Journal of Applied Genetics (2022) 63:527–533
https://doi.org/10.1007/s13353-022-00704-0

HUMAN GENETICS • ORIGINAL PAPER

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

Mateusz Kołomański[1] · Joanna Szyda[1] · Magdalena Frąszczak[1] · Magda Mielczarek[1]

Received: 28 December 2021 / Revised: 22 March 2022 / Accepted: 5 May 2022 / Published online: 20 May 2022

Abstract

Copy number variants (CNVs) may cover up to 12% of the whole genome and have substantial impact on phenotypes. We used 5867 duplications and 33,181 deletions available from the *1000 Genomes Project* to characterise genomic regions vulnerable to CNV formation and to identify sequence features characteristic for those regions. 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 and downstream of deletion regions, the percentage of low-complexity sequences was not different from the randomised data. In deletions and upstream of CNVs, it was higher, while for downstream of duplications, it was lower as compared to random sequences. The majority of CNVs intersected with genic regions — mainly with introns. 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, which indicates that these large-scaled variants play important biological roles.

Keywords 1000 Genomes Project · Copy number variants · DNA sequence complexity · GC content

Introduction

The *1000 Genomes Project*, finished in 2015, resulted in 2504 sequenced genomes of individuals representing 26 populations as well as in the identification of over 88 million of polymorphisms (1000 Genomes Project Consortium et al. 2015). 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 is single nucleotide polymorphisms (SNPs) — about 84.7 million. Copy number variations (CNVs) defined as deletions and duplications longer than 50 bp are less common that SNPs, but because of their length, they constitute up to 12% of the human genome (Redon et al. 2006). It is known that CNVs are not randomly distributed in eukaryotic genomes, but the biological mechanism of their genomic distribution is not fully understood (Nguyen et al. 2006; Makino et al. 2013). Certainly, there exist a considerable variation in CNV breakpoint location among individuals from the same species, as demonstrated, e.g. by Nicholas et al. (2009) for individuals representing several domestic dog (*Canis familiaris*) breeds. DNA sequence composition is one of the factors triggering the formation of CNV. Repeats of A/T nucleotides and sequences promoting the formation of hairpin structures were observed to mark CNV breakpoints in *Plasmodium falciparum* (Huckaby et al. 2019). Conversely, in mammals (domestic dogs), CNV breakpoints were enriched in G and C nucleotides (Berglund et al. 2012).

The aim of this study was to characterise DNA structure in regions of human genome that are susceptible to structural duplications or deletions. We searched for DNA sequence features promoting the formation of CNVs and the patterns of functional annotations of such deleted and duplicated regions.

Communicated by Michal Witt

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

[1] Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

Material and methods

Dataset

The human reference genome GRCh38 was downloaded from the National Center for Biotechnology Information database (NCBI Resource Coordinators 2018). 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*. Primary data resulted from oligonucleotide genotyping, whole genome and exome sequencing. Nine software packages were used to identify large-scale genomic variants including Breakdancer (Chen et al. 2009), Delly (Rausch et al. 2012), Variation Hunter (Hormozdiari et al. 2010), CNVnator (Abyzov et al. 2011), ReadDepth (Miller et al. 2011), Genome STRIP (Mills et al. 2011), Pindel (Ye et al. 2009), MELT (Gardner et al. 2017) and Dinumt (Dayama et al. 2014) and their call sets were merged. Selected variants 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% (1000 Genomes Project Consortium et al. 2015). Since a combination of filtering, calling and validation methods is a recommended approach to obtain reliable large-scale variants (Butty et al. 2020; Gabrielaite et al. 2021), we considered the 1000 Genomes Project Consortium calls as a *high confidence dataset*. In our study, from all available high confidence variants (copy number variants, indels, insertions, inversions and mobile elements), only CNVs defined as duplications or deletions were extracted. Overlapping CNVs were considered independently, resulting in 5867 tandem duplications and 33,181 deletions. Length of duplications ranged between 3006 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 3774 bp and mean of 12,143 ± 34,749 bp (Fig. 1).

