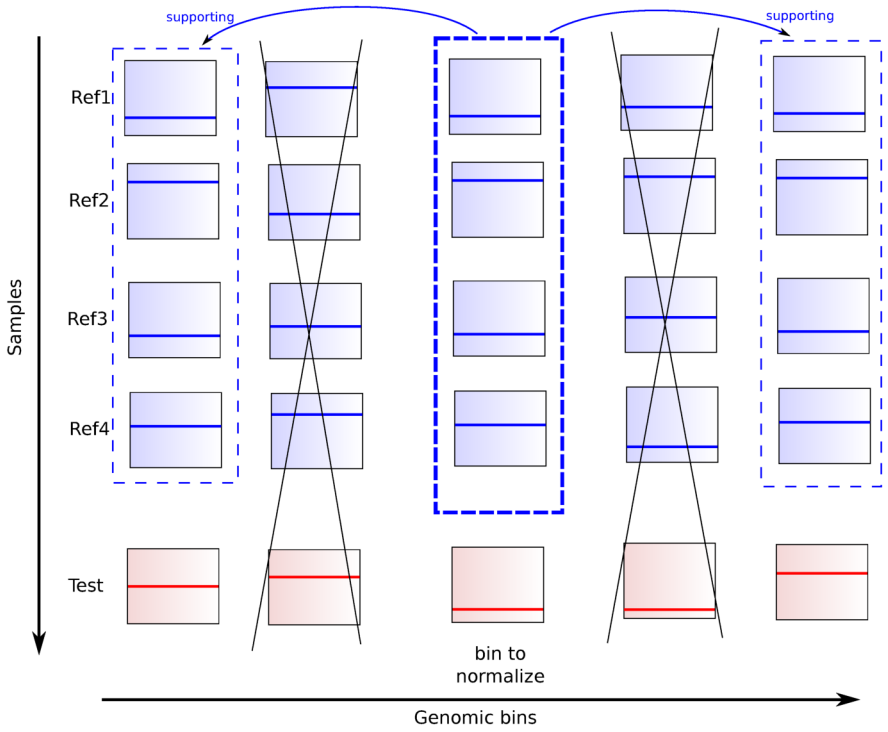


Figure S3.19: **Targeted normalization.** The coverage across the reference samples (blue) in the bin to normalize is used to find supporting bins across the genome. These supporting bins only are used to compute the normalization factor. The same supporting bins will be used to normalize the bin count in a test sample (red).

Supplementary Information

Epilepsy patients and sequencing

Ethics and patients recruitment CENet is a Genome Canada and Genome Québec funded initiative that aims to bring personalized medicine in the treatment of epilepsy. Patients were recruited through two main recruitment sites at the Centre Hospitalier Universitaire de Montréal (CHUM) and the Sick Kids Hospital in Toronto. This study was approved by the Research Ethics Board at the Sick Kids Hospital (REB number `1000033784`) and the ethics committee at the Centre Hospitalier Universitaire de Montréal (project number `2003-1394,ND02.058-BSP(CA)`). Before their inclusion in this study, patients 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 (±6.7). Supplementary Table S3.3 presents a detailed description of the clinical features for the various individuals recruited in this study. DNA was extracted from blood DNA exclusively. 301 unrelated healthy parents of other probands from CENet were also included in this study and used as controls.

Libraries preparation and sequencing gDNA was cleaned up using ZR-96 DNA Clean & Concentrator^{TM}-5 Kit (Zymo) prior to being quantified using the Quant-iT^{TM} PicoGreen® dsDNA Assay Kit (Life Technologies) and its integrity assessed on agarose gels. Libraries were generated using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina) according to the manufacturer's recommendations. Libraries were quantified using the Quant-iT^{TM} PicoGreen® dsDNA Assay Kit (Life Technologies) and the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems). Average size fragment was determined using a LabChip GX (PerkinElmer) instrument.

The libraries were first denatured in 0.05N NaOH and then were diluted to 8pM using HT1 buffer. The clustering was done on a Illumina cBot and the flowcell was run on a HiSeq 2500 for 2x125 cycles (paired-end mode) using v4 chemistry and following the manufacturer's instructions. A phiX library was used as a control and mixed with libraries at 1% level. The Illumina control software was HCS 2.2.58, the real-time analysis program was RTA v. 1.18.64. Program bcl2fastq v1.8.4 was then used to demultiplex samples and generate fastq reads. The average coverage was 37.6x ± 5.6x. The filtered reads were aligned to reference Homo_sapiens assembly b37. Each readset was aligned using BWA[203] which creates a Binary Alignment Map file (.bam). Then, all readset BAM files from the same sample are merged into a single global BAM file using Picard. Insertion and deletion realignment was performed on regions where multiple base mismatches were preferred over INDELs by the aligner since it appears to be less costly for the algorithm. Such regions were found to introduce false positive variant calls which may be filtered out by realigning those regions properly. Once local regions were realigned, the read mate coordinates of the aligned reads needed to be recalculated since the reads are realigned at positions that differ from their original alignment. Fixing the read mate positions is performed using Picard. Aligned reads were marked as duplicates if they have the same 5' alignment positions (for both mates in the case of paired-end reads). All but the best pair (based on alignment score) were marked as a duplicate in the `.bam`

file. Duplicates reads were excluded in the subsequent analysis. Marking duplicates was performed using Picard.

Testing for technical bias in WGS

To investigate the bias in read depth (RD), we first fragmented the genome in non-overlapping bins of 5 Kbp. The number of properly mapped reads was used as RD measure, defined as read pairs with correct orientation and insert size, and a mapping quality of 30 (Phred score) or more. In each sample, GC bias was corrected by fitting a Loess model between the bin's 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 when no bias is present. 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 is 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 100/N % of the genome (with N the number of samples). If an experiment was more affected by technical bias, it would be more often extreme. The same analysis was performed monitoring the lowest coverage.

The same analysis was ran after correcting the coverage in the Twin dataset using the QDNAseq pipeline[190]. The reads were counted in 5 Kbp bins using the function `binReadCounts`. GC bias and mappability were corrected using the following functions (with default parameters): `applyFilters`, `estimateCorrection`, `correctBins`, `normalizeBins`, `smoothOutlierBins`.

PopSV

Binning and coverage measure The genome is fragmented in non-overlapping consecutive bins of fixed size (5 Kbp). 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. Read pairs with correct orientation and insert size, and a mapping quality of 30 (Phred score) or more are considered properly mapped. 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 the epilepsy study and the Twins dataset we used all the samples as reference. In the renal cancer dataset we used the normal samples as reference. For each dataset, a Principal Component Analysis (PCA) was performed across samples on the counts normalized globally (median/variance adjusted). The resulting first two principal components are used to verify the homogeneity of the reference samples. In the presence of extreme outliers or clear sub-groups, a more cautious analysis would be recommended. For example, outliers can remain in the set of reference samples but flagged as they might potentially harbor more false calls later. Independent analysis in each of the identified sub-group is also a solution, especially when the same samples are to be used as reference. Although our three datasets showed different levels of homogeneity, we did not need to exclude samples or split the analysis. The effect of weak outlier samples was either corrected by the normalization step or integrated in the population-view. Moreover, the principal components were used to select one control sample from the final set of reference samples. This sample is used in the normalization step as a baseline to normalize other samples against. We picked the sample closest to the centroid of the reference samples in the Principal Component space.

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 exactly the same in each sample. After witnessing the presence of uncharacterized sample-specific variation, we implemented a more suited normalization. Targeted normalization uses information across the set of reference samples to identify similar bins across the genome and normalize their counts separately (Fig. S3.19). 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. Hence, other regions with similar bias patterns are used to correct for 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, but a subset of regions with specific bias might not be corrected. 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 regions of 100 consecutive bins, we counted how many times the RD in a sample was higher than the median across sample. If the ranks 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 region has non-random ranks. The resulting QC metric is simply the proportion of regions 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 does not 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-score represents how different the read count in the tested sample is from the reference samples. It is simply: $z = \frac{BC_t^b - mean(BC_{ref}^b)}{sd(BC_{ref}^b)}$ 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 that 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, later used to estimate the null distribution variance. Here the null Z-score distribution is assumed to be centered on 0 and 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. Indeed, once the plateau is reached, additional trimming does not 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 does not affect the P-value estimation. Multiple testing correction is performed after pooling all the P-values.

Segmentation, copy number estimation and other metrics Following the significance estimation, consecutive bins with abnormal coverage are merged into a call. Consecutive or nearby abnormal bins (e.g. one bin size apart) are merged into one variant if in the same direction (deletion or duplication). In PopSV's R package, the P-values can also be segmented using circular binary segmentation[204].

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. The copy number is 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 be 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 benchmark, we used the entire set of calls.

Validation and benchmark of PopSV

We compared PopSV to FREEC[40], CNVnator[41] and cn.MOPS[42], three 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 while CNVnator uses a mean-shift technique inspired from image processing. cn.MOPS considers simultaneously several samples and detects copy number variation using a Poisson model and a Bayesian approach. We also ran LUMPY[67] which uses an orthogonal mapping signal: the insert size, orientation and split mapping of paired reads.

