bioRxiv preprint doi: https://doi.org/10.1101/015156; this version posted February 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discovery of large genomic inversions using pooled clone sequencing

Marzieh Eslami Rasekh [1], Giorgia Chiatante [2], Mattia Miroballo [2], Joyce Tang [3], Mario Ventura [2], Chris T. Amemiya [3], Evan E. Eichler [4], Francesca Antonacci [2,*] and Can Alkan [1, *]

[1]Department of Computer Engineering, Bilkent University, 06800 Ankara, Turkey
[2]Department of Biology, University of Bari, Via Orabona 4, 70125 Bari, Italy
[3]Benaroya Research Institute, 1201 Ninth Avenue, 98101 Seattle, WA, United States
[4]Department of Genome Sciences and Howard Hughes Medical Institute, University of Washington, 3720 15th Avenue NE, 98195 Seattle, WA, United States

ABSTRACT

Motivation: There are many different forms of genomic structural variation that can be broadly classified as copy number variation (CNV) and balanced rearrangements. Although many algorithms are now available in the literature that aim to characterize CNVs, discovery of balanced rearrangements (inversions and translocations) remains an open problem. This is mainly because the breakpoints of such events typically lie within segmental duplications and common repeats, which reduce the mappability of short reads. The 1000 Genomes Project spearheaded the development of several methods to identify inversions, however, they are limited to relatively short inversions, and there are currently no available algorithms to discover large inversions using high throughput sequencing technologies (HTS).

Results: Here we propose to use a sequencing method (Kitzman *et al.*, 2011) originally developed to improve haplotype resolution to characterize large genomic inversions. This method, called pooled clone sequencing, merges the advantages of clone based sequencing approach with the speed and cost efficiency of HTS technologies. Using data generated with pooled clone sequencing method, we developed a novel algorithm, dipSeq, to discover large inversions (>500 Kbp). We show the power of dipSeq first on simulated data, and then apply it to the genome of a HapMap individual (NA12878). We were able to accurately discover all previously known and experimentally validated large inversions in the same genome. We also identified a novel inversion, and confirmed using fluorescent in situ hybridization.

Availability: Implementation of the dipSeq algorithm is available at https://github.com/BilkentCompGen/dipseq

Contact: calkan@cs.bilkent.edu.tr, francesca.antonacci@uniba.it

1 INTRODUCTION

Genomic structural variants are defined as alterations in the DNA that affect >50bp that may delete, insert, duplicate, invert, or move genomic sequence (Alkan *et al.*, 2011). Structural variation (SV) are shown to be common in human genomes (Iafrate *et al.*, 2004; Sebat *et al.*, 2004), which caused increased interest in the characterization of both normal (Tuzun *et al.*, 2005; Kidd *et al.*, 2008; Mills *et al.*, 2011), and disease-causing large variants (Sharp *et al.*, 2006; Antonacci *et al.*, 2010). Furthermore, SVs are known to be one of the driving forces of creation of new haplotypes (Steinberg *et al.*, 2012), and evolution (Ventura *et al.*, 2011).

SVs were initially identified using bacterial artificial chromosome (BAC) arrays (Iafrate *et al.*, 2004; Sebat *et al.*, 2004; Redon *et al.*, 2006), then SNP arrays (Redon *et al.*, 2006; McCarroll *et al.*, 2006) and array comparative genomic hybridization (Conrad *et al.*, 2010). A more detailed map of structural variation was made possible using fosmid end sequencing (Tuzun *et al.*, 2005; Kidd *et al.*, 2008), however this method was not cost effective due to the use of the Sanger method. Introduction of high throughput sequencing (HTS) finally made it possible to screen the genomes of many (Korbel *et al.*, 2007; Alkan *et al.*, 2009; Hormozdiari *et al.*, 2009; Yoon *et al.*, 2009) to thousands (Mills *et al.*, 2011) of individuals.

Although there are now many algorithms to discover and genotype SV using HTS data (Medvedev *et al.*, 2009; Alkan *et al.*, 2011), they mainly focus on copy number variation (CNV), which change the amount of DNA, such as deletions, duplications, insertions, and transpositions. Other types of SV, namely *balanced rearrangements* such as inversions and translocations are harder to detect due to the fact that their breakpoints usually lie within complex repeats, reducing mappability. Balanced rearrangements also do not alter the read depth, which makes the use of read depth signature (Medvedev *et al.*, 2009; Yoon *et al.*, 2009; Alkan *et al.*, 2009) irrelevant for their detection. Therefore, very few attempts to characterize inversions are reliable only for small inversions (∼10-50 Kb) (Hormozdiari *et al.*, 2009; Rausch *et al.*, 2012; Layer *et al.*, 2014; Chaisson *et al.*, 2014), and exhibit high false discovery rates in translocation call sets (Talkowski *et al.*, 2011). Characterization of large genomic inversions using HTS remains an open problem.

