Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 253 – 261
© TÜBİTAK
doi:10.3906/elk-1903-135

Automatic characterization of copy number polymorphism using high throughput sequencing

Can ALKAN[*]
Department of Computer Engineering, Faculty of Engineering, Bilkent University, Bilkent, Ankara, Turkey

Received: 22.03.2019 • **Accepted/Published Online:** 25.07.2019 • **Final Version:** 27.01.2020

Abstract: Genome structural variation, broadly defined as alterations longer than 50 bp, are important sources for genetic variation among humans, including those that cause complex diseases such as autism, developmental delay, and schizophrenia. Although there has been considerable progress in characterizing structural variation since the beginnings of the 1000 Genomes Project, one form of structural variation called segmental duplications (SDs) remained largely understudied in large cohorts. This is mostly because SDs cannot be accurately discovered using the alignment files generated with standard read mapping tools. Instead, they can only be found when multiple map locations are considered. There is still a single algorithm available for SD discovery, which includes various tools and scripts that are not portable and are difficult to use. Additionally, this algorithm relies on a priori information for regions where no structural variations are discovered in large number of genomes. Therefore, there is a need for fully automated, portable, and user-friendly tools to make SD characterization a part of genome analyses. Here we introduce such an algorithm and efficient implementation, called mrCaNaVaR, that aims to fill this gap in genome analysis toolbox.

Key words: Genomics, copy number polymorphism, whole genome sequencing, containers

1. Introduction

The changes in DNA sequences are classified depending on their size and organization. The smallest form of genomic variation, called single nucleotide variation (SNV), are single basepair substitutions between two segments of DNA sequences [1], typically called sample and reference. There can be also insertions and deletions of short sequences (1–50 bp), named indel polymorphisms [2]. Other forms of genomic variation include expansion and contraction of short tandem repeats (microsatellite polymorphisms) [3], balanced rearrangements such as inversions [4] and translocations [5], and copy number variation (CNV) [6]. CNVs, by definition, alter the amount of DNA material in cells, and they can be deletions, insertions, and duplications of genomic segments [7], as well as mobile element retrotranspositions [8]. The 1000 Genomes Project that ran between 2008 and 2015 generated the most comprehensive map of genomic variation in the genomes of 2504 individuals from 26 populations, aiming to characterize genetic diversity within the human species [9–11].

CNVs, insertions, mobile element retrotranspositions, inversions, and translocations are also collectively referred to as structural variations [12]. Interest in characterizing structural variations (SVs) has increased dramatically since the inception of high-throughput sequencing (HTS) [13], especially when it is understood that SNVs do not explain the heritability of most complex diseases alone [14]. However, the fact that SVs are typically in repeat-rich regions of the genome [15] makes it very difficult to accurately discover and genotype

[*]Correspondence: calkan@cs.bilkent.edu.tr

SVs. There are a plethora of algorithms that aim to characterize SVs that use one of the four sequence mapping signatures, namely, read pair, read depth, split reads, and assembly [16]. Some of the earlier methods that were also used in the 1000 Genomes Project [11] include VariationHunter [17, 18] (read pair), CNVnator [19] (read depth), Pindel [20] (split reads), and NovelSeq [21] (assembly). More modern algorithms try to integrate multiple sequence signatures and characterize several forms of SVs, such as TARDIS [22, 23], LUMPY [24], TIDDIT [25], and Manta [26].

Among the CNVs, segmental duplications (SDs) are defined as large segments of DNA (typically >1 kbp) that are copied to other regions of the genome at high sequence identity ($>90\%$) [27]. Similarly, deletions are losses of DNA that stretches longer than 50 bp [17]. Both CNV types have been associated with various human diseases [28–30] and shown to play important roles in great ape evolution [31, 32]. Although characterization of deletions are now regarded as a solved problem [11, 33, 34], discovering segmental duplications and predicting copy numbers of genes is omitted in most large scale studies due to the additional computational costs. Any read that originates from a duplicated segment aligns to multiple regions of the reference genome [35]; however, the standard read mapping process in HTS analysis keeps only one mapping selected randomly among other potential map locations [36]. Therefore, SD discovery is not part of routine genome analyses, and performed only when required by the specifics of the project at hand.