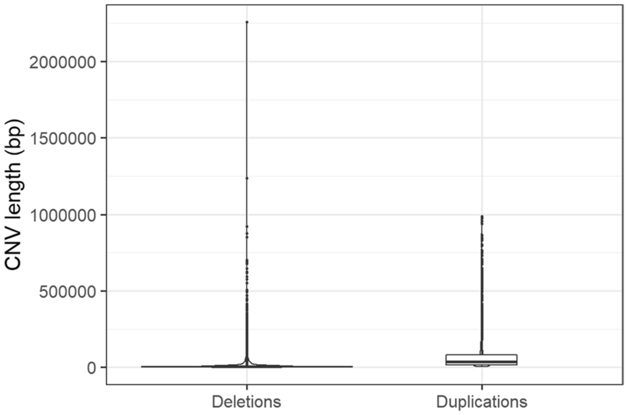

Fig. 1 Duplication and deletion length (bp)

Reference genome sequence features

The Samtools software (Li and Durbin 2009) was used to extract regions covered by CNVs from the GRCh38 reference genome. Moreover, coordinates of reference sequences flanking CNVs (100 nucleotides upstream and downstream of each deletion and duplication) were extracted. These regions were considered in the context of unknown nucleotides (denoted as “N”), Guanine-Cytosine pairs, sequence complexity and functional annotation. In order to compare regions covered by CNVs with random genomic sequences, we selected random region coordinates and extracted the regions from the reference genome, using the Samtools software. The process was repeated to match the actual numbers of analysed CNVs, so that four sets of random sequences were selected: (i) *set 1* contained 5859 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 (3774 bp); (iii) *set 3* contained 5867 sequences of 100-bp length and was used for comparisons with sequences upstream and downstream of duplications; (iv) *set 4* contained 33,181 sequences of 100-bp length and was used for comparison with sequences upstream and downstream of deletions. All sequences containing unknown nucleotides were excluded. 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 pair content of high confidence CNV-related sequences were compared with the corresponding randomised sets, i.e. high confidence duplications and set 1, high confidence deletions and set 2, flanking regions of high confidence duplications and set 3, and flanking regions of high confidence deletions and set 4. It was done using the Wilcoxon-Mann–Whitney test, with H_0 stating that the distributions of GC content are equal. The normalised Wilcoxon-Mann–Whitney test statistic is given by:

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

where $V = \sum_{j=1}^{m} S_j - \frac{m\bullet(m+1)}{2}$, S_j denotes ranks corresponding to the GC pair percentage classes in the random sequences, n is a count of deletion/duplication/flanking CNV regions and m is a count of sets with random sequences. Statistical analysis and figures were done in R package (R Core Team 2013).

Sequence complexity

Sequence complexity of the entire reference genome was estimated using the sDust software (Morgulis et al. 2006). The overlap between low-complexity regions defined by sDust and CNV-related regions was determined by using the bedtools software (Quinlan and Hall 2010) for high confidence 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 the same way as GC pair content by using the Wilcoxon-Mann–Whitney test.

Functional annotation

The Variant Effect Predictor (VEP) software (McLaren et al. 2016) was used for the functional annotation of CNVs. Gene Ontology enrichment (Mi et al. 2019) was tested using the Fisher's exact test with the false discovery rate (FDR). Moreover, significantly enriched signalling pathways from the Panther (Mi et al. 2019) database were identified using the KOBAS tool (Xie et al. 2011) applying the Fisher's exact test with FDR.

Results

Reference genome sequence features.

Unknown nucleotide (N) content

Among all of the regions of the human reference genome GRCh38 (Schneider et al. 2017) covered by CNVs, only eight duplications (Supplementary Information S1) and six deletions (Supplementary Information S2) contained unknown nucleotides. Percentage of unknown nucleotides in duplications ranged between 0.002 and 22.06% and in deletions, it varied between 0.0004 and 63.21%. Note that in three deletions, unknown nucleotides constituted over 50% of the whole length. Some regions contained a fixed number of unknown nucleotides (i.e. 100 or 50,000 Ns); what represents the fact that the actual number of unknown nucleotides in the reference cannot be determined. In regions flanking CNVs, only one sequence, located upstream of deletion, contained 17 unknown nucleotides (chromosome: 18, start: 9984, end: 10,083).

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% ± 5.83) and deletions (41.08% ± 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% (Table 1). The visual examination of GC pair content distributions in CNVs, presented on Figs. S3a and S4a 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 (Figs. S5a, S5b, S6a and S6b), indicating no link between CNV breakpoint formation and the GC content. The distributions of GC pair contents of duplications ($P = 0.004$) and deletions ($P = 7.955 \cdot 10^{-12}$) significantly differed from the contents of corresponding randomised sequences. In particular, high confidence deletions contained less GC pairs than random regions ($P = 3.977 \cdot 10^{-12}$), while duplications were enriched in GC pairs as compared to a randomised set of sequences ($P = 0.0024$). High confidence deletion flanking regions contained more GC pairs than the corresponding randomised sequences, i.e. $P = 1.5 \cdot 10^{-10}$ for upstream and $P = 1.259 \cdot 10^{-21}$ for downstream regions. The same situation was observed for downstream region of duplications ($P = 1.74 \cdot 10^{-9}$), but for upstream, it was lower than in random case ($P = 0.014$). graphical representation of Randomised duplications (Fig. S3b), deletions (Fig. S4b) and their flanking regions (Figs. S5c and S6c) GC content distributions are provided in the supplementary material.