FREEC and CNVnator were 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 across the entire genome, the minimum telocentromeric distance was set to 0. The remaining parameters were set to default. Of note an additional run with slightly looser parameter ('breakPointThreshold=0.6') was performed to get a larger set of calls used in some parts of the in silico validation analysis to deal with borderline significant calls. CNVnator also corrects internally for GC bias. We used default parameters. For the analysis using higher confidence calls, we used calls with either 'eval1' or 'eval2' lower than 10-5 (instead of the default 0.05). cn.MOPS 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. For LUMPY, the discordant reads were extracted from the BAMs using the recommended commands. Split-reads were obtained by running YAHA[66] with default parameters. All the CNVs (deletions and duplications) larger than 300 bp were kept for the upcoming analysis. Calls with 5 or more supporting reads were considered high-confidence.

First, we compared the frequency at which a region is affected by a CNV using the calls from the different methods. In order to investigate how many systematic calls are present in a typical run, we compare the frequency distributions on average per sample. In figure S3.4, the bars represents the average proportion of a sample's calls in each frequency range.

Then, the samples were clustered using the CNV calls. The distance between two samples A and B is defined as : $1 - 2\frac{|VAB|}{(|VA|+|VB|)}$ where VA represents the variants found in sample A, VAB 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 called 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 ¡10 reads).

To assess the performance of each method, we measured the number of CNVs identified in each twin that were also found in the matching 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 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. Even if we removed systematic calls, the most frequent calls in the cohort are more likely to look replicated by chance, compared to rare calls. To normalize for this effect, we use the frequency distribution to compute the number of replicated 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 replicated calls by chance is subtracted to the original number of replicated calls to give an adjusted measure of sensitivity. Although we do not know the true number of variant, this number of replicated calls is used to compare the different methods. When possible, the low-quality calls were also gradually filtered to explore the effect on the replication metrics. For CNVnator, we used the minimum of the eval1 and eval2 columns, with lower values corresponding to higher quality calls. For LUMPY, the number of supporting reads was used. For PopSV, we filtered calls based on adjusted P-values.

In addition to their replication, we compared which regions were called by several methods. For each of the calls found in less than 50% of the samples, we overlapped the region with calls from other methods in the same sample. If calls from another method overlapped we considered the call shared and saved which methods shared the call. To focus on on high quality calls we considered calls found by at least two methods and computed the proportion of calls from one method found by each of the

other methods. This metric captures how much each method recovers high-quality calls from a second method.

Concordance between different bin sizes We compared calls using small bins (500 bp) and calls using larger bins (5 Kbp). In theory, calls from the 5 Kbp analysis should be supported by many 500 bp calls. We also expect large stretches of 500 bp calls to be detected in the 5 Kbp analysis. This comparison is informative as it explores the quality of the calls, the size of detectable events and the resolution for different bin sizes. 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. Overall, we find that 5 Kbp calls are well supported by 500 bp calls, with only 14% of the 5 Kbp bins not supported by any 500 bp bin (Fig. S3.9a). 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 to test if the lack of support is due to lower confidence or real discordancy 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. Even for these unsupported 5 Kbp calls, we find that the 500 bp bins RD was consistently enriched (or depleted) although not enough to be called with confidence (Fig. S3.9b and S3.9c). 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 Kbp 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? We find that the concordance gradually increases until the 500 bp calls reach 5 Kbp in size where the concordance rises to nearly 100% (Fig. S3.9d). 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 Kbp calls and they likely corresponds to fragmented larger calls. Considering the trade-off between bin size and noise, this suggests running PopSV with a few bin sizes to better capture variants of different sizes.

CNV detection in the CENet cohorts

CNVs were called using PopSV using 5 Kbp bins and all the samples from both the epilepsy and control cohorts as reference. We annotated the frequency of the CNVs using germline CNV calls from the Twin and cancer datasets (internal database) as well as four public CNV databases:

- CNVs from Phase 1 of the 1000 Genomes Project as identified by GenomeSTRiP[33].
- SV from the 1000 Genomes Project phase 3[5].
- Genome of Netherlands[107].
- CNVs from the Simons Genome Diversity Project[192].

CNVs were annotated with the maximum frequency in the databases. For each CNV to annotate, any overlapping CNV in the CNV databases were considered. This is a stringent criterion that ensures that the entire regions of a rare CNV, for example, is never affected by common CNVs in the databases. Hence, a rare CNV is defined as present in less than 1% of the samples in each of the five CNV databases.

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.

CNV enrichment in exonic region and around epilepsy genes

Enrichment in exons 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 a hundred times. For each sub-sampling and each cohort, control regions are selected to fit the size distribution of the CNV catalog and the overlap with centromere, telomeres and assembly gaps (details in the next section).

Then, we computed the proportion of CNV and control regions that overlap an exon. The fold-enrichment is the ratio of these proportions and represents how much more/less of the CNVs overlap an exon compared to the control regions. The boxplot in Fig. 3.2c shows the distribution of the 100 sub-sampling in each cohort.

To test if the difference observed between the cohort is significant, the *cohort* labels were permuted 10,000 times and the difference in median across the 100 sub-sampling was saved. The resulting P-value was computed as $\frac{1+d}{1+N}$ where d is the number of times the permuted difference was greater or equal to the observed difference, and N is the number of permutations.

The same analysis was repeated for exons from genes with a probability of loss-of-function intolerance[193] higher than 0.9. These genes were called *LoF intolerant genes* in Fig. 3.2c. Small (¡ 50Kbp) and large (¿50 Kbp) CNVs were analyzed separately. The analysis was repeated using rare CNVs only.

Selecting control regions The control regions must have the same size distribution as the regions they are derived from (e.g. CNVs in a CNV catalog). We also controlled for the overlap with centromere, telomeres and assembly gaps (CTGs) to avoid selecting control regions in assembly gaps where no CNV or annotation is available. To select control regions, thousands of bases were first randomly chosen in the genome. The distance between each base and the nearest CTG was then computed. At this point, selecting 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:

$$\left\{b, O_{CTG}(d^b_{CTG} - \frac{S_r}{2}) < 0\right\} \tag{6.1}$$

with O_{CTG} equals 1 if the original region overlaps with a CTG, -1 if not; d^b_{CTG} is the distance between base b and the nearest CTG; and S_r is the size of the original

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

Recurrence of rare exonic CNVs In each cohort, we retrieved the CNV catalog of rare (¡1% in all 5 public datasets) exonic CNVs. We annotated each CNV with its recurrence in the cohort. We then evaluated the proportion of the CNVs in the catalog that are private (i.e. seen in only one sample), or seen in X samples or more. This cumulative proportion of CNVs is shown in Fig. S3.12a. The control cohort was down-sampled a thousand times to the same sample size as the epilepsy cohort. These down-sampling provided a confidence interval (ribbon in Fig. S3.12a) and an empirical P-value.

We performed the same analysis after removing the top 20 samples with the most non-private rare exonic CNVs (Fig. S3.12b). With this analysis, we tried to remove the potential effect of a few extreme samples.

We also repeated the analysis using only French-Canadians individuals, to ensure that the observed differences are not caused by rare population-specific variants (Fig. S3.12b).

CNVs and epilepsy genes We used the list of genes associated with epilepsy from the EpilepsyGene resource[118] 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. We noticed that the epilepsy genes tend to be large, and genes hit by CNVs also (Fig. S3.13a). This could lead to a spurious association so we also performed a permutation approach that controls for the size of the 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. The permutation P-value was computed as $\frac{1+d}{1+N}$ where d is the number of times the number of epilepsy genes in the random set of genes was greater or equal to the one in genes hit by CNVs, and N is the number of permutations. Using this sampling approach we tested different sets of CNVs: deletion or duplications of different frequencies in the epilepsy cohort, control individuals and samples from the twin study.

To investigate rare non-coding CNV close to known epilepsy genes, we counted how many patients have such a CNV at different distance thresholds. For example, how many patients had a rare non-coding CNV at 10 Kbp of an epilepsy gene's exon or closer. We compared this cumulative distribution to the control cohort, after down-sampling it to the sample size of the epilepsy cohort. Down-sampling was also used to produce a confidence interval, represented by the ribbon in Fig. 3.3c). This analysis was repeated using deletions only. Each epilepsy gene was also tested for an excess of rare non-coding deletions in patients versus controls using a Fisher test.

In order to retrieve non-coding CNV that might have a functional impact, we downloaded eQTLs associated with the epilepsy genes, as well as DNase 1 hypersensitive sites associated with the promoter of epilepsy genes. The eQTLs are provided by the GTEx project[195]. Pairs of associated DNase 1 hypersensitive sites and associated genes[196] were downloaded at http://www.uwencode.org/proj/Science_Maurano_Humbert_et_al/data/genomewideCorrs_above0.7_promoterPlusMinus500kb_

withGeneNames_35celltypeCategories.bed8.gz.