Currently there is only one algorithm publicly available to characterize SDs [37]. It has been used to analyze genomes of humans [38], great apes [39], cattle [40], water buffalo [41], dogs [42], cats [43], and grapevine [44]. However, in each of these studies, the existing code needed to be modified for the genome at hand, and the entire analysis pipeline was composed of multiple steps that were not streamlined, often requiring manual intervention. The lack of a fully automated and streamlined SD characterization tool made it difficult to fully understand the copy number spectrum in many organisms.

Here we introduce mrCaNaVaR, a comprehensive tool for SD characterization and copy number prediction. Different from its predecessor implementation [37], it is straightforward to apply to nonhuman genomes as it requires only a reference genome and the coordinates for the assembly gaps that are released as part of draft and finished genome assemblies. For read mapping, it uses the mrsFAST aligner that we have previously developed specifically for tracking multiple map locations for accurate CNV discovery [45]. mrCaNaVaR can take as input raw FASTQ files, or alignment files in BAM [46] or CRAM [47] format generated with any read mapper such as BWA-MEM [48]. arXiv 2013; arXiv:13033997. and Bowtie2 [49]. We provide BAM and CRAM file support to enable mrCaNaVaR use for data sets where the raw FASTQ files are deleted after alignment. We also package both mrCaNaVaR and mrsFAST into a single Docker container to enable seamless portability to different environments, operating systems, and cloud infrastructures. Source code, Docker file, and workflow scripts for mrCaNaVaR are available at https://github.com/BilkentCompGen/mrcanavar and a prebuilt Docker image is available at https://hub.docker.com/r/alkanlab/mrcanavar.

2. Materials and methods

The main algorithm behind mrCaNaVaR was previously described [37]; however, we introduce new improvements (Figure). Briefly, using the map locations generated by the mrsFAST aligner, the algorithm first calculates the read depth distribution over diploid control regions (no known CNVs). It then applies the LOESS smoothing technique [50] to correct for sequencing biases related to regions with high and low G+C content with respect to A+T [51]. The resulting distribution is expected to be Poisson [37], and any deviations from the distribution flags potential duplicated or deleted segments. Next, the algorithm merges CNV regions that overlap to report putative duplications and deletions. Finally, it calculates genome-wide copy numbers over nonoverlapping intervals of 1 kb, by simply dividing the read depth over each region by the average read depth in control regions.

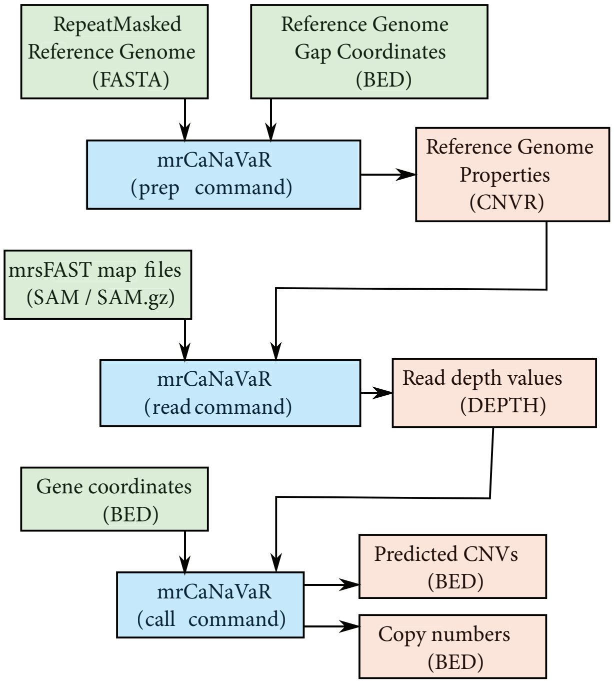
RepeatMasked Reference Genome (FASTA)
Reference Genome Gap Coordinates (BED)
mrCaNaVaR (prep command)
Reference Genome Properties (CNVR)
mrsFAST map files (SAM / SAM.gz)
mrCaNaVaR (read command)
Read depth values (DEPTH)
Gene coordinates (BED)
mrCaNaVaR (call command)
Predicted CNVs (BED)
Copy numbers (BED)

Figure 1: Steps of calling CNVs and copy numbers using mrCaNaVaR. We first generate a file (extension `.cnvr`) that contains G+C ratios within overlapping sliding windows using `mrcanavar prep` command. This step is needed only once for each new reference genome, and the resulting `.cnvr` file can be reused for other samples. Next, `mrcanavar read` command loads the `SAM/SAM.gz` files generated by mrsFAST for the sample to be analyzed and saves the G+C normalized read depth values in a native file format with `.depth` extension. Finally, `mrcanavar call` command loads the `.depth` file and optionally the list of gene coordinates to predict copy numbers and CNVs (both deletions and duplications). The additional `mrcanavar-auto` wrapper merges all of these steps into a single command that also includes mrsFAST mapping.