The common method to discover inversions is to analyze the read pair signature (Medvedev *et al.*, 2009; Alkan *et al.*, 2011), where the mapping strand of the read pairs spanning the inversion breakpoints will be different from what is expected (Figure 1). For example, the Illumina platform generates read pairs from opposing strands, however, if the DNA fragment spans an inversion breakpoint, they will both be mapped to the same strand. They will also be separated from each other by a distance approximately same with the inversion

*to whom correspondence should be addressed

bioRxiv preprint doi: https://doi.org/10.1101/015156; this version posted February 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

size. When the inversion is large, the *real* mapping distance between pairs also increases, therefore increasing the chance of incorrect mapping due to the common repeats that lie in between. Many algorithms were developed using this read pair signature, such as VariationHunter (Hormozdiari *et al.*, 2009), inGap (Qi and Zhao, 2011), and HYDRA (Quinlan *et al.*, 2010). Another algorithm, INVY (Rausch *et al.*, 2012) also uses split read signature to better identify inversions. Most recently, LUMPY (Layer *et al.*, 2014) was developed that integrates multiple sequence signatures, including read alignments, and prior knowledge into a probabilistic framework. However, all the aforementioned tools have limitations on the inversion size due to the challenges explained above.

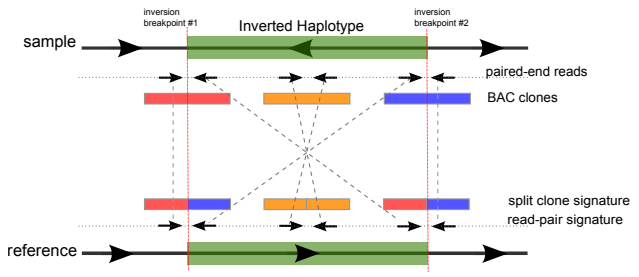

Fig. 1. Sequence signatures used by the dipSeq algorithm. In the presence of an inverted haplotype in the sequenced genome, we look for both read pair and split clone signatures. Paired end reads that span the inversion breakpoints will be mapped to the same strand with a large distance between them, instead of the concordant read pairs that map to opposing strands (Medvedev *et al.*, 2009; Alkan *et al.*, 2011). Large insert clones will show mapping properties similar to the split read sequence signature (Ye *et al.*, 2009), but since we do not have the full clone sequence, or sufficient coverage to assemble clones, we interrogate lengths of contiguous read mapping (Methods).

The HTS platforms generate data at very high rates with minimal cost. However, since both the HTS reads (100-150 bp for Illumina), and the DNA fragments are very short (350-500 bp), the mappability of the HTS data is dramatically reduced in repeat-rich regions that harbor most of the inversion breakpoints. On the contrary, the now-largely-abandoned method of clone-by-clone sequencing (International Human Genome Sequencing Consortium, 2001) enables data observation from much larger genomic intervals (40-to-200 Kb), but the associated costs are substantially higher. A sequencing method, called *pooled clone sequencing* (PCS) aims to combine the advantages of clone-by-clone sequencing, with the cost and time efficiency offered by the HTS platforms (Kitzman *et al.*, 2011) (see Methods). Although pooled clone sequencing was developed to improve haplotype phasing and to characterize large haplotype blocks, we propose a new algorithm, *dipSeq*, that utilizes PCS to discover large genomic inversions (>500 Kb). dipSeq proves its potential when tested on simulated data, and it is able to discover previously characterized large inversions (2-5 Mb) in the genome of a human individual (NA12878), using pooled BAC sequence data. dipSeq is *theoretically* compatible with all similarly constructed pooled sequence data, such as the TruSeq Synthetic Long-Reads (Moleculo) (Kuleshov *et al.*, 2014), or the Complete Genomics LFR Technology (Peters *et al.*, 2012), provided that the pooled large DNA fragment sizes follow a Gaussian distribution.

