DNA sequence features underlying large-scale duplications and deletions in human

Mateusz Kołomański

Wrocław University of Environmental and Life Sciences: Uniwersytet Przyrodniczy we Wrocławiu
<https://orcid.org/0000-0001-7424-3919>

Joanna Szyda

Wrocław University of Environmental and Life Sciences: Uniwersytet Przyrodniczy we Wrocławiu
<https://orcid.org/0000-0001-9688-0193>

Magdalena Frąszczak

Wrocław University of Environmental and Life Sciences: Uniwersytet Przyrodniczy we Wrocławiu
<https://orcid.org/0000-0002-8012-4980>

Magda Mielczarek (✉ magda.mielczarek@upwr.edu.pl)

Uniwersytet Przyrodniczy we Wrocławiu <https://orcid.org/0000-0002-1086-9119>

Research article

Keywords: Genomes Project, Copy Number Variants, DNA sequence complexity, GC content

Posted Date: April 26th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-458335/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

Copy Number Variants (CNVs) may cover up to 12% of the whole genome and have substantial impact on phenotypes. We used 5 867 duplications and 33 181 deletions available from the *1000 Genomes Project* to characterize genomic regions vulnerable to CNV formation and to identify sequence features characteristic for those regions.

Results

Only 14 CNVs contained unknown nucleotides, which reflected the high quality of the analysed data. The GC content for deletions was lower and for duplications was higher than for randomly selected regions. In regions flanking deletions and downstream of duplications content was higher than in the random sequences, but upstream of duplication content was lower. In duplications, the percentage of low complexity sequences was not different from the randomized data, but in deletions it was higher. Moreover, it was significantly higher upstream of CNVs as compared to random sequences. Conversely, it was lower downstream of duplications. The majority of CNVs intersected with genic regions - mainly with introns.

Conclusions

GC content may be associated with CNV formation and CNVs, especially duplications are initiated in low complexity regions. Moreover, CNVs located or overlapped with introns indicate their role in shaping intron variability. Genic CNV regions were enriched in many essential biological processes such as cell adhesion, synaptic transmission, transport, cytoskeleton organization, immune response and metabolic mechanisms what indicates that this large-scaled variants play important biological roles.

Background

One of the biggest achievements of modern genetics and genomics was sequencing of full human genome for the first time. The *Human Genome Project* begun in 1990 and lasted thirteen years. During that time, great progress in computational biology has been made. The *Project* was carried out across three phases and each was an improvement in DNA data generation and analysis. A multitude of methods has been used for raw data processing, variant calling, as well as for variant filtering and validation. The project had excessive impact on our understanding of genes' function as well as on the origins of genetic diseases (1). A few years after the *Human Genome Project* had been completed, even more ambitious studies have begun. The *1000 Genomes Project* aimed to sequence at least 1 000 human genomes in order to study genetic diversity of humans. It finished in 2015 and resulted in 2 504 sequenced genomes of individuals representing 26 populations as well as in the identification of over 88 million of genetic variants (2). The study found out that an individual human genome differs from the reference genome at 4–5 million sites. The most common type of polymorphisms are Single Nucleotide

Polymorphisms (SNPs) – about 84.7 million SNPs with relatively high frequency have been identified in human populations (2). Although SNPs impact many human phenotypes from complex traits such as height (3) to disease susceptibility (4, 5), large scale structural polymorphisms cover higher percentage of genomic sequence than SNPs underscoring their important role in genetic diversity (6, 7). Structural Variants (SVs) are longer than 50 bp and form inversions, translocations, insertions, deletions or duplications (8). Genomic regions covered by large-scale duplications and deletions, also called Copy Number Variants (CNVs), constitute up to 12% of the whole genome (6). CNVs can be inherited or formed *de novo* (9). Adults show significantly higher abundance of DNA rearrangements than infants, which suggests that CNVs accumulate during lifetime (10). CNVs have been proven to impact various aspects of human health (7). Several studies reported increased frequency of CNVs in patients with congenital heart disease (11) and breast cancer (12), excess of rare single gene duplications and deletions was observed in patients with schizophrenia (13).

