Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Toolkit for automated and rapid discovery of structural variants

Arda Soylev [a], Can Kockan [a,1], Fereydoun Hormozdiari [b,*], Can Alkan [a,*]

[a] Department of Computer Engineering, Bilkent University, Ankara, Turkey
[b] Department of Biochemistry and Molecular Medicine, MIND Institute and UC-Davis Genome Center, University of California, Davis, CA, United States

ARTICLE INFO

Article history:
Received 16 February 2017
Received in revised form 20 May 2017
Accepted 30 May 2017
Available online 2 June 2017

Keywords:
Structural variation
High throughput sequencing
Combinatorial algorithms

ABSTRACT

Structural variations (SV) are broadly defined as genomic alterations that affect >50 bp of DNA, which are shown to have significant effect on evolution and disease. The advent of high throughput sequencing (HTS) technologies and the ability to perform whole genome sequencing (WGS), makes it feasible to study these variants in depth. However, discovery of all forms of SV using WGS has proven to be challenging as the short reads produced by the predominant HTS platforms (<200 bp for current technologies) and the fact that most genomes include large amounts of repeats make it very difficult to unambiguously map and accurately characterize such variants. Furthermore, existing tools for SV discovery are primarily developed for only a few of the SV types, which may have conflicting sequence signatures (i.e. read pairs, read depth, split reads) with other, untargeted SV classes. Here we are introduce a new framework, TARDIS, which combines multiple read signatures into a single package to characterize most SV types simultaneously, while preventing such conflicts. TARDIS also has a modular structure that makes it easy to extend for the discovery of additional forms of SV.

1. Introduction

Genome structural variations (SVs), defined as genomic alterations >50 bp [1,2], play major roles in both genome evolution [3] and pathogenesis of diseases of genomic origin such as schizophrenia, epilepsy, and autism [4]. Although *-by count-* less number of SVs are found in each human genome with respect to the reference than single nucleotide polymorphisms (SNPs), the total number of affected basepairs by SVs far exceed those affected by SNPs [2]. It is, therefore, of utmost importance to accurately and comprehensively characterize all forms of SVs, including copy number variants (CNVs, i.e. deletions, insertions and duplications), mobile element insertions, and balanced rearrangements (inversions and translocations).

Algorithm development for structural variation discovery and genotyping using high throughput sequencing (HTS) data was accelerated during the 1000 Genomes Project [2,5,6]. Briefly, all algorithms use one or several of four basic read mapping *signatures*: read pair, split read, read depth, and assembly [1]. The detection accuracy of using each sequence signature differs depending on the type, size, and the underlying sequence properties of genomic location of the SV. Therefore, although the first few SV discovery algorithms focused on using a single sequence signature [7–14], more recent SV callers use multiple signatures [15–19]. However, most SV calling algorithms aim to characterize one or a few types of SV, and they do not try to resolve conflicting SV within the same locations, or sequence signature that signal more than one type of SV.

Here we introduce TARDIS, a toolkit for automated and rapid discovery of SVs. TARDIS integrates read pair, read depth, and split read (using soft clipped mappings) sequence signatures to discover several types of SV, while resolving ambiguities among different putative SVs: 1) at the same locations signaled by different sequence signatures, and 2) in different locations signaled by the same mapping information. TARDIS is fully automated and requires no user intervention. Additionally, it is suitable for cloud use as the memory footprint is low. The current version is capable of characterizing deletions, small novel insertions, tandem duplications, inversions, and mobile element retrotransposition.

TARDIS is implemented in C using HTSLib (http://www.htslib.org), and it is freely available at https://github.com/BilkentCompGen/tardis.

2. Methods

We have previously developed some of the first tools to discover various types of SV that also incorporate multi-mapping of

[*] Corresponding authors.
E-mail addresses: fhormozd@ucdavis.edu (F. Hormozdiari), calkan@cs.bilkent.edu.tr (C. Alkan).
[1] Current address: School of Informatics and Computing, Indiana University, Bloomington, IN, United States.

http://dx.doi.org/10.1016/j.ymeth.2017.05.030
1046-2023/© 2017 Elsevier Inc. All rights reserved.

reads, such as mrCaNaVaR/mrFAST [20], VariationHunter [8], VariationHunter-CR [13], NovelSeq [21], Pamir [22], and CommonLAW [23]. All of these tools use a similar objective function for SV discovery although they are developed to discover different types of SV under different conditions (e.g. single vs. multi-sample) using different sequence signatures [1,12]. We now further improve our algorithms for SV detection and integrate them into a single package (TARDIS) that can simultaneously characterize different forms of SVs using read pairs, read depth, and split reads. TARDIS is a user-friendly single executable with a potential to be easily extended for discovering additional forms of complex SV (e.g. translocations) and for supporting different sequencing technologies such as linked read sequencing [24] and long read sequencing (i.e., PacBio, nanopore). However, the current version of TARDIS is developed only for whole genome sequencing (WGS) data generated with the Illumina platform, and in the remainder of the paper we assume the input is Illumina WGS. Below we first define the terminology and then provide problem formulation and our solution.