Here we introduce the following improvements:

1. The existing implementation requires a set of genomic regions to be passed as input. This makes it especially difficult to characterize CNVs in nonhuman genomes, and even with newer versions of the human reference genomes, as obtaining a set of diploid regions is a cumbersome task since one must collect all known CNVs in the genomes of a given species. Additionally, the set of known CNVs is not necessarily complete; genome of any given individual might harbor additional CNVs, which might skew the expected Poisson distribution. The new mrCaNaVaR bypasses the need for control regions, by simply analyzing the read depth over the entire genome, removes the outlier regions with too high or too low read depth until a Poisson distribution is achieved.

2. The previous version is made up of several shell scripts, Perl scripts, and C programs that are manually run one after the other. mrCaNaVaR is a fully-integrated single program implemented in C, which makes it very easy for nonexperts to analyze large number of genomes.

3. We also provide a wrapper for mrCaNaVaR, which also automates the read mapping step with mrsFAST, and feeds the read map locations to mrCaNaVaR without any user intervention. The wrapper also supports BAM and CRAM files. We also modified mrsFAST code to support standard input and output ports to remove the necessity of creating large files. This is especially important to decrease the storage requirements and helps reduce the costs when cloud platforms are used.

4. Finally, we packaged the entire pipeline into a single Docker container, and we additionally provide workflow scripts written in Common Workflow Language [52].

We implemented the entire mrCaNaVaR package in ANSI C language without the use of any external libraries.

3. Results

We tested mrCaNaVaR in two directions. First, we calculated the computational requirements of the entire mrCaNaVaR run, including mrsFAST remapping using three different data sets at different depth of coverage values. Next, we tested mrCaNaVaR on nine human genome data sets, previously sequenced and analyzed by the Human Genome Structural Variation Consortium [53] using various long and short read sequencing technologies, optical mapping [54], and StrandSeq protocol [55].

3.1. Computational requirements

We tested the analysis pipeline using three human genome data sets at three different depths of coverage. We obtained the HG00096 genome data set (5× coverage) from the 1000 Genomes Project [11], the LP6005441 genome (42× coverage) from the Simons Genome Diversity Project [56], and the HG00512 genome (82× coverage) from the Human Genome Structural Variation Consortium [53].

Table 1 summarizes the run time and memory requirements of characterizing CNVs and predicting copy numbers using mrsFAST and mrCaNaVaR. We ran all experiments on the Amazon Web Services (AWS) cloud using a `c5.4xlarge` *spot* instance. Table 1 also lists the end-to-end running costs of each experiment. Human genome whole genome sequence data is typically generated at ≈30–35× depth of coverage; therefore, the LP6005441 experiment demonstrates the most common scenario. We observed that, as expected, most of the run time and the memory requirements are spent at mrsFAST remapping. On the other hand, mrCaNaVaR needed less than 10 min even with the largest data set and used less than 177 MB main memory in the worst case. Overall, the average cost (42× coverage) for mrsFAST and mrCaNaVaR analysis on the AWS cloud was $2.05. We note that this cost may be further reduced by adjusting the number of threads and peak memory usage for the mrsFAST read mapping step.

3.2. Biological sample comparisons

In our second test we aimed to assess the prediction performance of mrCaNaVaR. Previously we had shown the accuracy of characterizing segmental duplications and gene copy numbers in three genomes together with experimental validation [37]. Here, we analyzed the genomes of nine human individuals from three populations and compared our predictions with orthogonal results obtained from multiple genome sequencing and optical mapping techniques.