2 OBSERVATION AND APPROACH

Kitzman *et al.* (2011) developed the pooled clone sequencing (PCS) method to improve haplotype phasing. Basically, genomic DNA is cloned into cloning vectors (fosmids, or BACs), which are then diluted to approximately 3% coverage of the diploid genome, and randomly placed into several number of pools (Methods). Next, the pools are barcoded, and sequenced using the Illumina platform. Note that, due to dilution and random generation of pools, it is expected that pools will not harbor overlapping clones within themselves (Kitzman *et al.*, 2011). We provide a method to approximately calculate the probability of having overlapping clones within a pool in the Supplementary Note.

Our approach to discover large (>500 Kbp) genomic inversion using PCS follows from the observation that, clones (BAC or fosmid) that span the inversion breakpoint will be split into two sections when mapped to the reference genome, also separated by a distance approximately the size of the inversion. We call this sequence signature as *split clones* (Figure 1, which is similar to the split read sequence signature used by several SV discovery tools such as DELLY (Rausch *et al.*, 2012) and Pindel (Ye *et al.*, 2009). Based on these observations, we developed a novel combinatorial algorithm and statistical heuristics called *dipSeq* (**d**iscover **i**nversions using **p**ooled **Seq**uencing). Briefly, dipSeq searches for both read pair and split clone sequence signatures using the mapping locations of pooled clone sequencing reads, and requires split clones from different pools to cluster at the same putative inversion breakpoints (Methods). Ambiguity due to multiple possible pairings of split clones are resolved using an approximation algorithm for the maximal quasi clique problem (Brunato *et al.*, 2008), and paired-end read support further assigns confidence score for the predicted inversion calls.

Assuming that read and mapping errors are negligible, the probability of discovering an inversion using dipSeq can also be estimated (Supplementary Note) given physical depth of coverage by the clones.

3 METHODS

3.1 Building pooled clone libraries

We first generated a single whole-genome BAC library with long inserts (∼140 kbp). This procedure is a modification of the original haplotyping method previously described by Kitzman et al. (2011), that generates fosmid libraries with ∼40 kbp inserts. Here we used BAC clones, since long inserts are required to span the large duplication blocks where inversion breakpoints typically map (Kidd *et al.*, 2008). The library was then randomly partitioned into pools such that each pool is essentially a haploid mixture of clones derived from either the maternal or paternal DNA at each genomic location. High-throughput sequencing of each pool provided haplotype information for each clone in that pool.

We used genomic DNA from a HapMap Project individual (NA12878) to construct the BAC library. High molecular weight DNA was isolated, partially EcoRI digested, and subcloned into

bioRxiv preprint doi: https://doi.org/10.1101/015156; this version posted February 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pCC1BAC vector (Epicentre) to create a ∼140 kbp insert library using previously described protocols (Smith *et al.*, 2010). We then split a portion of this library to 3 sets of 96 pools each, with 230 clones per pool for set 1, 389 clones per pool for set 2 and 153 clones per pool for set 3. Each pool was expanded by direct liquid outgrowth after infection. We next constructed 96 barcoded sequencing libraries per each set, for a total of 288 sequencing libraries (Adey *et al.*, 2010). Libraries from each set were indexed with barcodes, combined and sequenced using the Illumina HiSeq platform (101bp paired-end reads). Upon sequencing a total of 74,112 clones (22,080 in set1, 37,344 in set 2 and 14,688 in set 3) we obtained 3.38× expected physical depth of coverage. After read mapping and reconstructing clones (Section 3.3), 87.58% of the genome was covered by one or more clones.

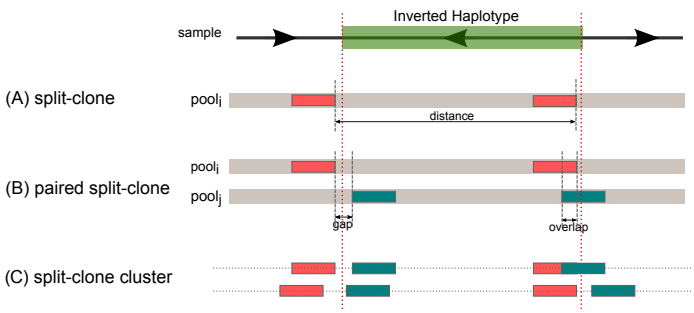

Fig. 2. Clustering split clones to detect inversions. (A) We first identify clone locations that are shorter than the expected clone size, but when paired with another short clone found in the same pool, the total length sums up to a full clone length. We call such clones as split clones. (B) We then cluster pairs of split clones that are mapped to approximately the same breakpoints. Note that due to read mapping errors and our clone reconstruction heuristics, a split clone may be identified as spanning a breakpoint. (C) Finally we cluster multiples of split clones from different pools if they agree on breakpoint location and the size of the inversion. *gap*: size of the region between the start and end locations of split clones from different pools. *overlap*: size of the overlapping region of split clones from different pools.