The aim of this study was to characterize regions of human genome that are susceptible to structural duplications or deletions. We studied the question, whether specific regions of the human genome are particularly vulnerable to the formation of CNVs, and if so, what distinguishes those regions from unchanged regions.

Results

Reference genome sequence features

Unknown nucleotides (N) content

Among all of the regions of the human reference genome GRCh38 (14) covered by CNVs, only 14 contained unknown nucleotides – eight of them were duplications (Table 1) and six of them were deletions (Table 2). Percentage of unknown nucleotides in duplications ranged between 0.002% and 22.06% and in deletions it varied between 0.0004% and 63.21%, with half of those being above 50%. It is worth to notice that some regions contained a fixed number of unknown nucleotides (i.e. 100 or 50 000 Ns). In regions flanking CNVs, only one sequence, located upstream of deletion, contained 17 unknown nucleotides (chromosome: 18, start: 9 984, end: 10 083).

Table 1
Unknown nucleotide (N) content in duplicated regions

Chromosome	Start	End	Length	N-content
2	32 403 872	33 107 363	703 491	2 000 (0.28%)
6	167 529 394	167 756 083	226 689	50 000 (22.06%)
7	136 433	405 502	269 069	2 396 (0.89%)
10	53 027 785	53 147 359	119 574	2 (0.002%)
10	131 249 138	132 223 150	974 012	100 (0.01%)
10	131 446 890	131 640 975	194 085	100 (0.05%)
17	425 092	895 536	470 444	41 515 (8.82%)
22	49 608 227	49 641 383	33 156	1 (0.003%)

Table 2
Unknown nucleotide content (N) in deleted regions

Chromosome	Start	End	Length	N-content
3	59 888 393	60 159 064	270 671	1 (0.0004%)
5	17 508 562	17 596 059	87 497	50 000 (57.14%)
5	17 526 780	17 605 880	79 100	50 000 (63.21%)
8	7 227 940	7 680 412	452 472	50 000 (11.05%)
10	38 529 525	38 614 773	85 248	43 431 (50.95%)
X	155 916 731	155 919 319	2 588	1 (0.04%)

GC content

All sequences containing unknown nucleotides were excluded from the GC-content analysis. The average content of GC pairs was very similar in duplications ($41.86\% \pm 5.83$) and deletions ($41.08\% \pm 6.15$). The lowest content was 29.08% in duplications and 21.27% in deletions, while the maximum contents were respectively 68.90% and 73.46%. The visual examination of GC pair content distributions in CNVs, presented on Figures S1a and S2a in the Supplementary material, demonstrated that both are skewed indicating an excess of low GC contents, while the regions flanking CNVs exhibit a more symmetric distribution (Figures S3a, S3b, S4a, S4b). The distributions of GC pair content of duplications ($P = 0.004$) and deletions ($P = 7.955 \cdot 10^{-12}$) significantly differed from the randomized set of sequences. In particular, deletions contained less GC pairs than random regions ($P = 3.977 \cdot 10^{-12}$), while duplications were enriched in GC pairs as compared to randomized set of sequences ($P = 0.0024$). Deletion flanking regions contained more GC pairs than the corresponding randomized sequences ($P = 1.5 \cdot 10^{-10}$ for regions

upstream deletions, and $P = 1.259 \cdot 10^{-21}$ for regions downstream deletions). The same situation was observed for the comparison of the distributions of GC pair content in regions downstream of duplications with the randomised set of sequences ($P = 1.74 \cdot 10^{-9}$), while the GC content upstream of duplications was higher than in random sequences ($P = 0.014$). Descriptive statistics for all regions were summarised in Table 3.

Table 3
Guanine-Cytosine pairs content [%] in the investigated regions.