A Kolmogorov-Smirnov test was used to compare the distance distributions in epilepsy patients versus controls. We also computed the odds ratio of having such a CNV for different distance thresholds between epilepsy patients and controls. For a distance d, we computed:

$$OR = \frac{S^{CNV}_{patient}}{S^{CNV}_{control}} / \frac{S^{noCNV}_{patient}}{S^{noCNV}_{control}}$$

where $S^{CNV}_{patient}$ is the number of patients with a rare non-coding CNV overlapping a functional region and located at d bp or less from the exon of a known epilepsy gene.

Appendix D

Supplementary material for chapter 4 and its corresponding manuscript: **Human copy number variants are enriched in regions of low mappability**.

Supplementary Tables

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

Table S4.1: **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.

Chr.	Start	End	CN	PCR product size	PCR product size when deletion	Validated	Gel	Sanger Sequencing
14	40098378	40100213	0	2586	751	Yes	Different bands	Yes: confirmed
5	85559864	85564846	1.05	5690	708	Yes	Different bands	Yes: confirmed
6	14299746	14299801	0.79	755	700	Yes	Double bands	No
7	153000055	153000246	1.76	1137	946	Yes	Double bands	Yes: confirmed
4	96401034	96401460	1.13	745	319	Yes	Double bands	No
16	34230052	34230512	1	1139	679	Yes	Double bands	No
16	8688137	8689592	1.02	2121	666	Yes	Double bands	Yes: confirmed
2	12018994	12022932	1.02	4291	353	Yes	Double bands	Yes: confirmed
3	121051576	121060845	1.14	9485	216	Yes	Double bands	No
3	54433855	54433912	0	952	895	Yes	One band	Yes: insertion
2	151031059	151038246	1.11	7485	298	Yes	Small band only	No
9	45462450	45462522	1.1	530	458	No	One band	No
7	63233184	63233261	1.33	390	313	No	One band	Yes: nothing
9	106371251	106371330	1.28	484	405	No	One band	No
16	20466400	20466487	1.27	393	306	No	One band	No
5	85559864	85564842	0.78	5690	712	No	One band	No
10	65703860	65708900	1.64	5430	390	No	One band	No
7	159117395	159122761	1.09	5909	543	No	One band	No
2	83066824	83068234	0.57	2097	687	NA	No amplification	No
13	35996202	35996254	1.13	546	494	NA	Non-specific	No
4	159799983	159801372	1.03	2313	924	NA	Non-specific	Yes: not clear
7	52963172	52964911	1.48	2316	577	NA	Non-specific	No
10	69323932	69326507	1.62	2795	220	NA	Non-specific	Yes: not clear
6	58618198	58624080	1.04	6518	636	NA	Non-specific	No

Table S4.2: **Experimental validation in low-coverage regions.** The result of the PCR validation was either concordant with PopSV call (Yes), discordant (No) or inconclusive (NA). In some cases, Sanger sequencing was performed. The *CN* column is the estimated copy-number of the deleted allele.

Homozygous deletion

deletion support	reference support	number of calls
0	0	11
0	1	1
1	0	1
2	0	12

Heterozygous deletion

deletion support	reference support	number of calls
0	0	18
0	1	10
0	2	7
1	0	6
1	1	4
1	2	4
2	0	10
2	1	3

Table S4.3: **Investigating low-mappability deletion calls with two CEPH12878 assemblies.** The first two columns represent the number of assemblies (0, 1 or 2) supporting the deleted allele or the reference allele. The third column shows the number of PopSV calls in each category.

CNV catalog	Samples	Variants			Avg Size (Kbp)	Proportion	Affected genome (Mbp)	
		Total	Per sample			<3 Kbp	Total	Per sample
			WG	*ELC*				
1000GP	2,504	41,979	1,024.44	2.22	6.00	0.68	580.03	6.14
deletion		36,102	975.32	2.21	4.67	0.72	342.97	4.56
duplication		8,503	48.26	0.00	32.54	0.00	331.48	1.57
GoNL	750	9,592	1,048.14	0.63	2.93	0.81	65.30	3.07
deletion		9,009	1,013.35	0.63	2.36	0.82	34.79	2.39
duplication		528	21.11	0.00	29.19	0.15	30.63	0.62
Handsaker 2015 (Genome STRiP)	847	8,657	212.03	1.88	27.80	0.00	196.57	5.89
deletion		5,961	145.78	0.56	21.64	0.00	108.03	3.15
duplication		3,469	66.26	1.32	41.35	0.00	118.28	2.74
Chiang 2017 (Genome STRiP)	148	7,932	828.49	9.23	6.35	0.42	73.20	5.26

Table S4.4: **Properties of events in public CNV catalogs.** Deletions, duplications and CNVs from four public catalogs. Variants with high frequency ($> 80\%$), variants on the chromosome X, and variants smaller than 300 bp were removed in order to compare with PopSV's numbers (Table 4.1). WG: whole genome; ELC: extremely low-coverage regions. The *Total* number of variants is the total number after collapsing recurrent variants. *Affected genome* represents the amount of the reference genome that overlaps at least one CNV.

Novel region	OMIM gene
1:25730001-25736000	RHCE
1:161640001-161645000	FCGR2B
1:207705001-207715000	CR1
1:207725001-207745000	CR1
5:68845001-68886000	OCLN
5:69360001-69365000	SMN2
5:69373001-69374000	SMN2
5:70165001-70222000	SMN1
5:70242001-70242500	SMN1
5:70246501-70258000	SMN1
6:29905001-29910000	HLA-A
6:31960001-31975000	C4A
6:31995001-31995500	C4B
6:32522001-32560000	HLA-DRB1
6:32590001-32602000	HLA-DQA1
6:32628501-32629000	HLA-DQB1
6:32630001-32634000	HLA-DQB1
7:39052001-39055500	POU6F2
7:74195001-74200000	NCF1

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

Table S4.5: **OMIM genes overlapping novel CNV regions of low-mappability** Novel CNV regions are polymorphic in more than 1% of the individuals across the three cohorts but absent from the 1000GP SV catalog[5]. OMIM genes are genes associated with a disease or phenotype in the OMIM Morbid Map (Online Mendelian Inheritance in Man; http://omim.org/).

Supplementary Figures

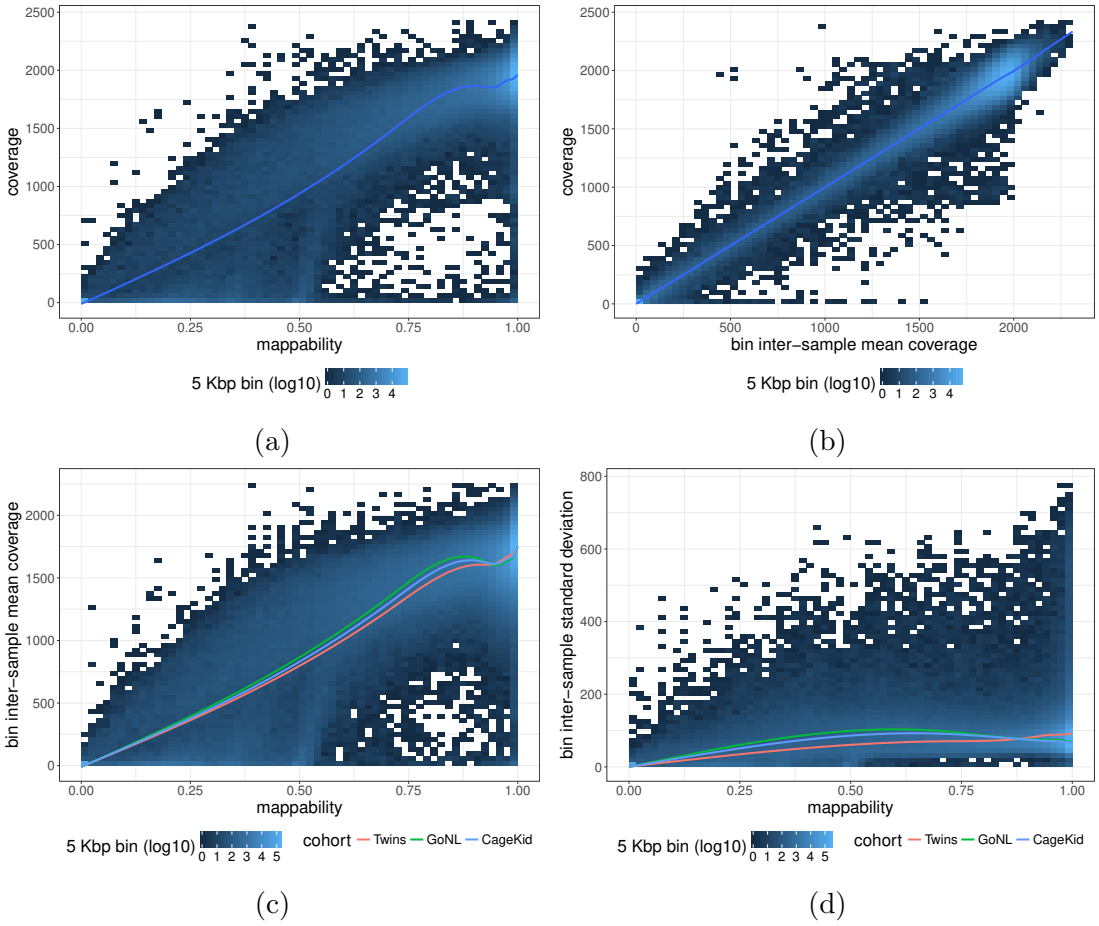
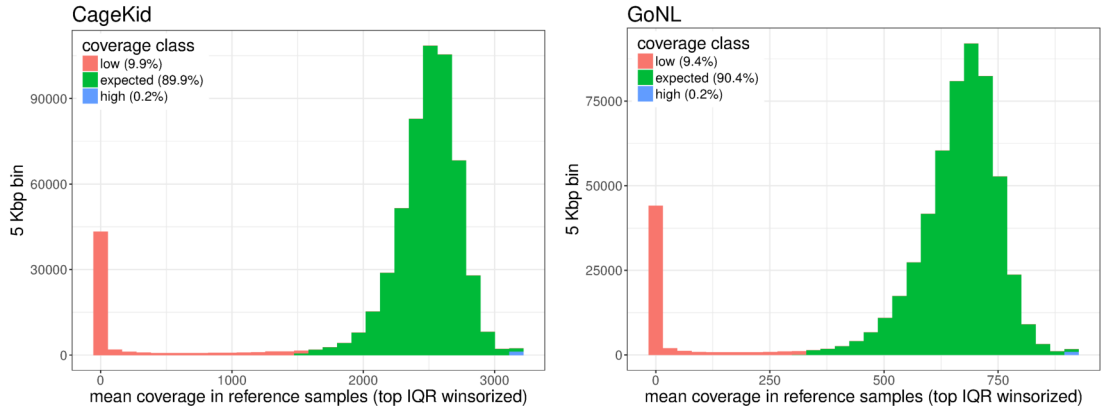