We first define some of the terms that we use in this paper below.

- *fragment size:* the Illumina WGS protocol generates paired-end reads from both ends of longer fragments. The lengths of these fragments are assumed to be sampled from a normal distribution. Therefore, in the absence of structural variants, mapping locations of the paired ends *span* within an interval $[\delta_{min}, \delta_{max}]$. Most (>90%) of paired-end reads are sampled from no-SV regions, therefore the fragment size distribution can be learned empirically for each WGS data set separately.
- *concordant reads:* a read pair is called *concordant* if they can be mapped to the reference genome as “expected”: (a) mapped to opposing strands where the upstream read is mapped to the forward strand and the downstream read is mapped to the reverse strand,[2] (b) the distance between ends is between the minimum and maximum expected fragment size.
- *discordant reads:* briefly, any non-concordant read pair is considered *discordant*. Note that, by definition, the discordant read pairs signal potential SVs. The sequence signature produced by these type of reads is known as read-pair signature [1,12].
- *split reads:* a read that can only be mapped to the reference genome by breaking into two sub-reads is called a *split-read*. These types of reads also indicate a potential SV or a short insertion or deletion (indel).
- *read depth:* number of reads that map within a region of the genome. Overall genome-wide read depth is also referred to as *depth of coverage*. It is expected that the number of reads that “cover” each base-pair to follow a Poisson distribution. Therefore, if the read depth over a certain region deviates significantly from this distribution, it signals for a potential copy number variation (CNV) [1,20,12].

2.1. Problem formulation

One of the main drawbacks of high-throughput sequencing technologies is that reads are usually very short (<200 bp). This results in mapping ambiguity as some reads may map to more than one location equally likely due to genomic repeats and segmental duplications [25]. Similar to our previous work [8,13,23], TARDIS uses the signatures explained above and it also considers all map locations of multi-mapping reads. However, TARDIS also has a *quick* mode, which considers only the best map location provided in the input BAM file. We formulate our problem formulation under the assumption of *maximum parsimony*.

As in VariationHunter [8] the objective function that TARDIS tries to optimize is also based on maximum parsimony. Briefly, TARDIS aims to minimize the total number of structural variation inferred from all discordant read pairs and split reads. We have previously showed that maximum parsimony SV discovery problem is NP-Complete [8] by reduction from the SET-COVER problem [26]. Additionally we provided a greedy algorithm with an approximation factor of $O(\log\ n)$ using only the read pair signature.

In addition to the read pair signature, TARDIS also uses read depth and split read signatures for SV discovery. Briefly, after clustering discordant read pairs (Section 2.2), we can assign weights to the clusters based on the GC%-normalized read depth within the inferred cluster coordinates (Section 2.3). Note that, since the read depth weights are calculated for each cluster once, and they mainly represent a score, the approximation ratio of the greedy algorithm does not change.

2.2. Maximal valid clusters of read pairs

We define a set of discordant read pairs that signal the same SV (i.e. same type and size) as a *valid cluster*. Similarly, we define a *maximal valid cluster* as a valid cluster where no additional discordant read pairs can be added without violating its validity. Valid clusters for some of the SV types are previously defined in [27,8,28].

2.3. Read-depth signature

We use read depth signature to score and eliminate likely false positive CNV calls (deletions). We model read depth distribution as Poisson, and we calculate the read depth of each putative SV as the summation of read depths for each base pair within the SV breakpoints. Other discrete binomial distributions have been suggested for modeling read depth such as the *negative binomial* distribution [29]. Calculation of the distribution function is implemented as a module in TARDIS, thus it can be replaced in upcoming versions.

Note that the summation of two Poisson distributions is also a Poisson distribution. Additionally, we use a statistical smoothing method (i.e. LOESS transformation) to normalize read depth values based on the GC% content as previously described elsewhere [20,30].