Region	Min	Mean	Max	Sd
Duplications	29.08	41.86	68.90	5.83
Set 1 (randomized duplications)	31.74	41.59	65.74	5.63
Upstream duplications	7.00	41.24	83.00	11.59
Downstream duplications	6.00	42.60	86.00	10.73
Set 3 (randomized up- and downstream duplications)	1.00	41.42	84.00	10.54
Deletions	21.27	41.08	73.46	6.15
Set 2 (randomized deletions)	20.56	41.54	77.19	6.53
Upstream deletions	0.00	41.82	84.00	10.53
Downstream deletions	0.00	42.05	89.00	10.47
Set 4 (randomized up- and downstream deletions)	0.00	41.41	89.00	10.66

Sequence complexity

4 798 406 low complexity regions (LCRs) were identified for the GRCh38 reference genome. Lengths of those regions varied between 7 bp and 25 072 bp with mean of 29 bp (± 56). All duplications and 93.93% of deletions contained LCRs. Median number of LCRs overlapped with duplications was 57 and with deletions was six (Fig. 1, Table 4), what on average made up 4.59% of duplication length and 4.66% of deletion length (Fig. 2, Table 4). Furthermore, 20.83% of sequences upstream of duplications and 16.52% of sequences downstream of duplications contained low complexity sequence. The corresponding statistics for regions flanking deletions were 20.37% of upstream sequences and 19.25% of downstream sequences. Among them, on average, 4.73% of the length of regions upstream of duplications, 3.47% of the length of regions downstream of duplications, 4.44% of the length of regions located upstream of deletions and 4.16% of the length of regions downstream of deletions. In the random sequence Set 1, 99.21% of sequences contained low-complexity regions, in Set 2 – 97.62%, in Set 3 – 18.12%, and in Set 4 – 18.61%. The distributions of all considering sets containing low complexity sequences significantly deviated from the normal distribution. The distributions of low-complexity sequence content in Set 1 and in duplications ($P = 0.106$) as well as between Set 4 and regions downstream of deletions ($P = 0.078$) did not differ. The percentage of low complexity sequences was significantly higher upstream of deletions (P

= $1.907 \cdot 10^{-8}$), upstream of duplications ($P = 8.982 \cdot 10^{-5}$) and within deletions ($P = 2.963 \cdot 10^{-19}$) than in corresponding randomized sets. Conversely, the distribution of low-complexity sequence content downstream of duplications was significantly lower than in Set 3 ($P = 0.007$).

Table 4
Content of low-complexity regions within CNV-related regions

Regions	Number of overlapped low-complexity regions			Content of low complexity regions (%)		
	Min.	Mean	Max.	Min.	Mean	Max.
Duplications	1	104	1 698	0.07	4.59	47.52
Set 1 (randomized duplications)	0	58	114	0.00	4.49	59.44
Upstream duplications	0	0	3	0.00	4.73	100.00
Downstream duplications	0	0	2	0.00	3.47	100.00
Set 3 (randomized up- and downstream duplications)	0	0	3	0.00	4.19	100.00
Deletions	0	6	3 769	0.00	4.66	100.00
Set 2 (randomized deletions)	0	6	25	0.00	4.55	98.07
Upstream deletions	0	0	3	0.00	4.44	100.00
Downstream deletions	0	0	3	0.00	4.16	100.00
Set 4 (randomized up- and downstream deletions)	0	0	4	0.00	4.36	100.00

Functional annotation of CNVs

The 5 867 identified duplications overlapped with 9 111 genes corresponded to 35 317 transcripts. The 33 181 identified deletions overlapped with 19 022 genes corresponded to 71 542 transcripts. The most common Sequence Ontology (SO) terms determined for **duplication regions** comprised: intron variants and transcript amplifications (Fig. 3a). For **deleted regions**, the most common SO terms were: feature truncations and intron variants (Fig. 3b). In the context of biological processes, genes containing duplications were significantly overrepresented in gene ontologies related to homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156), modulation of chemical synaptic transmission (GO:0050804), cytoskeleton organization (GO:0007010) and in the Cadherin signalling pathway (P00012) as well as underrepresented for complement activation, classical pathway (GO:0006958), including immune response (GO:0006955). Genes containing deletions were significantly overrepresented in GO terms related to transport (GO:0006810), cellular component organization (GO:0016043) and regulation of cellular process (GO:0050794).