3.2 Read mapping

We first map the paired-end reads generated for each pool separately to the human reference genome assembly (GRCh37). Our dipSeq algorithm does not depend on any specific aligner, but in this study we used both BWA (Li and Durbin, 2009), and mrFAST (Xin *et al.*, 2013). We then separate the read pairs that map in the same orientation (i.e. read pair signature for inversions using Illumina), and those that map concordantly (within 4 standard deviations of the average fragment span size) into separate files to facilitate easier clone reconstruction (Section 3.3), and calculating read pair support (Section 3.4).

3.3 Reconstructing clones

We use only the concordantly mapped read pairs to infer the locations of clones. However, due to the low depth and breadth of coverage, it is not always possible to observe a continuous mapping of read pairs that collectively span genomic intervals within expected size of BAC clones. To overcome this issue, we apply several heuristics to identify clone locations. We use a sliding window approach, where we first look for regions to be covered by at least 50%. We use such regions that are ≥6.5Kbp as *seeds*. We then extend these seed windows using any read pairs that map to its flanking regions with a distance of at most 1.5 Kbp. Although the parameters we used here may seem arbitrary, in fact were obtained by applying an optimization grid on simulated BAC data (Section 4.1, and Supplementary Note) where we required >80% of the BACs to be recovered. This algorithm runs in $O(n \log n)$ time for sorting the reads, and amortized run time of $O(n)$ for reconstructing the clones, where n is the number of reads.

3.4 Inversion Discovery

After the identification of read pairs with inversion signature (i.e. mapping to same strand), and the predicted clone locations, we then look for potential split clones in each pool. We first extract the locations of *inferred* clones that are smaller than expected clone length, and look for those pairs of small clones (Figure 2) where the summation of their lengths is within an expected size range ($\mu \pm 3\sigma$). We also require the distance between the split clones to be within the inversion size limits we are trying to discover. In this study we set this parameter to 500 Kbp - 10 Mbp. Therefore, two regions r_k and r_l are predicted to be a split clone, denoted as SC_{r_k,r_l} if:

$$|r_k| < \mu_{clone},\ |r_l| < \mu_{clone}$$

$$\mu - 3\sigma \leq |r_k| + |r_l| \leq \mu + 3\sigma$$

$$min_inv_size \leq |r_k.start - r_l.start| \leq max_inv_size$$

Assuming the inferred clone locations are sorted by mapping locations, our algorithm can detect split clones in $O(n)$ amortized run time, where n is the number of inferred clones. However, the constant coefficient increases rapidly with the average read coverage.

We build inversion clusters by identifying two split clone pairs from different pools that are compatible (i.e. same breakpoint locations and inversion size). Due to both mapping errors and biases caused by our sliding window approach we permit a gap or overlap between the split clones to be paired (Figure 2b). We expect the inversion breakpoints lie between these gaps. Two split clones SC_{r_k,r_l} and $SC_{r_{k'},r_{l'}}$ are *compatible* to be in the same paired split clone (PSC) set, assuming $r_k/r_{k'}$ are located upstream of $r_l/r_{l'}$, if:

$$-max_overlap < r_{k'}.start - r_k.start < max_gap; \quad or\ vice\ versa$$

$$-max_overlap < r_{l'}.start - r_l.start < max_gap; \quad or\ vice\ versa$$

Here we set $max_overlap = 2 \times (\mu_{fragment} + 4\sigma_{fragment})$, where $\mu_{fragment}$ is the average fragment length, and $\sigma_{fragment}$ is the standard deviation of the paired-end sequencing data ($max_overlap$=2 Kbp for our data set). We also set max_gap to maximum expected clone length (∼200 Kbp for BACs). Note that adding more split clones to the same cluster will narrow down the gap size in breakpoint estimate. However, not all of the split clones we identify signal an inversion event. In an ideal case, where there are no mapping errors, other forms of structural variation, or areas with low mappability may also show themselves as split clone

bioRxiv preprint doi: https://doi.org/10.1101/015156; this version posted February 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