Figure S4.1: **Coverage, mappability and population-based measures.** a-b) Read coverage in a sample (y-axis) versus mappability (a) or the inter-sample average coverage (b). c-d) Inter-sample mean (c) and standard deviation (d) were fitted against the mappability in each cohort separately. The tiles represent all cohorts pooled together.



(a) (b)

Figure S4.2: **Average coverage in reference samples in the CageKid (a) and GoNL (b) datasets.**

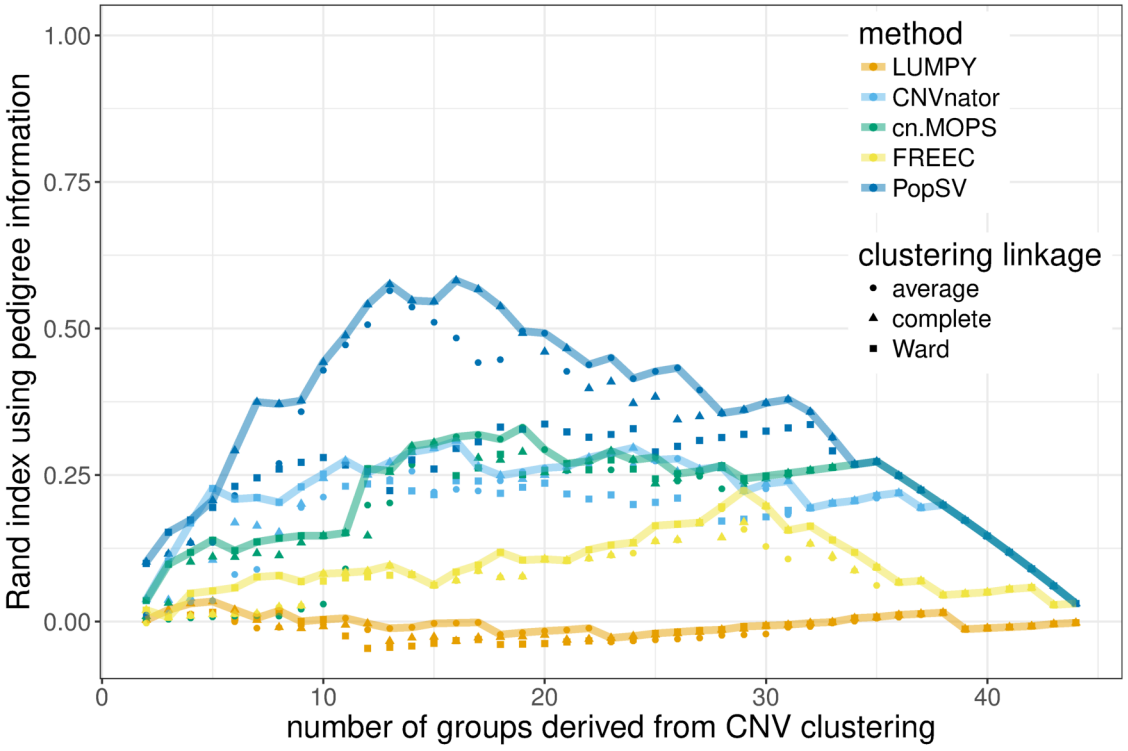


Figure S4.3: **Rand index between the pedigree information and the dendogram from CNV calls in low-coverage regions.** The dendogram for CNV-based clustering was cut at different levels (x-axis) and the groups compared to the pedigree (family-level) with the Rand index (y-axis). For each method, the line highlights the best performance across three linkage criteria.

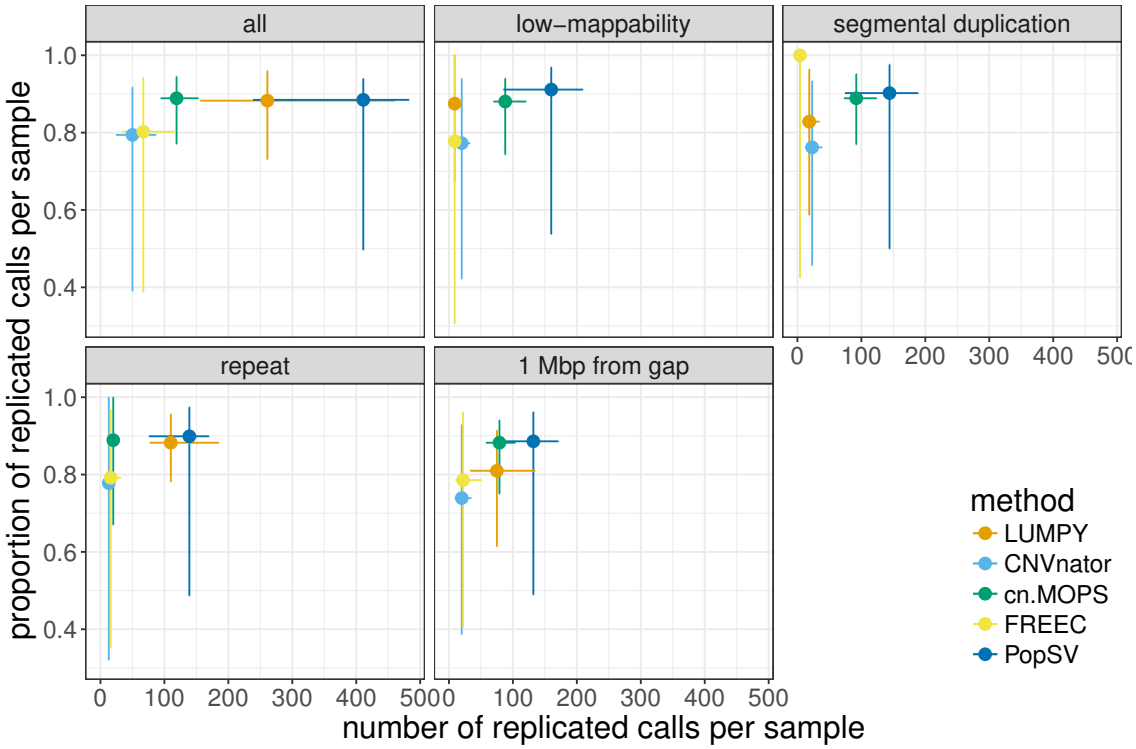


Figure S4.4: **PopSV's performance in low-mappability regions in CageKid dataset.** Proportion and number of calls replicated in the paired tumor. The point shows the median value per sample, the error bars the 95% confidence interval.