Discussion

CNVs are one of the major source of human genetic diversity (15, 16). In recent years many studies on CNVs in the human genome have been published. Vast majority of them regarded clinical applications, such as the impact of CNVs on arrhythmogenic pathologies (17), autism and psychosis (18) as well as on cancer (19). However, only few studies have focussed on genome architectures underlying CNVs (20).

Our study revealed that the distribution of GC pairs differs between CNV regions and randomised sets of sequences. The GC content for deletions was lower and for duplications - higher than in random regions. Rigau et al. (21) observed that deleted regions have significantly higher GC content to that of the introns where they are located and that loss causes a significant decrease of the overall GC content of the introns. In our study, the majority of deletions was annotated to introns what may explain the GC content imbalance (22). Since genic regions contain more GC we can hypothesize that deletions are functionally more severe type of structural variation than duplications and therefore deletions appear more seldom in genic regions than the randomized average, while duplications appear more often in genic regions than the randomized average. Because duplications appear more often in GC rich regions (i.e. genes) they have some evolutionary advantage (23). It is also worth to notice, that according to Dittwald et al. (24) GC content, among others, is positively correlated with the frequency of nonallelic homologous recombination (NAHR) being a common cause of CNV formation. According to Romiguier et al. (25), GC-rich sequences are prone to deletions because base composition imbalance trigger replication slippage. On the other hand, Chen et al. (26) did not report a difference in GC content between CNV regions and autosomal average. Considering regions flanking CNVs, our study demonstrated that the GC content is higher than in the corresponding randomized sequences upstream and downstream of deletions. The same was observed for the comparison of distributions of GC pair content in regions downstream of duplications with the randomised set of sequence, but not for upstream regions of duplications where GC content is lower than in random case. Bose et al. (27) investigated breakpoint regions for various SVs, including CNV, and concluded that all SV types had a higher GC percentage than the genome average.

LCRs are defined as sequences composed of a lower diversity of nucleotides compared to other areas of the genome. Our study demonstrated that sequences in deleted regions have significantly higher LCRs content compared to randomised data, the same as for regions located upstream of CNVs. Conversely, the LCR content was lower downstream of duplications. Barski et al. (28) investigated sequence complexity in regions flanking CNV in *Bos taurus*. The study concluded that duplications and deletions preferentially form in regions of low complexity. CNVs also appear to be enriched in regions of low mappability, as well as within satellites and Short Tandem Repeats (29), all of those characterised by low complexity. Chen et al. (20) postulated that low-copy and high-copy repeats can induce DNA instability, resulting in errors in replication and repair mechanisms and consequently leading to the formation of CNVs.

Functional annotation revealed that majority of CNVs were located within gene regions. Similar observation was made by Chen et al. (26) for population-specific CNVs. Higher gene density in regions

covered by CNVs then in random genome regions has been also highlighted by Johansson and Feuk (30). According to Rigau et al. (21), intronic deletions are the most frequent CNVs in protein-coding genes in humans, while deletions overlapping exons are less frequent than expected by chance. Therefore, it was also suggested that intronic CNVs contribute to the variability of gene expression and splicing in human populations. The homophilic cell adhesion identified as an ontology over-represented in deletions in our study was also reported for genes with somatic duplications in placenta by Kasak et al. (31). Moreover, Morello et al. (32) observed that synaptic transmission, an ontology over-represented in deletions in our study, was the most highly enriched term in CNV-driven differentially expressed genes in a sporadic form of amyotrophic lateral sclerosis. Involvement of CNVs in immune response mechanisms have already been reported by Perry et al. (33) and, the same as in our study, genes with immune response functions were overrepresented in human CNV regions (6). Deleted genes were significantly overrepresented in GO terms related to transport, cellular component organization and regulation of cellular process, which indicates that deletions significantly affect essential cellular mechanisms (34). Duplicated genes were enriched in the Cadherin signalling pathway, which is involved in multiple biological processes, such as development, neurogenesis, cell adhesion, and inflammation. It's enrichment has been reported in the context of many diseases including cancer (35).