Next, we calculate the probability $P(RD|CN = i)$[3] for each putative deletion within breakpoint intervals (B_l, B_r) as follows. We first calculate the *expected read depth* (denoted as E_{RD}) within the deletion breakpoints normalized with respect to its GC% content using a sliding window of size 100 bp. Here, the expected read depth refers to “normal” read depth (i.e. no CNV).

We then calculate for every region the copy number corrected (i.e. $CN = i$) expected read depth as

$$E_i = \frac{E_{RD} \times i}{2}$$

We also denote observed read-depth as O. Thus assuming Poisson distribution we calculate the probability $P(RD|CN = i)$ as:

$$P(RD|CN = i) = \frac{{E_i}^O \times e^{-E_i}}{O!}$$

We consider a deletion prediction to be correct if the likelihood of the observed read depth is significantly higher for a copy number that supports a deletion (i.e. CN = 0 or CN = 1) compared to that of CN> 1. More formally, we calculate the deletion likelihood assuming the copy number is bounded by 10.

[2] This is correct for most Illumina WGS data sets, however, there are alternative library preparation protocols with different strand rules.

[3] RD: read depth, CN: copy number, and *i* denotes an integer for copy number.

Table 1
Simulation results. We show the true and false discovery rates (TDR and FDR) of TARDIS without soft clipped reads and under different minimum read pair (RP) cut off values; and LUMPY, and DELLY at different depths of coverage from 5X to 40X. TARDIS consistently demonstrates low FDR, and its TDR is comparable to others.

Coverage	TARDIS-noSC (RP > 0)		TARDIS-noSC (RP> 2)		TARDIS-noSC (RP > 4)		LUMPY		DELLY	
	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR
5X	0.01	**0.48**	0.005	0.37	**0.004**	0.16	0.02	0.29	0.03	0.26
10X	0.02	**0.65**	0.009	0.58	**0.002**	0.44	0.02	0.59	0.04	0.56
20X	0.02	0.73	0.007	0.69	**0.001**	0.64	0.03	**0.75**	0.06	0.74
30X	0.04	0.75	0.015	0.71	**0.002**	0.68	0.04	0.80	0.08	**0.80**
40X	0.05	0.76	0.017	0.72	**0.002**	0.70	0.06	0.81	0.07	**0.83**

The best performing values are represented in bold.

$$\rho = \frac{P(RD|CN=0)P(CN=0) + P(RD|CN=1)P(CN=1)}{\sum_{i=2}^{i=10} P(RD|CN=i)P(CN=i)}$$

Note that the prior probability in the equation above (P(CN = i)) can be calculated using the previously identified copy number distribution profiles characterized in the genomes of individuals of the same species. For example, for the human genomes we can use the copy number profiles characterized by the 1000 Genomes Project [5] as prior values. In this paper we assumed that all prior values are the same. In addition, we use the log likelihood ratios as a metric to rank the predicted deletions, denoted as ρ. We performed parameter sweep to determine a good value for this threshold to optimize both true and false discovery rates (TDR and FDR) using simulations, which resulted in selecting $\rho \geqslant 2$.

2.4. Split read signature

Different from our previous algorithms, TARDIS also considers split read signal using soft clipped reads (>10 bp clips) in the input BAM file. TARDIS first tries to remap the clipped region to the reference genome to eliminate any mismappings in the original input. In order to establish consistency between clustering discordant read pairs and split reads, and also to account for possible incorrect mappings of very short segments, we treat split reads as a special case of discordant read pairs. We consider two splits of a split read as two ends of a read pair with a tight fragment size distribution (e.g. $\delta_{min} = 0$ and $\delta_{max} = 20$). This approach helps formulate a very similar framework for clustering split reads, and make it straightforward to include split reads and read pairs in the same SV clusters.

3. Results

3.1. Simulations

We first performed simulation experiments to benchmark the accuracy of TARDIS for deletion discovery and to compare it against two of the state-of-the-art SV discovery tools, LUMPY [18] and DELLY [17]. We used the VarSim [31] tool to simulate realistic structural variants and corresponding WGS reads. We show in 1 the benchmark results for TARDIS without incorporating the soft clipped reads (denoted TARDIS-noSC) at different depths of coverage.

3.2. Real data

We applied TARDIS to three real data sets. Here we opted for those that were sequenced at high depth using the Illumina platform, but also were sequenced using long reads generated with the single molecule real time (SMRT) technology (i.e. PacBio). Our motivation for choosing these samples was to be able to cross-validate and compare our calls predicted with an orthogonal technology.