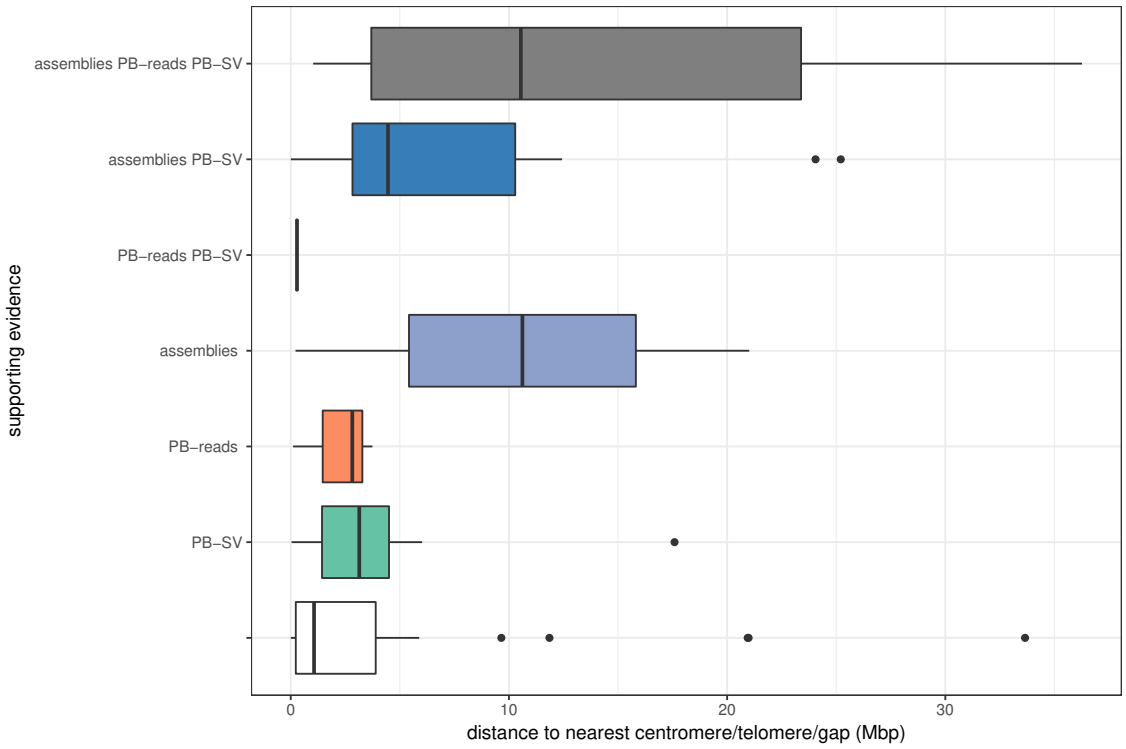


Figure S4.5: **Distance to assembly gaps and supporting evidence from long-read sequencing in CEPH12878.** Deletions in low-mappability regions were grouped by their supporting evidence (y-axis and colors). *assemblies*: deletion observed in at least one of the two public assemblies. *PB-SV*: overlap with a structural variant called from the PacBio reads[59]. *PB-reads*: deletion observed in the local assembly or consensus of the PacBio reads. Variants with no support are represented by the white boxplot.

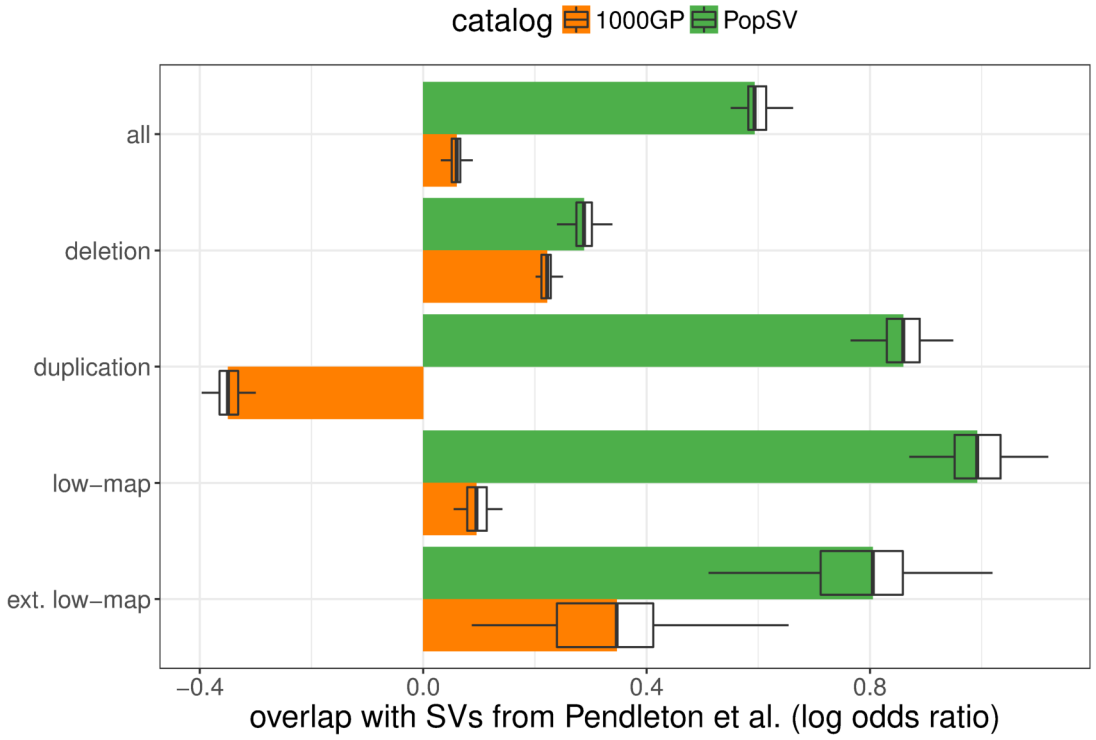


Figure S4.6: **Overlap between PopSV catalog and calls from Pendleton et al..** Recurrent calls were collapsed in each catalog (i.e PopSV and the 1000 Genomes Project (1000GP)). The proportion of the collapsed calls overlapping calls from Pendleton et al.[59] was computed. The fold-enrichment is produced by drawing control regions with similar size distribution as Pendleton's calls. *low-map*: calls in low-mappability regions; *ext. low-map*: calls in extremely low-mappability regions.

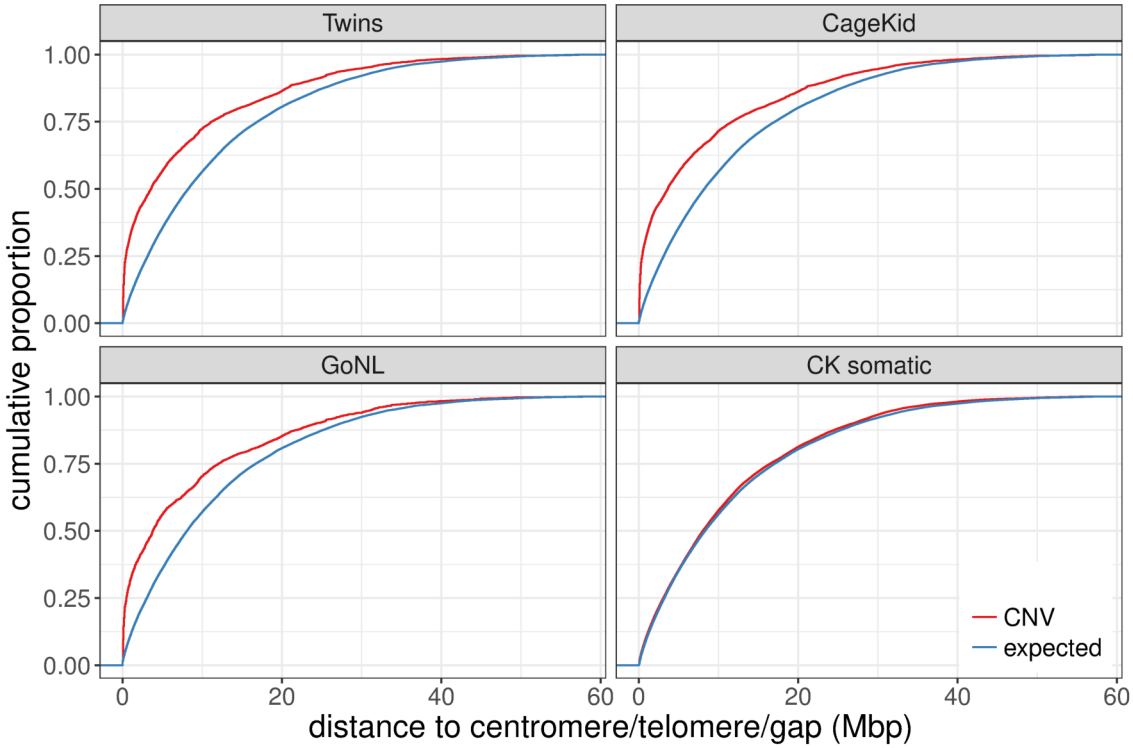


Figure S4.7: **Distance to a centromere, telomere or assembly gap.** The y-axis represents the cumulative proportion of the affected genome. The *expected* curve is computed from uniformly distributed genomic regions with matched size.