We downloaded the human genome data sets from the Human Genome Structural Variation (HGSV) Consortium [53]. The depth of coverage of all samples were in the range 80–82×. We also downloaded the

Table 1: Computational requirements of mrsFAST and mrCaNaVaR.

		mrsFAST		mrCaNaVaR		
Sample	Coverage	Time	Peak memory	Time	Peak memory	AWS cost
HG00096 [11]	5×	0 h 35 min	20 GB	3 min 14 s	138 MB	$0.32
LP6005441 [56]	42×	3 h 58 min	20 GB	9 min 47 s	177 MB	$2.05
HG00512 [53]	82×	9 h 54 min	28 GB	11 min 06 s	138 MB	$5.12

We ran all experiments on an AWS `c5.4xlarge` instance using 16 threads for mrsFAST mapping (Intel Xeon Platinum 8000 processor at clock speed of up to 3.5 GHz). mrCaNaVaR is a single-threaded application. Note that the memory requirements of mrsFAST aligner can be adjusted using the `--mem` parameter. AWS: Amazon Web Services.

structural variation call sets generated by the consortium using the same samples from database of large scale genomic variants [57] (dbVar ID: nstd152[1]).

We remapped the reads to the human reference genome (GRCh38) using mrsFAST, and we then used mrCaNaVaR to call segmental duplications and large deletions (both >10 kb). As expected, most of the segmental duplications (>90%) we detected were common; i.e. present in the reference genome [27, 58, 59]). The HGSV project did not specifically aim to characterize large segmental duplications; however, the call set included an average of 760 kb of segmental duplications larger than 10 kb per genome that are not present in the reference genome. We observed that mrCaNaVaR called 20%–40% of such SDs. Additionally, we identified an average of 17 Mb of SDs per genome. We summarize our findings in Table 2. Similarly, we compared our deletion predictions with the HGSV call set. In total we predicted 150 kb to 2 Mb of deletions in the genomes we analyzed and compared with the deletion calls released by the HGSV Consortium.

Table 2: Segmental duplication and large deletion prediction results from 9 human genomes.

	Segmental duplications (>10 kb)				Deletions (>10 kb)		
Sample	Predicted	Common SDs	Known [53]	Novel	Predicted	Known [53]	Novel
HG00512	212.4 Mb	192.5 Mb	0.183 Mb	19.7 Mb	1.8 Mb	0.392 Mb	1.4 Mb
HG00513	181.5 Mb	163.5 Mb	0.285 Mb	17.7 Mb	0.5 Mb	0.440 Mb	0.06 Mb
HG00514	181.3 Mb	163.5 Mb	0.140 Mb	17.7 Mb	0.4 Mb	0.352 Mb	0.05 Mb
HG00731	211.5 Mb	192.4 Mb	0.191 Mb	18.9 Mb	1.8 Mb	0.151 Mb	1.65 Mb
HG00732	187.3 Mb	170.1 Mb	0.148 Mb	17 Mb	0.2 Mb	0.085 Mb	0.11 Mb
HG00733	184.3 Mb	166.9 Mb	0.086 Mb	17.3 Mb	0.15 Mb	0.151 Mb	0 Mb
NA19238	188.4 Mb	171.5 Mb	0.415 Mb	16.5 Mb	0.22 Mb	0.184 Mb	0.04 Mb
NA19239	201.3 Mb	182.7 Mb	0.179 Mb	18.4 Mb	2 Mb	0.138 Mb	1.93 Mb
NA19240	186.3 Mb	169.1 Mb	0.378 Mb	16.8 Mb	0.15 Mb	0.164 Mb	0 Mb

We obtained all genomes and known CNV call sets from [53]. Here we compare SDs with the reference genome and the known CNV calls. We only report the intersection of nonreference SDs with those reported in [53]. All samples have ≈ 82× depth of coverage.

[1] Chaisson et al. (2019). Human Genome Structural Variation Consortium Call Set [online]. Website https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd152/ [accessed 15 March 2019].

4. Conclusion

In this paper we introduced mrCaNaVaR, an extension to our previous algorithm to characterize large CNVs. The implementation of mrCaNaVaR is also completely novel, efficient, and user-friendly. We also built a Docker image that contains all dependencies, including a modified version of the mrsFAST aligner. Finally, we provide workflow scripts written in the Common Workflow Language (CWL) that can be run using any CWL execution tool[2]. Especially Docker and CWL provide straightforward portability to cloud environments, which in turn will make SD characterization a routine part of genome analyses.

Acknowledgment

We thank Y. Küçük for his help in coding parts of the mrCaNaVaR tool. This study is supported by a Marie Curie Career Integration Grant (PCIG-2011–303772) to C.A. under the Seventh Framework Programme. We also acknowledge support from The Science Academy of Turkey, under the BAGEP program.