Two of the WGS data sets we used were generated from haploid cell lines, namely CHM1 and CHM13. Illumina WGS was previously generated by [32] and PacBio data was reported in [33]. There also exists SV call sets for the same cell lines using PacBio data using the SMRT-SV algorithm [33]. The third data set we used was generated from the genome of a HapMap individual (NA12878). Similarly we used Illumina WGS [34] to characterize SVs using TARDIS, and PacBio data set [35] to compare and cross-validate.

3.2.1. Deletions

We first compared the deletions (>100 bp) we characterized in CHM1 and CHM13 genomes using TARDIS with call sets generated using LUMPY [18] and DELLY [17]. We required >50% reciprocal overlap for two deletions to be considered the same using BEDTools [14]. Additionally, under the assumption that the deletions called in corresponding PacBio data sets [33] are the gold standard, we calculated TDR and FDR for each call set (Fig. 1). We found in both experiments that TARDIS showed the lowest FDR among the three tools we tested with comparable sensitivity.

Next we compared the deletions detected in the genome of NA12878 using TARDIS and LUMPY (Fig. 2a). In the same figure we also provide the size distribution of deletions predicted by TARDIS. As expected, we observed peaks at 300 bp and 5900 bp, corresponding to Alu and L1 deletions, respectively.

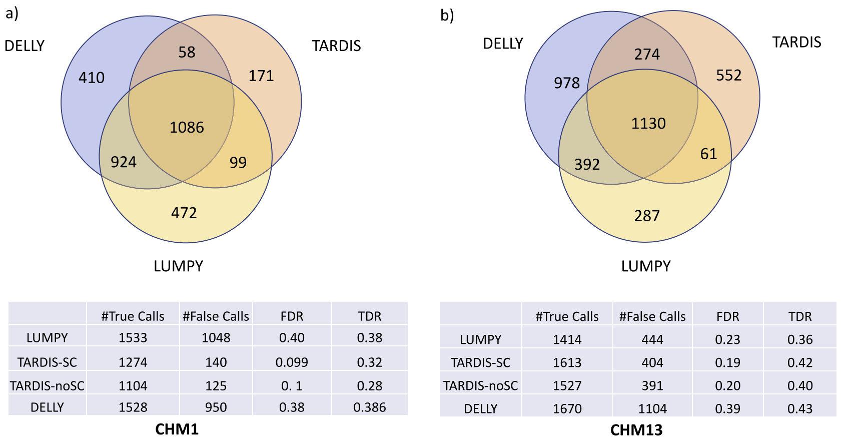
3.2.2. Mobile element insertions (MEI)

We also evaluated the performance of TARDIS in mobile element insertion discovery using the CHM1 and CHM13 genomes and compared to the orthogonal PacBio predictions (Fig. 3). We note that the MEI events TARDIS characterized but missing in PacBio data may indeed be real and simply false negatives in the PacBio predictions. Comparison of the additional MEI found by TARDIS with the previously known polymorphic MEI from dbRIP, showed that over 30% of these additional MEI are indeed correct. Further analysis also revealed that most of the MEI that TARDIS missed were found within other repeats, which makes it very challenging to accurately map short reads.

3.3. Time and memory usage

Finally we report the computational resources needed to run TARDIS, LUMPY, and DELLY in Table 2. We benchmarked all three tools on the same BAM file generated from the CHM1 genome (40X, mapped to reference human genome GRCh37). TARDIS completed the analysis substantially faster than LUMPY and DELLY, however it also required more memory. This is because TARDIS considers potentially multi-mapping reads, thus it has to analyze the entire genome. In contrast, LUMPY and DELLY perform chromosome-by-chromosome analysis, which requires lower memory footprint. Note that the speed and memory requirement were calculated using the same computing server.[4]

[4] Intel(R) Xeon(R) CPU E7- 4830 @ 2.13 GHz: 4 CPUs * 8 cores each = 32cores total 512 GB RAM.



	#True Calls	#False Calls	FDR	TDR
LUMPY	1533	1048	0.40	0.38
TARDIS-SC	1274	140	0.099	0.32
TARDIS-noSC	1104	125	0. 1	0.28
DELLY	1528	950	0.38	0.386

CHM1

	#True Calls	#False Calls	FDR	TDR
LUMPY	1414	444	0.23	0.36
TARDIS-SC	1613	404	0.19	0.42
TARDIS-noSC	1527	391	0.20	0.40
DELLY	1670	1104	0.39	0.43

CHM13

Fig. 1. Comparison of CHM1 and CHM13 deletions (>100 bp) between TARDIS, LUMPY, and DELLY calls. We also provide false and true discovery rate (FDR and TDR) estimations under the assumption that orthogonal PacBio predictions [33] provide the gold standard.