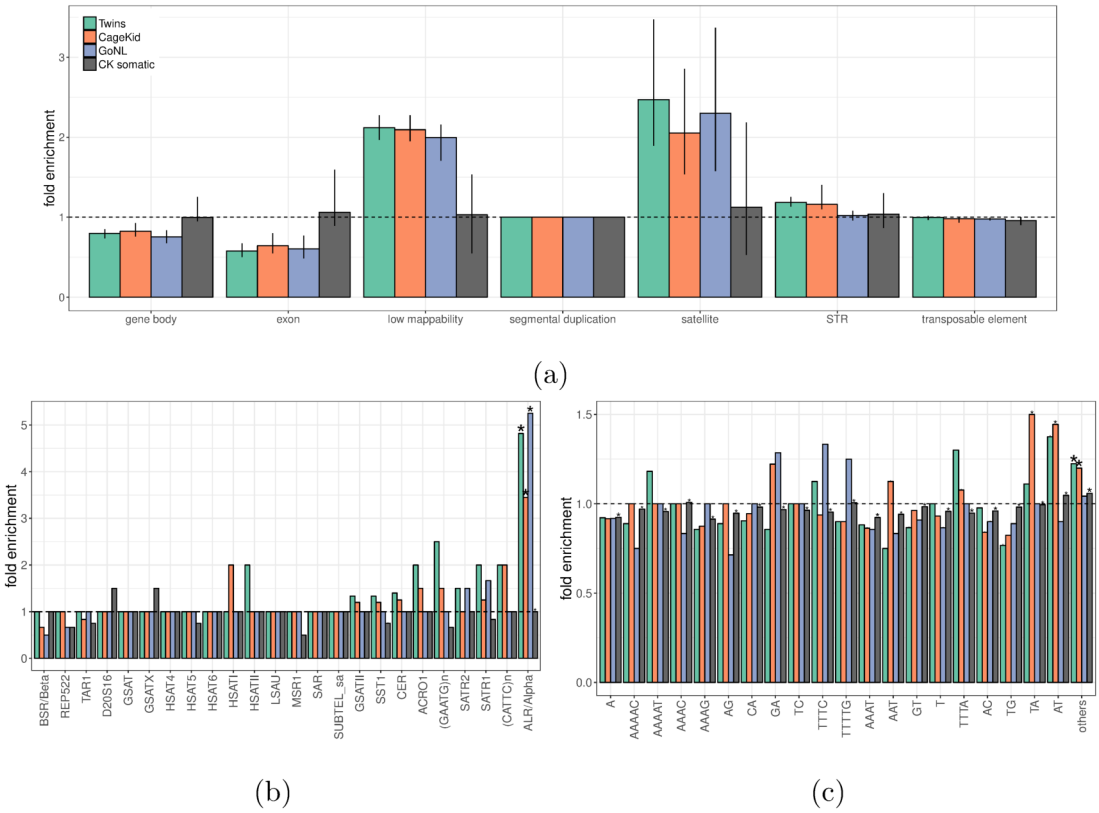
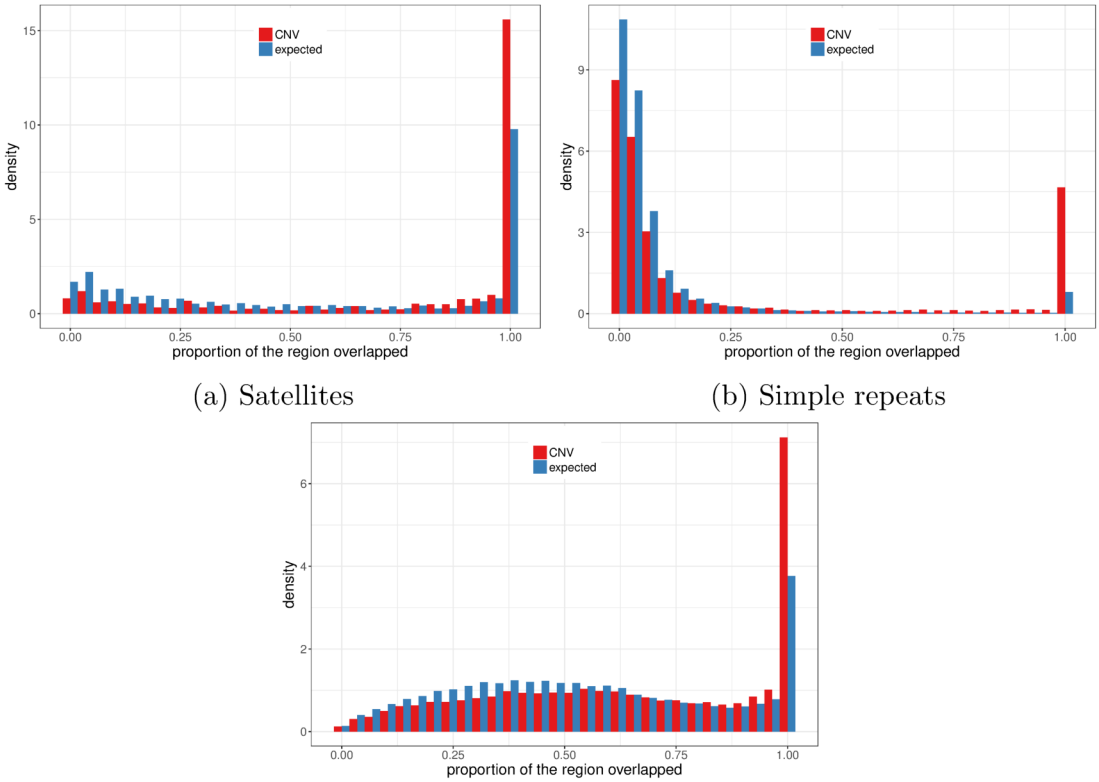


Figure S4.8: **CNVs enrichment after controlling for segmental duplication overlap and distance to CTG.** Enrichment of CNVs in a) different genomic features, b) satellite families and c) simple repeats in the different cohorts (colors). Bars show the median fold enrichment across samples compared to control regions. The star represents significant enrichment from the logistic regression.



(a) Satellites

(b) Simple repeats

(c) Transposable elements

Figure S4.9: **Overlap between CNVs and repeats.** The histograms represent the proportion of the CNV region that overlaps a) a satellite, b) a simple repeat or c) a transposable element, when they do overlap. The *expected* distribution is computed from the control regions used for the enrichment analysis.

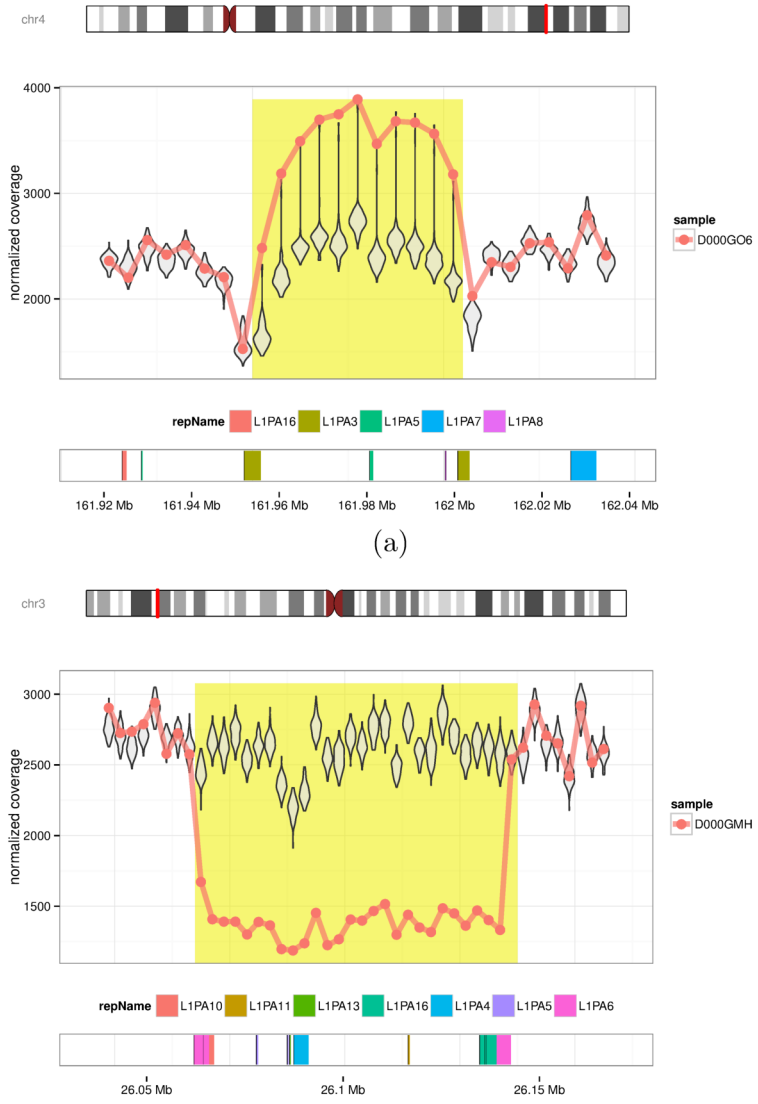


Figure S4.10: **Polymorphism likely caused by non-homologous allelic recombination between L1PA repeats.** Examples of CNV likely caused by non-allelic homologous recombination between two L1PA3 repeats (a) or L1PA6 (b). The line and points represent the coverage of one sample with a duplication (a) or a deletion (b), highlighted in yellow; the violin plots represent the distribution of the coverage in the reference samples.