References

[1] Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 2011; 12 (6): 443-451. doi: 10.1038/nrg2986

[2] Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Research 2006; 16 (9): 1182-1190. doi: 10.1101/gr.4565806

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

[4] Eslami Rasekh M, Chiatante G, Miroballo M, Tang J, Ventura M et al. Discovery of large genomic inversions using long range information. BMC Genomics 2017; 18 (1): 65. doi: 10.1186/s12864-016-3444-1

[5] Talkowski ME, Ernst C, Heilbut A, Chiang C, Hanscom C et al. Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. American Journal of Human Genetics 2011; 88 (4): 469-481. doi: 10.1016/j.ajhg.2011.03.013

[6] Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK et al. Detection of large-scale variation in the human genome. Nature Genetics 2004; 36 (9): 949-951. doi: 10.1038/ng1416

[7] Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA et al. Copy number variation: new insights in genome diversity. Genome Research 2006; 16 (8): 949-961. doi: 10.1101/gr.3677206

[8] Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nature Reviews Genetics 2002; 3 (5): 370-379. doi: 10.1038/nrg798

[9] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 2010; 467 (7319): 1061-1073. doi: 10.1038/nature09534

[10] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491 (7422): 56-65. doi: 10.1038/nature11632

[11] The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015; 526 (7571): 68-74. doi: 10.1038/nature15393

[12] Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA et al. Fine-scale structural variation of the human genome. Nature Genetics 2005; 37 (7): 727-732. doi: 10.1038/ng1562

[2] Common Workflow Language Working Group (2019). Common Workflow Language Implementations [online]. Website https://www.commonwl.org/ [accessed 10 March 2019].

[13] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nature Reviews Genetics 2011; 12 (5): 363-376. doi: 10.1038/nrg2958

[14] Eichler EE, Flint J, Gibson G, Kong A, Leal SM et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics 2010; 11 (6): 446-450. doi: 10.1038/nrg2809

[15] Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N et al. Mapping and sequencing of structural variation from eight human genomes. Nature 2008; 453 (7191): 56-64. doi: 10.1038/nature06862

[16] Medvedev P, Brudno M. Ab initio whole genome shotgun assembly with mated short reads. In: RECOMB 2008 International Conference on Research in Computational Molecular Biology; Singapore; 2008. pp. 50-64. doi: 10.1007/978-3-540-78839-3_5

[17] Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Research 2009; 19 (7): 1270-1278. doi: 10.1101/gr.088633.108

[18] Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D et al. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 2010; 26 (12): i350-i357. doi: 10.1093/bioinformatics/btq216

[19] Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research 2011; 21 (6): 974-984. doi: 10.1101/gr.114876.110

[20] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009; 25 (21):2865-2871. doi: 10.1093/bioinformatics/btp394

[21] Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I et al. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics 2010; 26 (10): 1277-1283. doi: 10.1093/bioinformatics/btq152

[22] Soylev A, Kockan C, Hormozdiari F, Alkan C. Toolkit for automated and rapid discovery of structural variants. Methods 2017; 129: 3-7. doi: 10.1016/j.ymeth.2017.05.030

[23] Soylev A, Le T, Amini H, Alkan C, Hormozdiari F. Discovery of tandem and interspersed segmental duplications using high throughput sequencing. Bioinformatics 2019 (in press). doi: 10.1093/bioinformatics/btz237

[24] Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biology 2014 15 (6): R84. doi: 10.1186/gb-2014-15-6-r84

[25] Eisfeldt J, Vezzi F, Olason P, Nilsson D, Lindstrand A. TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. F1000Research 2017; 6: 664. doi: 10.12688/f1000research.11168.2

[26] Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 2016; 32 (8): 1220-1222. doi: 10.1093/bioinformatics/btv710

[27] Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental duplications: organization and impact within the current human genome project assembly. Genome Research 2001; 11 (6): 1005-1017. doi: 10.1101/gr.187101

[28] Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. American Journal of Human Genetics 2007; 80 (6): 1037-1054. doi: 10.1086/518257

[29] Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annual Review of Medicine 2010; 61: 437-455. doi: 10.1146/annurev-med-100708-204735

[30] Girirajan S, Dennis MY, Baker C, Malig M, Coe BP et al. Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. American Journal of Human Genetics 2013; 92 (2): 221-237. doi: 10.1016/j.ajhg.2012.12.016

[31] Bailey JA, Eichler EE. Primate segmental duplications: crucibles of evolution, diversity and disease. Nature Reviews Genetics 2006; 7 (7): 552-564. doi: 10.1038/nrg1895