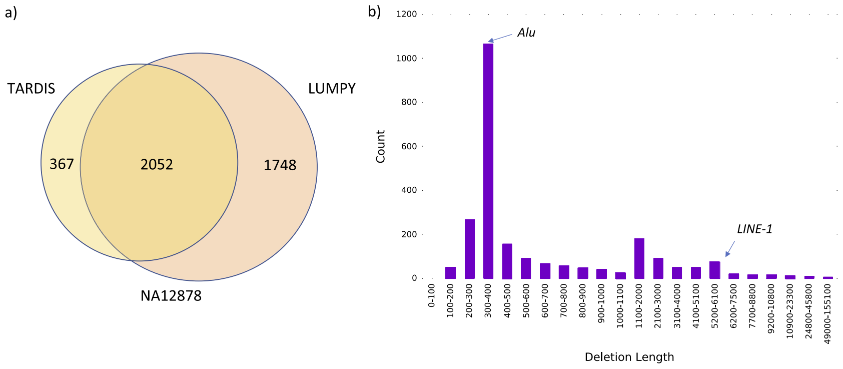


Fig. 2. Comparison of NA12878 deletions (>100 bp) between TARDIS and LUMPY calls (a). We also provide a deletion length histogram (b), demonstrating the expected peaks at 300 bp (Alu) and 5,900 bp (L1) deletions.

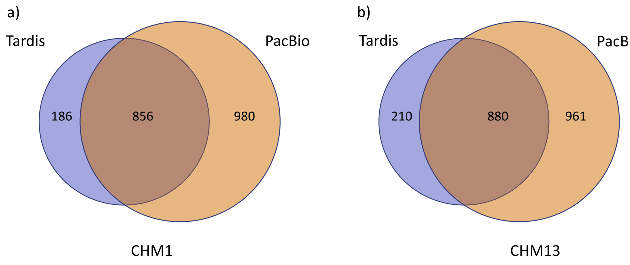


Fig. 3. Alu insertions predicted in the CHM1 and CHM13 genomes and compared against an orthogonal PacBio data set [33].

Table 2
Performance comparison of different tools for SV discovery in the CHM1 genome (40X depth of coverage).

Tools	CPU time	Peak memory usage(GB)
TARDIS	4 h 04 m	11 GB
LUMPY	8 h 41 m	7 GB
DELLY	32 h 19 m	0.3 GB

4. Discussion

In this paper we introduced TARDIS, a tool for easy and accurate structural variation discovery using whole genome shotgun sequencing based on the principles for SV discovery under maximum parsimony. TARDIS also is able to use multi-mapping reads to improve SV detection sensitivity in highly repetitive regions. Our experiments on real data and simulations demonstrated that TARDIS achieves better specificity than the state of the art methods for SV discovery and it is comparable to others in terms sensitivity. We have implemented TARDIS to allow easy extensions to discover other forms of complex SV such as inverted duplications and translocations.

Acknowledgments

We would like to thank V. Bhakhar, A. Tekat, B. Orabi, and R. Shahidi Nejad for their help in coding parts of the TARDIS software, and I. Hajirasouliha and C. Ricketts for extensive testing and bug reports. This study is partially supported by a TÜBİTAK grant (215E172) to C.A.

References

[1] C. Alkan, B.P. Coe, E.E. Eichler, Genome structural variation discovery and genotyping, Nat. Rev. Genet. 12 (5) (2011) 363–376, http://dx.doi.org/10.1038/nrg2958.

[2] R.E. Mills, K. Walter, C. Stewart, R.E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R.K. Cheetham, A. Chinwalla, D.F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L.M. Iakoucheva, Z. Iqbal, S. Kang, J.M. Kidd, M.K. Konkel, J. Korn, E. Khurana, D. Kural, H.Y.K. Lam, J. Leng, R. Li, Y. Li, C.-Y. Lin, R. Luo, X.J. Mu, J. Nemesh, H.E. Peckham, T. Rausch, A. Scally, X. Shi, M.P. Stromberg, A.M. Stütz, A.E. Urban, J.A. Walker, J. Wu, Y. Zhang, Z.D. Zhang, M.A. Batzer, L. Ding, G.T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E.E. Eichler, M.B. Gerstein, M.E. Hurles, C. Lee, S.A. McCarroll, J.O. Korbel, G. Project, Mapping copy number variation by population-scale genome sequencing, Nature 470 (7332) (2011) 59–65, http://dx.doi.org/10.1038/nature09708.