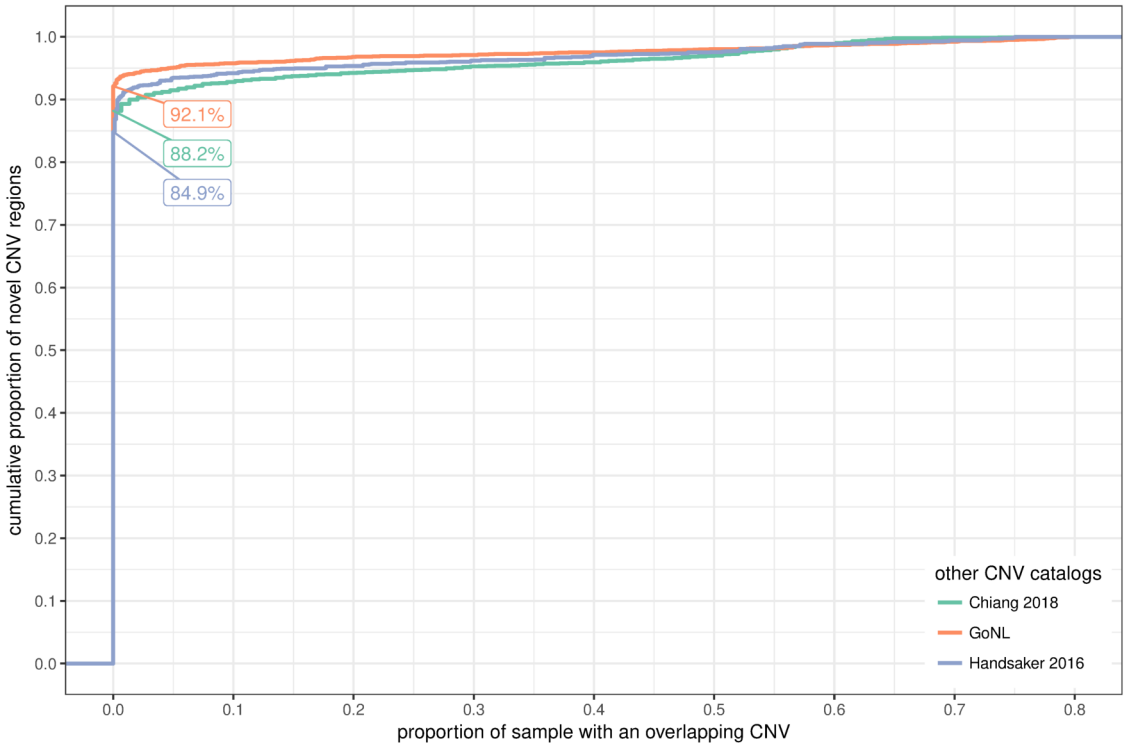


Figure S4.11: **Novel CNV regions and CNVs in other public catalogs.** Cumulative proportion of the 3,455 novel regions (y-axis) that overlap CNVs in different public CNV catalogs (colour) depending on their frequency in the public catalog (x-axis). The labels highlight the proportion of novel regions that don't overlap any CNV in the corresponding public catalog. Novel regions were defined as overlapping a CNV in more than 1% of the individuals but absent from the 1000GP catalog.

Supplementary Information

Data

Twin study All patients gave informed consent in written form to participate in the Quebec Study of Newborn Twins[189]. 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[143]. 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.[107]. This data have been derived from different sample collections:

- The LifeLines Cohort Study (http://www.lifelines.nl/), supported by the Netherlands Organization of Scientific Research (NWO, grant 175.010.2007.006), the Dutch government's Economic Structure Enhancing Fund (FES), the Ministry of Economic Affairs, the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the Northern Netherlands Collaboration of Provinces (SNN), the Province of Groningen, the University Medical Center Groningen, the University of Groningen, the Dutch Kidney Foundation and Dutch Diabetes Research Foundation.
- The EMC Ergo Study (http://www.ergo-onderzoek.nl/wp/).
- The LUMC Longevity Study, supported by the Innovation-Oriented Research Program on Genomics (SenterNovem IGE01014 and IGE05007), the Centre for Medical Systems Biology and the National Institute for Healthy Ageing (Grant 05040202 and 05060810).
- VU Netherlands Twin Register (http://www.tweelingenregister.org/).

In short, samples were sequenced on an Illumina HiSeq 2000 instrument (91-bp paired-end reads, 500-bp insert size). We downloaded the aligned read sequences

(BAM) for the 500 parents in the data set. We further performed indel realignment using GATK 3.2.2, adjusted pairs coordinates with Samtools 0.1.19, marked duplicates with Picard 1.118, and performed base recalibration (GATK 3.2.2). The average sequencing depth was 14x (minimum 9x / maximum 59x).

Genomic annotations Gencode annotation (V19) was directly downloaded from the consortium FTP server at ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz. Other genomic annotations were downloaded from the UCSC database[218] server at http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database. 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 / Short Tandem Repeats	`simpleRepeat.txt.gz`
RepeatMasker	`rmsk.txt.gz`

Read count across the genome

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

RD and mappability estimates

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

The mappability track[206] was downloaded from UCSC[218] (`wgEncodeCrgMapabilityAlign100mer.bw`) and the average mappability was computed for each bin. One sample was randomly selected and we compared its coverage with the mappability estimates. We then computed the mean and standard deviation of the coverage in each bin across the other samples and compared it with the sample coverage. We also compared the inter-sample average with the mappability estimates.

To compute Z-scores that integrates the observed coverage variation we used two approaches. The first modeled the coverage metrics (average or standard deviation)

using the mappability estimates and computed a Z-score from the predicted coverage and global standard deviation. A generalized additive model was fitted using a cubic regression spline on the mappability estimates (mgcv R package). In the second approach, Z-scores were computed using the inter-sample average and standard deviation. The normality of these two Z-score distributions were compared in term of excess kurtosis and skewness. For the kurtosis and skewness computation, we removed outlier Z-scores with an absolute value greater than 10. These bins could be regions of CNV and would bias the estimates. The Z-score distributions were also compared in bins from 10 different mappability intervals.

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

CNV detection with PopSV

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

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

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

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

CNV calling After targeted normalization the coverage in each sample is compared to the coverage in the reference samples. A Z-score is computed and translated into a P-value that is then corrected for multiple testing. Consecutive bins with sig-

nificant excess or lack of reads are merged and returned as potential duplication or deletion. Copy number estimates are derived from the coverage across the bin and the average coverage across the reference samples. However, it is important to note that the definition of a variant is different from other methods. Here a variant is defined by the major allele in the population rather than the reference genome state. Most of the genome is in a diploid state compared to the reference genome and sufficiently covered by sequencing reads that the copy number state can be correctly estimated by PopSV's population-based approach. However, highly polymorphic variants are called relative to the major allele in the population and additional efforts are required to assess the copy number state. Variants in extremely low-mappability regions are also difficult to fully characterize and might be caused by rare insertion in the reference genome or complex alleles. Nonetheless, PopSV can efficiently detect the presence of CNV in any situation. More details are available in the method paper[2].

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

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

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

Validation and benchmark

Running FREEC, CNVnator, cn.MOPS and LUMPY FREEC[40] segments the RD values of a sample using a LASSO-based algorithm. It was run on each sample separately, starting from the BAM file, using the same bin sizes as for PopSV. FREEC internally corrects the RD for GC and mappability bias. In order to compare its performance in low-mappability region, the minimum *"telocentromeric"* distance was set to 0. The remaining parameters were set to default. Of note an additional run with slightly looser parameter (`breakPointThreshold=0.6`) was performed to get a larger set of calls used in some parts of the *in silico* validation analysis to deal with borderline significant calls.

CNVnator[41] uses a mean-shift technique inspired from image processing. It was run on each sample separately, starting from the BAM file, using the same bin sizes as for PopSV. CNVnator also corrects internally for GC bias and we used default parameters. For the analysis using higher confidence calls, we used calls with either 'eval1' or 'eval2' lower than 10^{-5} (instead of the default 0.05).

cn.MOPS[42] considers simultaneously several samples and detects copy number variation using a Poisson model and a Bayesian approach. It was run on the same GC-corrected bin counts used for PopSV. All the samples are analyzed jointly. Of note an additional run with slightly looser parameter (`upperThreshold=0.32` and `lowerThreshold=-0.42`) was performed to get a larger set of calls used in some parts of the *in silico* validation analysis to deal with borderline significant calls.

LUMPY[67] which uses an orthogonal mapping signal: the insert size, orientation and split mapping of paired reads. The discordant reads were extracted from the BAMs using the recommended commands. Split-reads were obtained by running YAHA[66] with default parameters. All the CNVs (deletions and duplications) larger than 300 bp were kept for the upcoming analysis. `BND` variants with both ends more than 300 bp apart in the same chromosome were also included as they could be CNVs lacking support to characterize their type properly. Calls with 5 or more supporting reads were considered high-confidence.