Sequence complexity

A total of 4,798,406 low-complexity regions (LCRs) were identified within the whole GRCh38 reference genome. Lengths of those regions varied between 7 and 25,072 bp with mean of 29 bp (± 56). All duplications and 93.93% of deletions contained within LCRs. Median number of LCRs overlapped with a single duplication was 57 and

Table 1 Guanine-cytosine pair content (%) in the investigated regions

Region	Min	Mean	Max	Sd
Duplications	**29.08**	**41.86**	**68.90**	**5.83**
Set 1 (randomised 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 (randomised upstream and downstream duplications)	1.00	41.42	84.00	10.54
Deletions	**21.27**	**41.08**	**73.46**	**6.15**
Set 2 (randomised 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 (randomised upstream and downstream deletions)	0.00	41.41	89.00	10.66

with a single deletion was six (Fig. 2, Table 2). On average, LCRs made up 4.59% of a duplication length and 4.66% of a deletion length (Fig. 3, Table 2). On the other hand, CNV breakpoint regions contained much less LCRs. Only, 20.83% of sequences upstream of duplications and 16.52% of sequences downstream of duplications contained a low-complexity region(s). Similarly among deletion breakpoints, we identified 20.37% of upstream sequences and 19.25% of downstream sequences with LCR. 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%. None of the empirically constructed frequency distributions in the considered regions deviated from the normal distribution. The distributions of low-complexity sequence content in randomised duplications and in high confidence duplications ($P=0.106$) as well as between randomised downstream deletions and regions downstream of deletions ($P=0.078$) did not differ. The percentage of low-complexity sequences was significantly higher upstream of deletions ($P=1.907\cdot10^{-8}$), upstream of duplications ($P=8.982\cdot10^{-5}$) and within deletions ($P=2.963\cdot10^{-19}$) than in corresponding randomised upstream and downstream regions. Conversely, the distribution of low-complexity sequence content downstream of duplications was significantly lower than in set 3 ($P=0.007$).

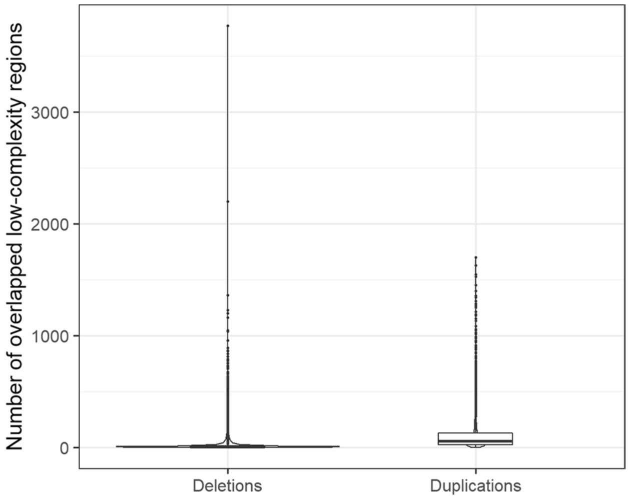

Fig. 2 The number of LCRs overlapped duplications and deletions

Functional annotation of CNVs

The 5867 duplications overlapped with 9111 genes corresponding to 35,317 transcripts. The 33,181 deletions overlapped with 19,022 genes corresponding to 71,542 transcripts. The most common Sequence Ontology (SO) terms determined for **duplications** comprised intron variants (SO:0,001,627) and transcript amplifications (SO:0,001,889) (Table 3). For **deletions**, the most common SO terms were as follows: feature truncations (SO:0,001,906) and intron variants (SO:0,001,627) (Table 3). In the context of biological processes, genes containing duplications were significantly overrepresented in ontologies related to homophilic cell adhesion via plasma membrane adhesion molecules (GO:0,007,156), modulation of chemical synaptic transmission (GO:0,050,804), cytoskeleton organization (GO:0,007,010) and in the Cadherin signalling pathway (P00012) as well as underrepresented for complement activation, classical pathway (GO:0,006,958), including immune response (GO:0,006,955). Genes containing deletions were significantly overrepresented in GO terms related