[3] J. Prado-Martinez, P.H. Sudmant, J.M. Kidd, H. Li, J.L. Kelley, B. Lorente-Galdos, K.R. Veeramah, A.E. Woerner, T.D. O'Connor, G. Santpere, A. Cagan, C. Theunert,

F. Casals, H. Laayouni, K. Munch, A. Hobolth, A.E. Halager, M. Malig, J. Hernandez-Rodriguez, I. Hernando-Herraez, K. Prüfer, M. Pybus, L. Johnstone, M. Lachmann, C. Alkan, D. Twigg, N. Petit, C. Baker, F. Hormozdiari, M. Fernandez-Callejo, M. Dabad, M.L. Wilson, L. Stevison, C. Camprubí, T. Carvalho, A. Ruiz-Herrera, L. Vives, M. Mele, T. Abello, I. Kondova, R.E. Bontrop, A. Pusey, F. Lankester, J.A. Kiyang, R.A. Bergl, E. Lonsdorf, S. Myers, M. Ventura, P. Gagneux, D. Comas, H. Siegismund, J. Blanc, L. Agueda-Calpena, M. Gut, L. Fulton, S.A. Tishkoff, J.C. Mullikin, R.K. Wilson, I.G. Gut, M.K. Gonder, O. A. Ryder, B.H. Hahn, A. Navarro, J.M. Akey, J. Bertranpetit, D. Reich, T. Mailund, M.H. Schierup, C. Hvilsom, A.M. Andrés, J.D. Wall, C.D. Bustamante, M.F. Hammer, E.E. Eichler, T. Marques-Bonet, Great ape genetic diversity and population history, Nature 499 (7459) (2013) 471–475.

[4] P. Stankiewicz, J.R. Lupski, Structural variation in the human genome and its role in disease, Annu. Rev. Med. 61 (2010) 437–455, http://dx.doi.org/10.1146/annurev-med-100708-204735.

[5] The 1000 Genomes Project Consortium, A global reference for human genetic variation, Nature 526 (7571) (2015) 68–74, http://dx.doi.org/10.1038/nature15393.

[6] P.H. Sudmant, S. Mallick, B.J. Nelson, F. Hormozdiari, N. Krumm, J. Huddleston, B. P. Coe, C. Baker, S. Nordenfelt, M. Bamshad, L.B. Jorde, O.L. Posukh, H. Sahakyan, W.S. Watkins, L. Yepiskoposyan, M.S. Abdullah, C.M. Bravi, C. Capelli, T. Hervig, J. T.S. Wee, C. Tyler-Smith, G. van Driem, I.G. Romero, A.R. Jha, S. Karachanak-Yankova, D. Toncheva, D. Comas, B. Henn, T. Kivisild, A. Ruiz-Linares, A. Sajantila, E. Metspalu, J. Parik, R. Villems, E.B. Starikovskaya, G. Ayodo, C.M. Beall, A. Di Rienzo, M.F. Hammer, R. Khusainova, E. Khusnutdinova, W. Klitz, C. Winkler, D. Labuda, M. Metspalu, S.A. Tishkoff, S. Dryomov, R. Sukernik, N. Patterson, D. Reich, E.E. Eichler, Global diversity, population stratification, and selection of human copy-number variation, Science 349 (6253) (2015). http://science.sciencemag.org/content/349/6253/aab3761.

[7] J.O. Korbel, A.E. Urban, J.P. Affourtit, B. Godwin, F. Grubert, J.F. Simons, P.M. Kim, D. Palejev, N.J. Carriero, L. Du, B.E. Taillon, Z. Chen, A. Tanzer, A.C.E. Saunders, J. Chi, F. Yang, N.P. Carter, M.E. Hurles, S.M. Weissman, T.T. Harkins, M.B. Gerstein, M. Egholm, M. Snyder, Paired-end mapping reveals extensive structural variation in the human genome, Science 318 (5849) (2007) 420–426, http://dx.doi.org/10.1126/science.1149504.