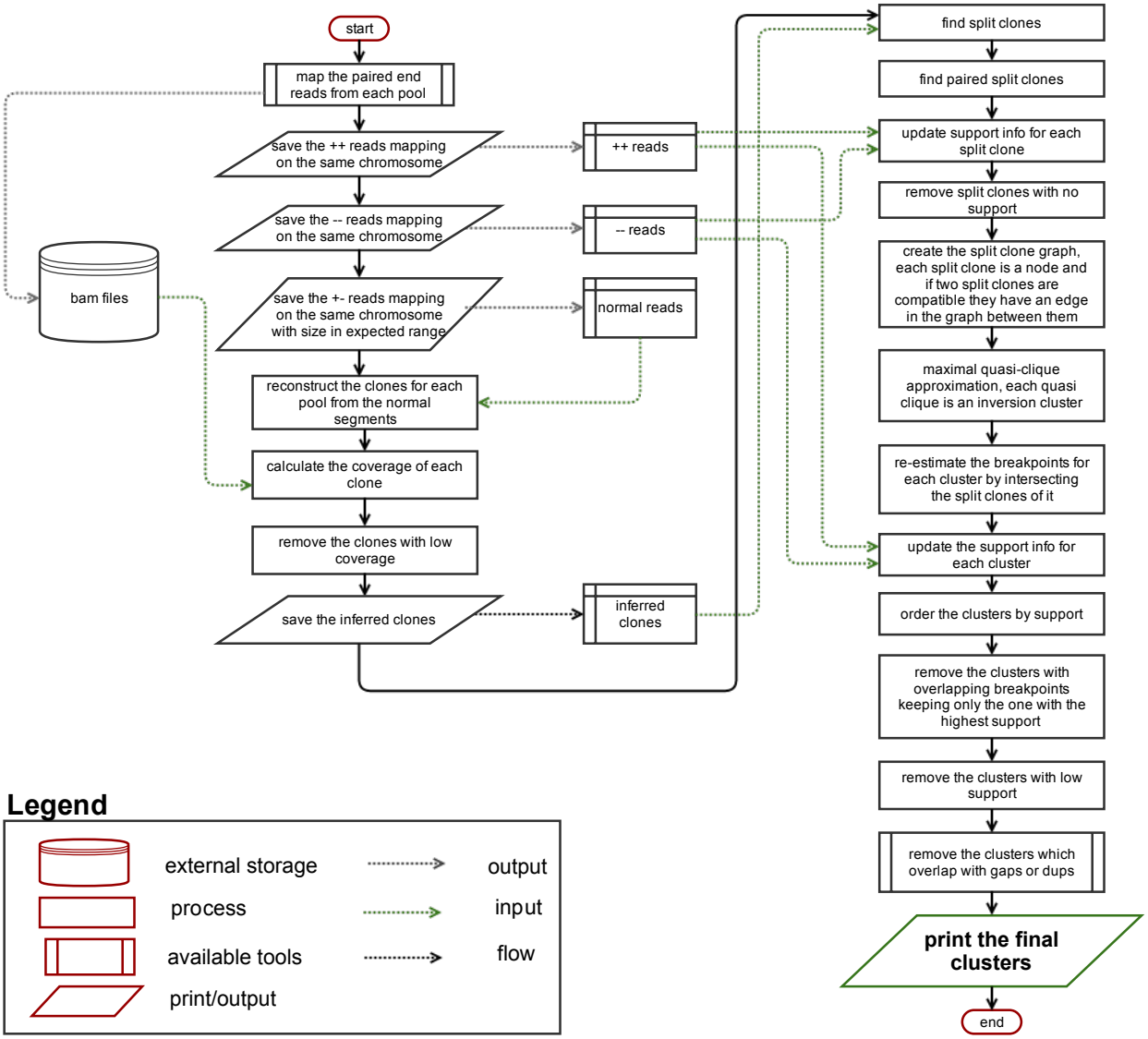

Fig. 3. The dipSeq algorithm. We start with mapping paired end reads for each pool. We then separate read pairs that map to the same strand, and generate two files for the two strands. We use the concordantly mapping reads to reconstruct clone locations, and calculate depth and breadth of coverage for the clones. Next, we remove clones with low coverage values, which result in the inferred clone locations. We identify the split clones from this list, and then find pairs of split clones (PSC), then we remove those with insufficient read pair support. Remaining PSCs are used to construct a graph, which we then use to find maximal quasi cliques that signal possible inversion locations. The clone and read pair support are then recalculated for the merged PSCs, those with low support and those that intersect with reference assembly gaps, or intersect with segmental duplications in both breakpoints are discarded.