Conclusions

In this study we analysed copy number duplications and deletions obtained within the frame of the *1000 Genomes Project* where a variety of methods was used to generate a reliable set of large-scaled polymorphisms. Our results indicated that GC content may be associated with CNV formation and CNVs, especially duplications are initiated in low complexity regions. Moreover, CNV were often located or overlapped with introns what indicates their role in shaping intron variability.

Material And Methods

Dataset

The human reference genome GRCh38 was downloaded from the National Center for Biotechnology Information database (36). Polymorphisms, including CNVs, were obtained within the frame of the 3rd phase of 1000 Genomes Project and are available from the European Bioinformatics Institute (<https://www.ebi.ac.uk>) under the ID: *estd214*. In *1 000 Genomes Project* a variety of methods was used to identify polymorphisms. Primary data resulted from oligonucleotide genotyping, low-coverage whole genome sequencing as well as exome sequencing, with complete genomics and high-coverage PCR-free sequencing of selected samples serving for validation. As much as nine programs were used for structural variant calling and the final set of SV was the merged output of all the software. SVs were then validated using various methods, including microarrays, PCR-free whole genome sequencing and PacBio sequencing, as well as PCR. The estimated false discovery rate for CNVs was below 5% (2).

In our study, from all available genetic polymorphisms, only CNVs considered as duplications or deletions were extracted. Overlapping CNVs were considered independently, resulting in 5 867 duplications and 33 181 deletions. Length of duplications ranged between 3 006 bp and 988 090 bp, with median of 37 036 bp and mean of 66 527±91 091 bp. Length of deletions ranged between 204 bp and 2 258 238 bp, with median of 3 774 bp and mean of 12 143±34 749 bp (Fig. 4).

Reference genome sequence features

The Samtools software (37) was used to extract regions covered by CNVs from the GRCh38 reference genome. Moreover, reference sequences flanking CNVs coordinates (100 nucleotides up- and downstream of each CNV) were extracted. These regions were considered in the context of unknown nucleotides (denoted as “N”), Guanine-Cytosine pairs, sequence complexity and functional annotation. Unknown nucleotides and Guanine-Cytosine pairs content were calculated in all considered sequences (i.e. CNVs and their flanking regions). In order to compare regions covered by CNVs with regions of the genome not affected by CNVs, random sequences were chosen. First, random coordinates were generated using Python the “random” module, then regions defined by those coordinates were extracted from the reference genome using the Samtools software. In total four sets of random sequences were generated: (i) **Set 1** contained 5 859 sequences of length equal to the median length of duplications (37 036 bp), (ii) **Set 2** contained 33 175 sequences of length equal to the median length of deletions (3 774 bp) , (iii) **Set 3** contained 5 867 sequences of 100 bp length and was used for comparisons with sequences up- and downstream of duplications, (iv) **Set 4** contained 33 181 sequences of 100 bp length and was used for comparison with sequences up- and downstream of deletions. All sequences containing unknown nucleotides were excluded and re-randomized. The number of random sequences in each Set matched the number of sequences in the corresponding CNV and flanking groups, as defined above. The percent of GC pairs was calculated for all sets of sequences using Python (38). The distributions of GC-content were tested for normality using the Kolmogorov test. The H_0 stating that the distributions of GC-content follow the normal distribution with mean and variance given by the considered data sets. The test statistics, which is defined as the supremum of difference between theoretical and empirical distribution, has the same distribution as the classical Kolmogorov statistics. Furthermore, the distributions of GC-pairs content of true CNV-related sequences were compared with the corresponding randomised Sets using the Wilcoxon–Mann–Whitney test, with H_0 stating that the distributions of GC-content are equal. The normalised test statistic is given by:

$$Z = \frac{V - \frac{n \cdot m}{2}}{\sqrt{\frac{n \cdot m \cdot (n + m + 1)}{12}}} \sim N(0,1)$$

$$V = \sum_{j=1}^m S_j - \frac{m \cdot (m+1)}{2}, S_j$$

where S_j denotes ranks corresponding to the GC-pairs percentage classes in the random sequences, n is a count of deletion/duplication/flanking CNVs regions and m is a count of sets with random sequences.

Sequence complexity

Sequence complexity of the entire reference genome was estimated using the sDust software (39). The overlap between low-complexity regions defined by sDust and CNV-related regions was determined by using the bedtools software (40) for true CNVs and flanking regions, as well as for the random Sets. The distributions of low-complexity sequence contents in CNV and flanking regions as well as in randomised data were compared using the Wilcoxon-Mann-Whitney test. Testing and visualisation were created using the R package (41).

Functional annotation

The Variant Effect Predictor (VEP) software (42) was used for the functional annotation of CNVs. Gene Ontology enrichment (35) was tested using the Fisher's Exact test with the False Discovery Rate (FDR). Moreover, significantly enriched signalling pathways from the Panther (35) database, was identified using the KOBAS tool (43) applying the Fisher's Exact test with FDR.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Polymorphisms, including CNVs, were obtained within the frame of the 3rd phase of 1000 Genomes Project and are available from the European Bioinformatics Institute (<https://www.ebi.ac.uk>) under the ID: *estd214*

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

MM designed the study. MK performed computations. MF suggested and performed the statistical analyses. MK and MM wrote the draft of manuscript. JS contributed to the concept of the study and improved the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge Poznan Supercomputing and Networking Centre for hosting the computations.

References

1. Rossiter BJ, Caskey CT. Impact of the human genome project on medical practice. *Ann Surg Oncol*. 1995 Jan;2(1):14–25.
2. 1000 Genomes Project Consortium, Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, et al. A global reference for human genetic variation. *Nature* [Internet]. 2015;526(7571):68–74. Available from: <https://doi.org/10.1038/nature15393>
3. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* [Internet]. 2010;42(7):565–9. Available from: <https://doi.org/10.1038/ng.608>
4. Kalia N, Kaur M, Sharma S, Singh J. A Comprehensive in Silico Analysis of Regulatory SNPs of Human CLEC7A Gene and Its Validation as Genotypic and Phenotypic Disease Marker in Recurrent Vulvovaginal Infections. *Front Cell Infect Microbiol*. 2018;8:65.
5. Ponomarenko P, Chadaeva I, Rasskazov DA, Sharypova E, Kashina E V, Drachkova I, et al. Candidate SNP Markers of Familial and Sporadic Alzheimer's Diseases Are Predicted by a Significant Change in the Affinity of TATA-Binding Protein for Human Gene Promoters. *Front Aging Neurosci* [Internet]. 2017 Jul 20;9:231. Available from: <https://pubmed.ncbi.nlm.nih.gov/28775688>
6. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. *Nature* [Internet]. 2006;444(7118):444–54. Available from: <https://doi.org/10.1038/nature05329>
7. Lee C, Scherer SW. The clinical context of copy number variation in the human genome. *Expert Rev Mol Med* [Internet]. 2010/03/09. 2010;12:e8. Available from: <https://www.cambridge.org/core/article/clinical-context-of-copy-number-variation-in-the-human-genome/EFDfEB7CEf4E7A42982038FC3F47FA50>
8. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. *Nature* [Internet]. 2011;470(7332):59–65. Available from: <https://doi.org/10.1038/nature09708>
9. Thapar A, Cooper M. Copy number variation: what is it and what has it told us about child psychiatric disorders? *J Am Acad Child Adolesc Psychiatry* [Internet]. 2013 Aug;52(8):772–4. Available from: <https://pubmed.ncbi.nlm.nih.gov/23880486>