[32] Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW et al. Evolution and diversity of copy number variation in the great ape lineage. Genome Research 2013; 23 (9): 1373-1382. doi: 10.1101/gr.158543.113

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

[34] Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N et al. Global diversity, population stratification, and selection of human copy-number variation. Science 2015; 349 (6253): aab3761. doi: 10.1126/science.aab3761

[35] Chiang DY, McCarroll SA. Mapping duplicated sequences. Nature Biotechnology 2009; 27 (11): 1001-1002. doi: 10.1038/nbt1109-1001

[36] Firtina C, Alkan C. On genomic repeats and reproducibility. Bioinformatics 2016; 32 (15): 2243-2247. doi: 10.1093/bioinformatics/btw139

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

[38] Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M et al. Diversity of human copy number variation and multicopy genes. Science 2010; 330 (6004): 641-646. doi: 10.1126/science.1197005

[39] Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL et al. Great ape genetic diversity and population history. Nature 2013; 499 (7459): 471-475. doi: 10.1038/nature12228

[40] Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF et al. Copy number variation of individual cattle genomes using next-generation sequencing. Genome Research 2012; 22 (4): 778-790. doi: 10.1101/gr.133967.111

[41] Liu S, Kang X, Catacchio CR, Liu M, Fang L et al. Computational detection and experimental validation of segmental duplications and associated copy number variations in water buffalo (Bubalus bubalis). Functional & Integrative Genomics 2019; 19 (3): 409-419. doi: 10.1007/s10142-019-00657-4

[42] Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E et al. Genome sequencing highlights the dynamic early history of dogs. PLoS Genetics 2014; 10 (1): e1004016. doi: 10.1371/journal.pgen.1004016

[43] Tamazian G, Simonov S, Dobrynin P, Makunin A, Logachev A et al. Annotated features of domestic cat - Felis catus genome. Gigascience 2014; 3: 13. doi: 10.1186/2047-217X-3-13

[44] Cardone MF, D'Addabbo P, Alkan C, Bergamini C, Catacchio CR et al. Inter-varietal structural variation in grapevine genomes. The Plant Journal 2016; 88 (4):648-661. doi: 10.1111/tpj.13274

[45] Hach F, Sarrafi I, Hormozdiari F, Alkan C, Eichler EE et al. mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. Nucleic Acids Research 2014; 42: W494-W500. doi: 10.1093/nar/gku370

[46] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25 (16): 2078-2079. doi: 10.1093/bioinformatics/btp325

[47] Fritz MHY, Leinonen R, Cochrane G, Birney E. Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Research 2011; 21 (5): 734-740. doi: 10.1101/gr.114819.110

[48] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 2013; arXiv:13033997.

[49] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods 2012; 9: 357-359. doi: 10.1038/nmeth.1923

[50] Cleveland WS. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 1979; 74 (368): 829-836. doi: 10.1080/01621459.1979.10481038

[51] Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nature Reviews Genetics 2004; 5 (5): 335-344. doi: 10.1038/nrg1325

[52] Leipzig J. A review of bioinformatic pipeline frameworks. Briefings in Bioinformatics 2017; 18 (3): 530-536. doi: 10.1093/bib/bbw020

[53] Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nature Communications 2019; 10 (1): 1784. doi: 10.1038/s41467-018-08148-z.

[54] Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. Nature Communications 2019; 10 (1): 1025. doi: 10.1038/s41467-019-08992-7

[55] Sanders AD, Hills M, Porubský D, Guryev V, Falconer E et al. Characterizing polymorphic inversions in human genomes by single-cell sequencing. Genome Research 2016; 26 (11): 1575-1587. doi: 10.1101/gr.201160.115

[56] Mallick S, Li H, Lipson M, Mathieson I, Gymrek M et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 2016; 538: 201-206. doi: 10.1038/nature18964

[57] Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD et al. dbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Research 2013; 41: D936-D941. doi: 10.1093/nar/gks1213

[58] Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV et al. Recent segmental duplications in the human genome. Science 2002; 297 (5583): 1003-1007. doi: 10.1126/science.1072047

[59] Numanagic I, Gökkaya AS, Zhang L, Berger B, Alkan C et al. Fast characterization of segmental duplications in genome assemblies. Bioinformatics 2018; 34 (17): i706-i714. doi: 10.1093/bioinformatics/bty586