signature for inversions. To ensure only split clones that signal a true inversion are detected, we also require read pair support for inversions (Medvedev *et al.*, 2009; Alkan *et al.*, 2011), and we discard any split clones that are not supported by read pairs. This step of the algorithm runs in $O(m+n)$, where m is the number of read pairs with inversion signature, and n is the number of split clones given than $m \gg n$.

Each pair of split clones gives a hint about the existence of an inverted haplotype. There may be many incorrectly identified split clone inversion signatures, or a short clone may have multiple potential "mate"s with similar properties. Therefore, clustering multiple split clone pairs that share inversion breakpoint locations and inversion lengths can help resolve the inversion breakpoints more accurately (Figure 2c). To both resolve ambiguities from multiple possible split clone pairings, and unambiguously identify inversions, we construct an undirected graph, where each PSC is a node, and an edge between two nodes indicates that share predicted breakpoints.

We initially formulated the inversion detection using split clones as a SET-COVER problem similar to VariationHunter, however, we

bioRxiv preprint doi: https://doi.org/10.1101/015156; this version posted February 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Inversions implanted on chromosome 1 for the simulation experiments.

ID	Start (bp)	End (bp)	Length (bp)	Genotype	SC_{1L}	SC_{1R}	$Findable_1$	SC_{2L}	SC_{2R}	$Findable_2$
Inv1	4,676,940	6,950,520	2,273,580	Het (P)	4	2	Y	0	3	N
Inv2	69,598,860	72,079,080	2,480,220	Het (M)	2	3	Y	10	6	Y
Inv3	76,232,700	82,398,900	6,166,200	Hom	7	6	Y	5+4	5+3	Y
Inv4	94,844,700	98,902,620	4,057,920	Hom	8	5	Y	3+4	5+2	Y
Inv5	107,694,120	109,006,800	1,312,680	Het (P)	1	4	Y	1	4	Y
Inv6	171,527,460	176,658000	5,130,540	Het (M)	2	7	Y	1	1	Y
Inv7	185,266,200	187,919,700	2,653,500	Hom	11	5	Y	2+3	3+2	Y
Inv8	190,600,560	198,012,420	7,411,860	Hom	6	7	Y	2+4	5+4	Y

SC_{1L}: number of split clones on the left breakpoint of the first simulation, SC_{2L} for the second simulation (complex rearrangements). The two split clone numbers for left/right breakpoints in the second simulation are separately shown for maternal and paternal homologs, as the deletions and duplications are simulated as heterozygous. $Findable_1$ shows if there are sufficient number of split clones spanning the inversion breakpoint in the first simulation. Implanted inversions may be on one of the homologs (genotype=Het), or both (genotype=Hom). P: paternal, M: maternal copy.

observed in both simulation and real data sets that due to segmental duplications and deletions around the breakpoints, SET-COVER approximation selected only one of the inversion breakpoints correctly (Supplementary Note). We therefore formulate the problem as finding maximal quasi cliques in the inversion cluster graph. This formulation allows some incomplete clusters, and tolerates some false split clones to be included in a true cluster, and as a result, increases flexibility and avoids getting stuck in a local optimum.

We construct a graph $G = (V, E)$ as follows. Each node in the graph denotes a pair of split clones (PSC) that are compatible with each other, as explained above, and each PSC will therefore represent a potential pair of inversion breakpoints. We put an edge between nodes if two different PSCs agree with breakpoint locations through simple intersection. Formally,

$$V = \{v_1, v_2, \ldots, v_m,\ \forall i,\ v_i\ denotes\ a\ paired\ split\ clone\}$$

$$E = \{(v_m, v_n) \mid breakpoints(v_m)\ intersect\ with\ breakpoints(v_n)\}$$

To find an approximate solution for the maximal quasi clique problem, we use an approximation algorithm previously suggested by Brunato *et al.* (2008), and we set the *tabu*, γ, and λ parameters to *10 rounds*, 50%, and 60%, respectively. We obtained the values for these parameters by another grid optimization on experimental graphs depicting worst case scenarios (Supplementary Note).