Clustering samples from the Twin study A distance between two samples A and B was defined as : $1 - 2\frac{|R_A \cap R_B|}{|R_A|+|R_B|}$ where R_A represents the regions called in sample A, $R_A \cap R_B$ the regions called in both A and B, and $|R|$ the cumulative size of the regions. Hence, the similarity between two samples is represented by the amount of sequence found in both divided by the average amount of sequence called. This distance is used for hierarchical clustering of the samples in the Twin dataset. The clustering was performed using only calls in regions with extremely low coverage (reference average $\leq$100 reads). Different linkage criteria (*average*, *complete* and *Ward*) were used for the exploration. In our dendograms we used the *average* linkage criterion. The concordance between the clustering and the pedigree was estimated by the Rand index, grouping the samples per family. For each method and linkage criteria, the Rand index was computed for every possible dendogram cut (x-axis in Figure S4.3).

Experimental validation

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

The presence of a deletion was tested by comparing the size of the amplified fragment in affected and control samples. If the affected sample showed a lower band than a control with a predicted 2 copies, the deletion was considered validated. On

the other hand if affected sample and controls had one similar band, the deletion was considered non-validated. Of note, the validation rate might be under-estimated because visual prediction of the breakpoint is not always accurate.

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

Analysis of CEPH12878

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

Comparison with assemblies We downloaded the genome assembly produced from short reads, Pacbio and BioNano reads[59] from ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/013/985/GCA_001013985.1_ASM101398v1/GCA_001013985.1_ASM101398v1_genomic.fna.gz. We also downloaded a second assembly that was used 10X Genomics linked reads instead of the Pacbio reads[215]. It is available at http://kwoklab.ucsf.edu/resources/nmeth_201604_NA12878_hybrid_assembly.fasta.gz

For each selected variant, we retrieved the two 50 Kbp flanking sequences in the reference genome and aligned them against the public assemblies with BLAST[216]. The output was parsed to identify regions with two flanks aligning in at least 1 Kbp of a contig. MUMmer plots[217] between the reference sequence and the contigs were visually inspected. The assembly supported PopSV calls when a deletion was visible in the expected region (between the flanks). The assembly supported the reference genome sequence when a contig crosses the variant without clear structural variant.

SV calls from a long-read sequencing study We downloaded the SV calls from the Pacbio reads and assembled contigs in Pendleton et al.[59]. The VCF file is publicly available at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/NA12878.sorted.vcf.gz. We overlapped PopSV calls with deletions from this SV catalogs. Because we used 5 Kbp bins for PopSV, at least 1 Kbp of a PopSV calls needed to overlap a deletion from Pendleton et al.[59] to be considered as sufficient support. Of note, the distribution of the overlap tended to be either null or higher than 1 Kbp supporting this choice.

Local assembly of Pacbio reads Corrected Pacbio reads from citetPendleton2015 were downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/corrected_reads_gt4kb.fasta. Each read was split in 200 bp fragments and mapped to the human reference genome (version hg19). From this mapping information we selected full Pacbio reads with at least one 200 bp mapping within a region of interest (with 30 Kbp flanks). For each region, the reads were mapped to the reference sequence with exonerate and we kept reads with partial mapping as they may support a SV. These reads were then assembled using Canu[242]. A consensus sequence was also derived for reads clustered by alignment breakpoint and the clustalo[243] software. The assembled contigs and consensus were mapped to the reference genome to identify a potential breakpoint. The two regions flanking the alignment breakpoint and the sequence spanning the breakpoint were mapped to the entire genome. We used the results of this genome-wide mapping to select the best candidates: assembled sequence whose flanks align uniquely to the region of interest and with reduced alignment quality for the "middle" sequence that spanned the breakpoint. Candidate contig/consensus were further visualized with MUMmer plots[217]. The assembly supported PopSV calls when a deletion was visible in the expected region (between the flanks).

Genomic patterns of CNVs

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

Computing global estimates of copy number variation In Table 4.1, a call in extremely low coverage region is overlapped at more than 90% by the *extremely low coverage* track. To compute the total number of calls, we collapsed calls with an overlap higher than 50%. The amount of sequence affected in a genome was

computed by merging all the variants in the cohort and counting the number of affected bases in this reference genome. After the merging step, each base of the genome either overlapped a merged variant or not. Each affected base was counted only once, even if it overlapped CNVs in several samples or with large copy number differences.

Comparison with public CNV catalogs The SV catalog from Sudmant et al.[5] was downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz. The CNV catalog from Handsaker et al.[33] was downloaded from http://www.broadinstitute.org/%7Ehandsake/mcnv_data/bulk/1000G_phase1_cnv_genotypes_phased_25Jul2014.genotypes.vcf.gz For the CNV catalog from Chiang et al.[36], we downloaded `GTEx_Analysis_2016-10-24_WholeGenomeSeq_147Indiv_SV_sites.vcf.gz` from the GTEx Data Portal at https://www.gtexportal.org/home/datasets (data release v6). The CNV catalog from Francioli et al.[107] was downloaded from https://molgenis26.target.rug.nl/downloads/gonl_public/variants/release6.1/20161013_GoNL_AF_genotyped_SVs.vcf.gz. We retrieved the set of autosomal deletion, duplication and CNVs. When comparing the global estimates of CNV with PopSV, we removed deletions smaller than 300 bp as well as variants with high frequency ($> 80\%$). This remaining SVs represent CNVs that could in theory be detected by PopSV's approach. Using this sub-set, we derived the number of variants, number of variants smaller than 3 Kbp, number of variants in *extremely low coverage* regions, and amount of genome affected. These number are computed exactly as the one presented in Table 4.1 for PopSV's results.

CNV frequency comparison The frequency at which a region is affected by a CNV is computed using calls from the 620 unrelated samples. The copy-number change is not taken into account in the computation and the frequency is derived for all the nucleotide that overlaps at least one CNV. Using each catalog we computed, for each base in the genome, the proportion of individuals with a CNV. This frequency measure facilitates the comparison of catalogs with different methods and resolution. We represented the distribution as a cumulative proportion distribution in Figure 4.3a. The graphs read as "how much of the total affected genome is called in at more than X% of the population". The frequency distribution was computed separately for deletions and duplications (and *CNV* in the 1000 Genomes Project catalog). Of note, the 1000 Genomes Project was down-sampled to 640 random individuals in order to give comparable frequency curves.

Comparison with CNV catalogs from long-read studies First, the SV catalog from Chaisson et al.[58] was downloaded from http://eichlerlab.gs.washington.edu/publications/chm1-structural-variation/. Recurrent calls were collapsed in both PopSV and the 1000 Genomes Project catalogs. PopSV's catalog corresponded to all germline calls in the Twin study, renal cancer dataset and GoNL. The 1000 Genomes Project catalog contained all the deletions, duplications and CNVs, no matter the size or frequency. The analysis was also performed separately on deletions, duplications, low-mappability regions and extremely low-mappability regions. For each comparison, we randomly selected control regions with sizes and overlap with assembly gaps similar to the SVs in Chaisson

et al.[58] (see Selecting control regions). A logistic regression tested the enrichment of CNVs in the Chaisson catalog versus the control regions. The regression was performed on 50 different sampling of the control regions for each comparison. The 50 samplings are represented by the boxplot in Figure 4.3b. We compared the estimates from the logistic regression. They represent the log odds ratio of a CNV overlapping the catalog from Chaisson. The same analysis was performed using the SV catalog from Pendleton et al.[59] downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/NA12878.sorted.vcf.gz.

Distance to centromere, telomere and assembly gaps The centromeres, telomeres and assembly gaps (CTGs) are annotated in the gap track from UCSC[218]. However, some chromosomes were missing telomere annotations. We defined them as the 10 Kbp region at the ends of chromosomes derived from the cytogenetic bands track.

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

Because this distribution changes with the size of the variants, we sampled random regions in the genome with similar sizes and computed the same distance distribution (see Selecting control regions). Thanks to this null distribution we were able to see if variants were located closer/further to CTG than expected by chance.

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

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

$$\left\{b, \forall\ feature\ f, O_f(d_f^b - \frac{S_r}{2}) < 0\right\} \tag{6.2}$$

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.

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

Enrichment in genomic features We tested different genomic features, starting with: genes, exons, low-mappability regions, segmental duplications, satellites, simple repeats and transposable elements. The different satellite families, frequent

simple repeat motives, transposable element families were also tested. We overlapped each genomic feature with CNVs and control regions. We then computed the fold change in proportion of regions overlapping a feature, in CNV versus control regions. A pseudo count was added when computing this ratio:

$$\text{Fold enrichment} = \frac{\frac{|CNV \cap Feature|+1}{N+1}}{\frac{|Control \cap Feature|+1}{N+1}} = \frac{|CNV \cap Feature| + 1}{|Control \cap Feature| + 1} \tag{6.3}$$

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

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

$$\log\left(\frac{P(\text{feature overlap})}{P(\text{no overlap})}\right) = \beta_0 + \beta_{CNV} \cdot CNV \tag{6.4}$$

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

To control for the enrichment in segmental duplication we used control regions with similar overlap profile (see Selecting control regions). We also added a variable representing the overlap with segmental duplication in the model:

$$\log\left(\frac{P(\text{feature overlap})}{P(\text{no overlap})}\right) = \beta_0 + \beta_{CNV} \cdot CNV + \beta_{SD} \cdot SD \tag{6.5}$$

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

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

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

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