[8] F. Hormozdiari, C. Alkan, E.E. Eichler, S.C. Sahinalp, Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes, Genome Res. 19 (7) (2009) 1270–1278, http://dx.doi.org/10.1101/gr.088633.108.

[9] K. Chen, J.W. Wallis, M.D. McLellan, D.E. Larson, J.M. Kalicki, C.S. Pohl, S.D. McGrath, M.C. Wendl, Q. Zhang, D.P. Locke, X. Shi, R.S. Fulton, T.J. Ley, R.K. Wilson, L. Ding, E.R. Mardis, BreakDancer: an algorithm for high-resolution mapping of genomic structural variation, Nat. Methods 6 (9) (2009) 677–681, http://dx.doi.org/10.1038/nmeth.1363.

[10] S. Lee, F. Hormozdiari, C. Alkan, M. Brudno, MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions, Nat. Methods 6 (7) (2009) 473–474, http://dx.doi.org/10.1038/nmeth.f.256.

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

[12] P. Medvedev, M. Stanciu, M. Brudno, Computational methods for discovering structural variation with next-generation sequencing, Nat. Methods 6 (11 Suppl) (2009) S13–S20, http://dx.doi.org/10.1038/nmeth.1374.

[13] F. Hormozdiari, I. Hajirasouliha, P. Dao, F. Hach, D. Yorukoglu, C. Alkan, E.E. Eichler, S.C. Sahinalp, Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery, Bioinformatics 26 (12) (2010) i350–i357, http://dx.doi.org/10.1093/bioinformatics/btq216.

[14] A.R. Quinlan, R.A. Clark, S. Sokolova, M.L. Leibowitz, Y. Zhang, M.E. Hurles, J.C. Mell, I.M. Hall, Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome, Genome Res. 20 (5) (2010) 623–635, http://dx.doi.org/10.1101/gr.102970.109.

[15] R.E. Handsaker, J.M. Korn, J. Nemesh, S.A. McCarroll, Discovery and genotyping of genome structural polymorphism by sequencing on a population scale, Nat. Genet. 43 (3) (2011) 269–276, http://dx.doi.org/10.1038/ng.768.

[16] T. Marschall, I.G. Costa, S. Canzar, M. Bauer, G.W. Klau, A. Schliep, A. Schönhuth, CLEVER: clique-enumerating variant finder, Bioinformatics 28 (22) (2012) 2875–2882, http://dx.doi.org/10.1093/bioinformatics/bts566.

[17] T. Rausch, T. Zichner, A. Schlattl, A.M. Stütz, V. Benes, J.O. Korbel, DELLY: structural variant discovery by integrated paired-end and split-read analysis, Bioinformatics 28 (18) (2012) i333–i339, http://dx.doi.org/10.1093/bioinformatics/bts378.

[18] R.M. Layer, C. Chiang, A.R. Quinlan, I.M. Hall, LUMPY: a probabilistic framework for structural variant discovery, Genome Biol. 15 (6) (2014) R84, http://dx.doi.org/10.1186/gb-2014-15-6-r84.

[19] S.S. Sindi, S. Onal, L.C. Peng, H.-T. Wu, B.J. Raphael, An integrative probabilistic model for identification of structural variation in sequencing data, Genome Biol. 13 (3) (2012) R22, http://dx.doi.org/10.1186/gb-2012-13-3-r22.

[20] C. Alkan, J.M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. O. Kitzman, C. Baker, M. Malig, O. Mutlu, S.C. Sahinalp, R.A. Gibbs, E.E. Eichler, Personalized copy number and segmental duplication maps using next-generation sequencing, Nat. Genet. 41 (10) (2009) 1061–1067, http://dx.doi.org/10.1038/ng.437.

[21] I. Hajirasouliha, F. Hormozdiari, C. Alkan, J.M. Kidd, I. Birol, E.E. Eichler, S.C. Sahinalp, Detection and characterization of novel sequence insertions using paired-end next-generation sequencing, Bioinformatics 26 (10) (2010) 1277–1283, http://dx.doi.org/10.1093/bioinformatics/btq152.

[22] P. Kavak, Y.-Y. Lin, I. Numanagić, H. Asghari, T. Güngör, C. Alkan, F. Hach, Discovery and genotyping of novel sequence insertions in many sequenced individuals, Bioinformatics to appear.

[23] F. Hormozdiari, F. Hach, S.C. Sahinalp, E.E. Eichler, C. Alkan, Sensitive and fast mapping of di-base encoded reads, Bioinformatics 27 (14) (2011) 1915–1921, http://dx.doi.org/10.1093/bioinformatics/btr303.