Table 2 Content of low-complexity regions (LCR) within CNV-related regions

Regions	Number of overlapped LCRs			Content of LCRs (%)		
	Min	Mean	Max	Min	Mean	Max
Duplications	**1**	**104**	**1 698**	**0.07**	**4.59**	**47.52**
Set 1 (randomised 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 (randomised upstream 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 (randomised 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 (randomised upstream and downstream deletions)	0	0	4	0.00	4.36	100.00

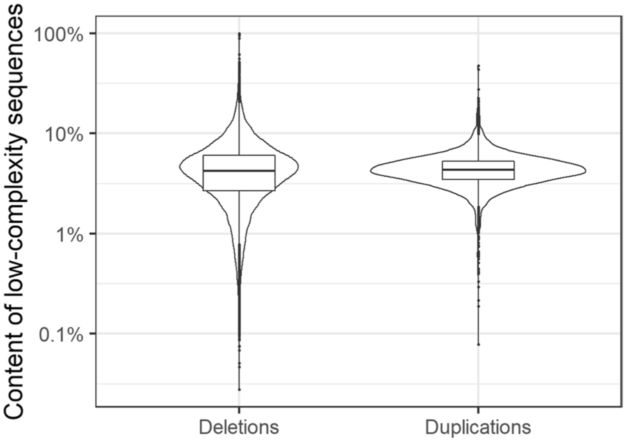

Fig. 3 Low-complexity sequence content in duplications and deletions (CNVs not overlapping any region are not presented)

to transport (GO:0,006,810), cellular component organization (GO:0,016,043) and regulation of cellular processes (GO:0,050,794).

Table 3 Functional annotation of duplications and deletions

Consequences for duplications	SO accession	Percent of variants
Intron variant	SO:0,001,627	25
Transcript amplification	SO:0,001,889	24
Coding sequence variant	SO:0,001,580	12
Feature elongation	SO:0,001,907	8
Non-coding transcript exon variant	SO:0,001,792	8
5-prime UTR variant	SO:0,001,623	7
3-prime UTR variant	SO:0,001,624	6
Upstream gene variant	SO:0,001,631	3
Downstream gene variant	SO:0,001,632	3
NMD transcript variant	SO:0,001,621	2
Non-coding transcript variant	SO:0,001,619	1
Other	-	1
Consequences for deletions	**SO accession**	**Percent of variants**
Feature truncation	SO:0,001,906	31
Intron variant	SO:0,001,627	31
Non-coding transcript variant	SO:0,001,619	10
Upstream gene variant	SO:0,001,631	4
Non-coding transcript exon variant	SO:0,001,792	4
Downstream gene variant	SO:0,001,632	4
Transcript ablation	SO:0,001,893	4
Intergenic variant	SO:0,001,628	3
Coding sequence variant	SO:0,001,580	3
NMD transcript variant	SO:0,001,621	3
5-prime UTR variant	SO:0,001,623	1
3-prime UTR variant	SO:0,001,624	1
Stop lost	SO:0,001,578	1

Discussion

Our study revealed a non-random distribution of GC pairs within CNVs and in CNV flanking regions. This could have been expected, following the hypothesis that the GC content serves as a tool for differentiation between intron (lower GC content) and exons (higher GC content) during splicing (Amit et al. 2012). In our study, we observed that the GC content of deletions was lower and of duplications — higher than in random genomic regions, what indicates that intronic regions are more prone to deletions, whereas exonic regions are more prone to duplications. However, a contradictory result was obtained for humans by Rigau et al. (2019) who observed that deleted regions had significantly higher GC content. In our study, the majority of deletions was annotated to introns what further supports the GC content imbalance (Aïssani and Bernardi 1991). Deletions in genic regions, containing more GC, are functionally more severe than duplications. Moreover, duplications which associate with GC-rich regions (i.e. exons) have some evolutionary advantage (Levasseur and Pontarotti 2011). It is also worth to notice that according to Dittwald et al. (2013), GC content is positively correlated with the frequency of nonallelic homologous recombination (NAHR) which is a common cause of CNV formation. According to Romiguier et al. (2010), GC-rich sequences are prone to deletions because base composition imbalance triggers replication slippage. On the other hand, Chen et al. (2011) did not report a difference in GC content between CNV regions and autosomal average. Our study also demonstrated a non-random GC content in CNV flanking regions, albeit without a consistent trend, i.e. enriched GC content in deletion breakpoints, but only downstream of duplications. Similarly, Bose et al. (2014) investigated SNV breakpoints and concluded that all SV types had a higher GC percentage than the genome average.