When a quasi clique is found, the nodes within the clique denote a set of PSCs that are clustered together to mark an inversion. The breakpoint of this cluster is obtained by intersecting its split clones using a heuristic based on read pair support and the gap size. Next, the read pair support for the breakpoints within a distance is recalculated using the discordant read pairs. All clusters are then checked for any overlap on one side of the breakpoints and only the one with larger read support to split clone support ratio is kept and the rest are discarded. We propose to use this ratio to ensure fairness for less covered regions due to either random mapping or repeated regions. We report the final clusters after removing those that intersect with duplications and assembly gaps (>40%).

In addition, when we use mrFAST to map the paired-end reads, we can also make use of the multiple possible mapping locations of the read pairs. The alternative map locations with the inversion signature then can be given to the dipSeq algorithm. However this file is not precise and applying a cutoff for the edit distance is recommended.

3.5 Experimental validation

The presence of an inversion was tested in the cell line of the NA12878 individual predicted to carry an inverted haplotype. Metaphase fluorescent in situ hybridization (FISH) validation was used for inversions larger than 2 Mbp using two probes located inside of the inversion. Interphase triple-color FISH was used to validate inversions smaller than 2 Mbp and larger than 500 Kbp using two probes inside and one outside the inversion.

4 RESULTS

We applied our algorithm to discover large inversions using two simulated and one real data set. The first simulation aims to both estimate the minimum read coverage requirements for accurate reconstruction of clones, and the effectiveness of our algorithms in large inversion discovery. We designed the second simulation to understand how dipSeq behaves in the presence of other structural variants that may have similar split clone signatures. We finally used dipSeq to discover large inversions in the genome of an individual of Northern European descent (NA12878). For the real data, we compared our results with the InvFEST database of known inversions (Martínez-Fundichely *et al.*, 2014), and we applied experimental validation for the novel inversion calls.

Table 2. Deletions implanted on chromosome 1.

No.	Locus (Mbp)	Length (Mbp)	Genotype	Site
1	4.5-4.67	0.17	Hom	Inv1
2	4.68-4.7	0.02	Hom	Inv1
3	6.5-6.9	0.4	Het (P)	Inv1
4	7.0-7.6	0.6	Het (P)	Inv1
5	65-69.5	4.5	Het (M)	Inv2
6	72-73	1	Het (P)	Inv2

Deletions are simulated as either heterozygous or homozygous (genotype, P: paternal, M: maternal copy for heterozygous simulations). Site: the ID of the closest implanted inversion (Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/015156; this version posted February 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4.1 Simulation Experiments

We designed two simulation experiments to test and demonstrate the power of dipSeq for inversion discovery.

Simulation 1. First, we randomly implanted 8 large inversions (500 Kbp to 10 Mbp) to the human reference genome (GRCh37) chromosome 1. Half of the simulated inversions were homozygous, and the remaining were heterozygous (Table 1). We then randomly selected BAC-sized intervals (μ = 150 Kbp, σ = 40 Kbp) from both chromosome 1 homologs at 4× physical coverage, which we then randomly placed into 288 pools.

We then simulated paired end reads of length 100 bp (fragment size μ = 600 bp, σ = 60 bp) using `wgsim`. We generated three different data sets at 3×, 5×, and 10× depth of coverage to investigate the effect of read depth in our inversion calls. We mapped the reads to the reference genome using both BWA and mrFAST aligners and applied our clone reconstruction method. We were able to correctly infer 87.18% and 86.40% of the clones that were not located on the breakpoints using the BWA and mrFAST alignments, respectively (Supplementary Note). Using the inferred clones, dipSeq could find all 8 inversions. It performed similarly in terms of sensitivity at all levels of depth of coverage, and returned no false positives.

Simulation 2. As a second simulation test, we explored dipSeq's performance when there are other forms of structural variation close to or intersecting the inversion breakpoints, therefore simulating complex rearrangements. We used the same simulated inversions (Table 1), and we additionally implanted deletions (Table 2) and duplications (Table 3). We also inserted two additional inverted duplications to test whether dipSeq would predict them as normal inversions. We then repeated our clone and paired end read simulation as explained above. However, due to random simulation, one of the inversion breakpoints were not "findable". i.e. no clones spanned the breakpoint (Table 1). After clone reconstruction, dipSeq was able to find all remaining 7 inversions correctly even at 3× sequence coverage. It is, however, interesting that when we increased the read depth to 15× and 20×, the predicted breakpoints for one of the inversions shifted a few basepairs from the real ones. This may be due to an increase in spurious mapping within repeat regions, thus increasing the chance of observing reads that falsely map in close proximity. In addition, dipSeq did not incorrectly identify inverted duplications as bona fide inversions. We provide our algorithm to prevent such incorrect calls in the Supplementary Note.