[24] Y. Mostovoy, M. Levy-Sakin, J. Lam, E.T. Lam, A.R. Hastie, P. Marks, J. Lee, C. Chu, C. Lin, Ž. Džakula, H. Cao, S.A. Schlebusch, K. Giorda, M. Schnall-Levin, J.D. Wall, P.-Y. Kwok, A hybrid approach for de novo human genome sequence assembly and phasing, Nat Methods 13 (2016) 587–590, http://dx.doi.org/10.1038/nmeth.3865.

[25] C. Firtina, C. Alkan, On genomic repeats and reproducibility, Bioinformatics 32 (15) (2016) 2243–2247, http://dx.doi.org/10.1093/bioinformatics/btw139.

[26] R.M. Karp, Reducibility among combinatorial problems, in: Complexity of Computer Computations, Springer, 1972, pp. 85–103.

[27] A. Bashir, S. Volik, C. Collins, V. Bafna, B.J. Raphael, Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer, PLoS Comput. Biol. 4 (4) (2008) e1000051, http://dx.doi.org/10.1371/journal.pcbi.1000051.

[28] S. Sindi, E. Helman, A. Bashir, B.J. Raphael, A geometric approach for classification and comparison of structural variants, Bioinformatics 25 (2009) i222–i230, http://dx.doi.org/10.1093/bioinformatics/btp208 (Oxford, England).

[29] C.A. Miller, O. Hampton, C. Coarfa, A. Milosavljevic, ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads, PloS One 6 (2011) e16327, http://dx.doi.org/10.1371/journal.pone.0016327.

[30] A. Abyzov, A.E. Urban, M. Snyder, M. Gerstein, CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing, Genome Res. 21 (6) (2011) 974–984, http://dx.doi.org/10.1101/gr.114876.110.

[31] J.C. Mu, M. Mohiyuddin, J. Li, N. Bani Asadi, M.B. Gerstein, A. Abyzov, W.H. Wong, H.Y.K. Lam, VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications, Bioinformatics 31 (9) (2015) 1469–1471, http://dx.doi.org/10.1093/bioinformatics/btu828.

[32] K.M. Steinberg, V.A. Schneider, T.A. Graves-Lindsay, R.S. Fulton, R. Agarwala, J. Huddleston, S.A. Shiryev, A. Morgulis, U. Surti, W.C. Warren, D.M. Church, E.E. Eichler, R.K. Wilson, Single haplotype assembly of the human genome from a hydatidiform mole, Genome Res. 24 (12) (2014) 2066–2076, http://dx.doi.org/10.1101/gr.180893.114.

[33] J. Huddleston, E.E. Eichler, An incomplete understanding of human genetic variation, Genetics 202 (4) (2016) 1251–1254, http://dx.doi.org/10.1534/genetics.115.180539.

[34] M.A. Eberle, E. Fritzilas, P. Krusche, M. Kallberg, B.L. Moore, M.A. Bekritsky, Z. Iqbal, H.-Y. Chuang, S.J. Humphray, A.L. Halpern, S. Kruglyak, E.H. Margulies, G. McVean, D.R. Bentley, A reference dataset of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree, Genome Res 27 (1) (2017) 157–164, http://dx.doi.org/10.1101/gr.210500.116.

[35] J.M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C.E. Mason, N. Alexander, E. Henaff, A.B.R. McIntyre, D. Chandramohan, F. Chen, E. Jaeger, A. Moshrefi, K. Pham, W. Stedman, T. Liang, M. Saghbini, Z. Dzakula, A. Hastie, H. Cao, G. Deikus, E. Schadt, R. Sebra, A. Bashir, R.M. Truty, C.C. Chang, N. Gulbahce, K. Zhao, S. Ghosh, F. Hyland, Y. Fu, M. Chaisson, C. Xiao, J. Trow, S. T. Sherry, A.W. Zaranek, M. Ball, J. Bobe, P. Estep, G.M. Church, P. Marks, S. Kyriazopoulou-Panagiotopoulou, G.X.Y. Zheng, M. Schnall-Levin, H.S. Ordonez, P.A. Mudivarti, K. Giorda, Y. Sheng, K.B. Rypdal, M. Salit, Extensive sequencing of seven human genomes to characterize benchmark reference materials, Sci. Data 3 (2016) 160025, http://dx.doi.org/10.1038/sdata.2016.25.