Also in terms of sequence complexity, a non-random pattern was revealed, with deletions being enriched with LCR, but without a consistent pattern in breakpoint regions. Barski et al. (2019) 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 (Nguyen et al. 2006; Monlong et al. 2018), all of those characterised by low complexity. Chen et al. (2014) 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 in introns. Similar observation was made by Chen et al. (2011) for population-specific CNVs. Higher gene density in regions covered by CNVs than in random genome regions was also highlighted by Johansson and Feuk (2011). Moreover, Nguyen et al. (2006) reasoned that large-scale DNA changes, if beneficial, they should be enriched in genes, especially those involved in fighting infection and sensing our environment. According to Rigau et al. (2019), 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. (2015). Moreover, Morello et al. (2019) 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 has already been reported by Perry et al. (2008) and, the same as in our study, genes with immune response functions were overrepresented in human CNV regions (Redon et al. 2006). 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 (Alloza et al. 2011). Duplicated genes were enriched in the Cadherin signalling pathway, which is involved in multiple biological processes, such as development, neurogenesis, cell adhesion and inflammation. Its enrichment has been reported in the context of many diseases including cancer (Mi et al. 2019).

In conclusions, genomic regions containing large-scale duplications and deletions, called copy number variants (CNVs), constitute a common source of genetic variation. In this study, we analysed duplications and deletions identified within the frame of the 1000 Genomes Project, in the context of identification of the unique DNA sequence features in CNV regions and of annotation of CNVs to functional segments of the human genome. We discovered that (i) guanine-cytosine content is associated with the formation of CNVs; (ii) duplications are initiated in low-complexity regions and (iii) CNVs are preferentially located within introns. Our findings provide a step towards more complete understanding of the human genomic landscape in the context of copy number variants.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13353-022-00704-0.

Acknowledgements We acknowledge Poznan Supercomputing and Networking Centre for hosting the computations. The APC (Article Processing Charge) is financed by Wroclaw University of Environmental and Life Sciences.

Author contribution 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.

Data availability All data were downloaded from the publicly available databases.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent to publish Not applicable.

Competing interests The authors declare no competing interests.

References

1000 Genomes Project Consortium, Auton A, Abecasis GR et al (2015) A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393

Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21:974–984. https://doi.org/10.1101/gr.114876.110

Aïssani B, Bernardi G (1991) CpG islands, genes and isochores in the genomes of vertebrates. Gene 106:185–195. https://doi.org/10.1016/0378-1119(91)90198-k

Alloza E, Al-Shahrour F, Cigudosa JC, Dopazo J (2011) A large scale survey reveals that chromosomal copy-number alterations significantly affect gene modules involved in cancer initiation and progression. BMC Med Genomics 4:37. https://doi.org/10.1186/1755-8794-4-37

Amit M, Donyo M, Hollander D et al (2012) Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep 1:543–556. https://doi.org/10.1016/j.celrep.2012.03.013

Barski P, Mielczarek M, Frąszczak M SJ (2019) DNA sequence features underlying copy number variants. Acta Sci Pol Zootech 25–30

Berglund J, Nevalainen EM, Molin A-M et al (2012) Novel origins of copy number variation in the dog genome. Genome Biol 13:R73. https://doi.org/10.1186/gb-2012-13-8-r73

Bose P, Hermetz KE, Conneely KN, Rudd MK (2014) Tandem repeats and G-rich sequences are enriched at human CNV breakpoints. PLoS ONE 9:e101607

Butty AM, Chud TCS, Miglior F et al (2020) High confidence copy number variants identified in Holstein dairy cattle from whole genome sequence and genotype array data. Sci Rep 10:8044. https://doi.org/10.1038/s41598-020-64680-3

Chen K, Wallis JW, McLellan MD et al (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6:677–681. https://doi.org/10.1038/nmeth.1363

Chen L, Zhou W, Zhang L, Zhang F (2014) Genome architecture and its roles in human copy number variation. Genomics Inform 12:136–144. https://doi.org/10.5808/GI.2014.12.4.136

Chen W, Hayward C, Wright AF et al (2011) Copy number variation across European populations. PLoS ONE 6:e23087–e23087. https://doi.org/10.1371/journal.pone.0023087

Dayama G, Emery SB, Kidd JM, Mills RE (2014) The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res 42:12640–12649. https://doi.org/10.1093/nar/gku1038