10. Flores M, Morales L, Gonzaga-Jauregui C, Domínguez-Vidaña R, Zepeda C, Yañez O, et al. Recurrent DNA inversion rearrangements in the human genome. *Proc Natl Acad Sci U S A*. 2007 Apr;104(15):6099–106.
11. Marian AJ. Copy number variants and the genetic enigma of congenital heart disease. *Circ Res* [Internet]. 2014 Oct 24;115(10):821–3. Available from: <https://pubmed.ncbi.nlm.nih.gov/25342769>
12. Walker LC, Wiggins GAR, Pearson JF. The Role of Constitutional Copy Number Variants in Breast Cancer. *Microarrays* (Basel, Switzerland). 2015 Sep;4(3):407–23.
13. Szatkiewicz JP, Fromer M, Nonneman RJ, Ancalade N, Johnson JS, Stahl EA, et al. Characterization of single gene copy number variants in schizophrenia. *bioRxiv* [Internet]. 2019 Jan 1;550863. Available from: <http://biorxiv.org/content/early/2019/02/15/550863.abstract>
14. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*. 2017 May;27(5):849–64.
15. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* [Internet]. 2009;10:451–81. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/19715442>
16. Zhang L, Wang J, Zhang C, Li D, Carvalho CMB, Ji H, et al. Efficient CNV breakpoint analysis reveals unexpected structural complexity and correlation of dosage-sensitive genes with clinical severity in genomic disorders. *Hum Mol Genet* [Internet]. 2017 May 15;26(10):1927–41. Available from: <https://pubmed.ncbi.nlm.nih.gov/28334874>
17. Mates J, Mademont-Soler I, Fernandez-Falgueras A, Sarquella-Brugada G, Cesar S, Arbelo E, et al. Sudden Cardiac Death and Copy Number Variants: What Do We Know after 10 Years of Genetic Analysis? *Forensic Sci Int Genet*. 2020 Jul;47:102281.
18. Larson F V, Arrand JR, Tantam D, Jones PB, Holland AJ. Copy number variants in people with autism spectrum disorders and co-morbid psychosis. *Eur J Med Genet*. 2018 Apr;61(4):230–4.
19. Hopman S, Merks J, Eussen H, Douben H, Snijder S, Hennekam R, et al. Structural genome variations in individuals with childhood cancer and tumour predisposition syndromes. *Eur J Cancer* [Internet]. 2013;49(9):2170–2178. Available from: <https://doi.org/10.1016/j.ejca.2013.02.002>
20. Chen L, Zhou W, Zhang L, Zhang F. Genome architecture and its roles in human copy number variation. *Genomics Inform*. 2014 Dec;12(4):136–44.
21. Rigau M, Juan D, Valencia A, Rico D. Intronic CNVs and gene expression variation in human populations. *PLoS Genet* [Internet]. 2019 Jan 24;15(1):e1007902–e1007902. Available from: <https://pubmed.ncbi.nlm.nih.gov/30677042>
22. Aïssani B, Bernardi G. CpG islands, genes and isochores in the genomes of vertebrates. *Gene*. 1991 Oct;106(2):185–95.
23. Levasseur A, Pontarotti P. The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. *Biol Direct* [Internet]. 2011;6(1):11. Available from: <https://doi.org/10.1186/1745-6150-6-11>

24. Dittwald P, Gambin T, Szafranski P, Li J, Amato S, Divon MY, et al. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Res* [Internet]. 2013/05/08. 2013 Sep;23(9):1395–409. Available from: <https://pubmed.ncbi.nlm.nih.gov/23657883>
25. Romiguier J, Ranwez V, Douzery EJP, Galtier N. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res* [Internet]. 2010/06/07. 2010 Aug;20(8):1001–9. Available from: <https://pubmed.ncbi.nlm.nih.gov/20530252>
26. Chen W, Hayward C, Wright AF, Hicks AA, Vitart V, Knott S, et al. Copy number variation across European populations. *PLoS One* [Internet]. 2011/08/04. 2011;6(8):e23087–e23087. Available from: <https://pubmed.ncbi.nlm.nih.gov/21829696>
27. Bose P, Hermetz KE, Conneely KN, Rudd MK. Tandem Repeats and G-Rich Sequences Are Enriched at Human CNV Breakpoints. *PLoS One* [Internet]. 2014 Jul 1;9(7):e101607. Available from: <https://doi.org/10.1371/journal.pone.0101607>
28. Barski P, Mielczarek M., Frąszczak M. SJ. DNA sequence features underlying copy number variants. *Acta Sci Pol Zootech.* 2019;(18(2)):25–30.
29. Monlong J, Cossette P, Meloche C, Rouleau G, Girard SL, Bourque G. Human copy number variants are enriched in regions of low mappability. *Nucleic Acids Res.* 2018 Aug;46(14):7236–49.
30. Johansson AC V, Feuk L. Characterization of copy number-stable regions in the human genome. *Hum Mutat.* 2011 Aug;32(8):947–55.
31. Kasak L, Rull K, Vaas P, Teesalu P, Laan M. Extensive load of somatic CNVs in the human placenta. *Sci Rep* [Internet]. 2015 Feb 10;5:8342. Available from: <https://pubmed.ncbi.nlm.nih.gov/25666259>
32. Morello G, Guarnaccia M, Spampinato AG, Salomone S, D'Agata V, Conforti FL, et al. Integrative multi-omic analysis identifies new drivers and pathways in molecularly distinct subtypes of ALS. *Sci Rep* [Internet]. 2019 Jul 10;9(1):9968. Available from: <https://pubmed.ncbi.nlm.nih.gov/31292500>
33. Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, et al. Copy number variation and evolution in humans and chimpanzees. *Genome Res.* 2008 Nov;18(11):1698–710.
34. Alloza E, Al-Shahrour F, Cigudosa JC, Dopazo J. A large scale survey reveals that chromosomal copy-number alterations significantly affect gene modules involved in cancer initiation and progression. *BMC Med Genomics* [Internet]. 2011 May 6;4:37. Available from: <https://pubmed.ncbi.nlm.nih.gov/21548942>
35. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* [Internet]. 2019 Jan 8;47(D1):D419–26. Available from: <https://doi.org/10.1093/nar/gky1038>
36. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2018 Jan;46(D1):D8–13.
37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* [Internet]. 2009/06/08. 2009 Aug 15;25(16):2078–9. Available

from: <https://www.ncbi.nlm.nih.gov/pubmed/19505943>

38. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
39. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol.* 2006 Jun;13(5):1028–40.
40. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* [Internet]. 2010/01/28. 2010 Mar 15;26(6):841–2. Available from: <https://pubmed.ncbi.nlm.nih.gov/20110278>
41. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2013. Available from: <http://www.r-project.org/>
42. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol* [Internet]. 2016;17(1):122. Available from: <https://doi.org/10.1186/s13059-016-0974-4>
43. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 2011 Jul;39(Web Server issue):W316-22.

Figures

Figure 1

Duplications and deletions length (bp) containing low-complexity regions.

Figure 2

Low-complexity sequence content in duplications and deletions (CNVs not overlapping any region are not presented).

Figure 3

Functional annotation of (a) duplications and (b) deletions.

Figure 4

Duplications and deletions length (bp).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplemementarymaterialS1S4.docx](#)