Dittwald P, Gambin T, Szafranski P et al (2013) NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. Genome Res 23:1395–1409. https://doi.org/10.1101/gr.152454.112

Gabrielaite M, Torp MH, Rasmussen MS, et al (2021) A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. Cancers (Basel) 13https://doi.org/10.3390/cancers13246283

Gardner EJ, Lam VK, Harris DN et al (2017) The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res 27:1916–1929. https://doi.org/10.1101/gr.218032.116

Hormozdiari F, Hajirasouliha I, Dao P et al (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 26:i350–i357. https://doi.org/10.1093/bioinformatics/btq216

Huckaby AC, Granum CS, Carey MA et al (2019) Complex DNA structures trigger copy number variation across the Plasmodium falciparum genome. Nucleic Acids Res 47:1615–1627. https://doi.org/10.1093/nar/gky1268

Johansson ACV, Feuk L (2011) Characterization of copy number-stable regions in the human genome. Hum Mutat 32:947–955. https://doi.org/10.1002/humu.21524

Kasak L, Rull K, Vaas P et al (2015) Extensive load of somatic CNVs in the human placenta. Sci Rep 5:8342. https://doi.org/10.1038/srep08342

Levasseur A, Pontarotti P (2011) The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. Biol Direct 6:11. https://doi.org/10.1186/1745-6150-6-11

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Makino T, McLysaght A, Kawata M (2013) Genome-wide deserts for copy number variation in vertebrates. Nat Commun 4:2283. https://doi.org/10.1038/ncomms3283

McLaren W, Gil L, Hunt SE et al (2016) The Ensembl Variant Effect Predictor. Genome Biol 17:122. https://doi.org/10.1186/s13059-016-0974-4

Mi H, Muruganujan A, Ebert D et al (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res 47:D419–D426. https://doi.org/10.1093/nar/gky1038

Miller CA, Hampton O, Coarfa C, Milosavljevic A (2011) ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. PLoS ONE 6:e16327. https://doi.org/10.1371/journal.pone.0016327

Mills RE, Walter K, Stewart C et al (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470:59–65. https://doi.org/10.1038/nature09708

Monlong J, Cossette P, Meloche C et al (2018) Human copy number variants are enriched in regions of low mappability. Nucleic Acids Res 46:7236–7249. https://doi.org/10.1093/nar/gky538

Morello G, Guarnaccia M, Spampinato AG et al (2019) Integrative multi-omic analysis identifies new drivers and pathways in molecularly distinct subtypes of ALS. Sci Rep 9:9968. https://doi.org/10.1038/s41598-019-46355-w

Morgulis A, Gertz EM, Schäffer AA, Agarwala R (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol 13:1028–1040. https://doi.org/10.1089/cmb.2006.13.1028

NCBI Resource Coordinators (2018) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 46:D8–D13. https://doi.org/10.1093/nar/gkx1095

Nguyen D-Q, Webber C, Ponting CP (2006) Bias of selection on human copy-number variants. PLOS Genet 2:e20

Nicholas TJ, Cheng Z, Ventura M et al (2009) The genomic architecture of segmental duplications and associated copy number variants in dogs. Genome Res 19:491–499. https://doi.org/10.1101/gr.084715.108

Perry GH, Yang F, Marques-Bonet T et al (2008) Copy number variation and evolution in humans and chimpanzees. Genome Res 18:1698–1710. https://doi.org/10.1101/gr.082016.108

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team (2013) R: a language and environment for statistical computing

Rausch T, Zichner T, Schlattl A et al (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Redon R, Ishikawa S, Fitch KR et al (2006) Global variation in copy number in the human genome. Nature 444:444–454. https://doi.org/10.1038/nature05329

Rigau M, Juan D, Valencia A, Rico D (2019) Intronic CNVs and gene expression variation in human populations. PLoS Genet 15:e1007902–e1007902. https://doi.org/10.1371/journal.pgen.1007902

Romiguier J, Ranwez V, Douzery EJP, Galtier N (2010) Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res 20:1001–1009. https://doi.org/10.1101/gr.104372.109

Schneider VA, Graves-Lindsay T, Howe K et al (2017) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res 27:849–864. https://doi.org/10.1101/gr.213611.116

Xie C, Mao X, Huang J et al (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res 39:W316–W322. https://doi.org/10.1093/nar/gkr483

Ye K, Schulz MH, Long Q et al (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865–2871. https://doi.org/10.1093/bioinformatics/btp394

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.