Whole genome sequencing based inversion calling We then tested the efficacy of using whole genome sequencing (WGS) based inversion discovery algorithms. For this purpose, we simulated WGS data sets, again using `wgsim`, at 3×, 5×, 10×, 15×, and 20×. Next we mapped the simulated reads to the reference human genome (GRCh37) with both BWA and mrFAST, to test the detection performance of three algorithms: INVY (Rausch *et al.*, 2012), LUMPY (Layer *et al.*, 2014), and VariationHunter (Hormozdiari *et al.*, 2009). We used the BWA alignments for INVY and LUMPY, and mrFAST alignments for VariationHunter, as per each tool's usage recommendations. As expected, INVY and LUMPY failed to discover any of the implanted inversion events, as they are mainly designed for finding shorter inversions. VariationHunter was able to identify Inv5 (Table 1), which may be due to VariationHunter's ability to incorporate all map locations, and a higher maximum inversion size threshold.

4.2 Real data set from NA12878

Next, we tested dipSeq using a real pooled clone sequencing data set generated from the genome of NA12878. We mapped paired-end reads from a total of 288 pools (Methods) using both BWA and mrFAST to the reference genome. Average fragment length of the paired end reads was ∼450bp, with a standard deviation of 98bp. Using our algorithms, we reconstructed the clone locations, which showed an average clone length of ∼140 Kbp and a standard deviation of 40 Kbp.

For inversion discovery, we set the minimum and maximum inversion size thresholds as 500 Kbp and 10 Mbp, respectively. Although it is theoretically possible to detect inversions as small as a typical BAC size (150-200 Kbp), we chose the minimum size as 500 Kbp due to the limitations of the FISH method we used for validation (Methods). We permitted a gap between paired regions up to 200 Kbp, and an overlap of 1Kbp (Figure 1b). After the initial split clone clustering and maximal quasi clique approximation (Methods), we filtered those regions without read pair signature support. We generated two main callsets using BWA and mrFAST, selected a total of 11 inversions with high support for experimental validation (Table 4). We then compared our predictions with the known inversions reported in the InvFEST database (Martínez-Fundichely *et al.*, 2014), and found that dipSeq could correctly identify all 3 inversions that are previously *validated* in the genome of the same individual; a 5 Mbp inversion in 8p23.1 (Antonacci *et al.*, 2009), a 1.5 Mbp inversion in 17q12 (Antonacci *et al.*, 2009), and a 2 Mbp inversion in 15q13.3 (Antonacci *et al.*, 2014) (Table 4). Out of the remaining 8 inversion predictions, 2 could not be tested due to the segmental duplications around the breakpoints. We tested the remaining using FISH experiments (Methods), and validated a novel inversion in 15q25 locus (Figure 4a,b). We also show the visualization of a previously characterized 15q13 inversion (InvFEST ID: HsInv1049) using the SAVANT browser (Fiume *et al.*, 2012) in Figure 4c. We also used dipSeq with different parameters and generated two more data sets, which are not extensively tested (Supplementary Table 3).

5 DISCUSSION

In this paper, we presented a novel algorithm, dipSeq, to characterize large genomic inversions using a new sequencing method initially developed to improve haplotype phasing. Although it suffers from high false positive rate using real data (Table 4), dipSeq was able to identify all previously validated inversion events, and also discover a novel variant. Furthermore, dipSeq performed better with simulated data, suggesting that the relatively poor performance with the NA12878 genome may be improved with higher depth of coverage.

There are multiple directions that we can take to further improve dipSeq. First, to reduce the false discovery rate, we may incorporate split read sequence signature (Ye *et al.*, 2009), and we may perform local *de novo* assembly around the predicted breakpoint intervals with an approach similar to TIGRA (Chen *et al.*, 2014). However, since both of these methods need high sequence coverage, they

bioRxiv preprint doi: https://doi.org/10.1101/015156; this version posted February 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

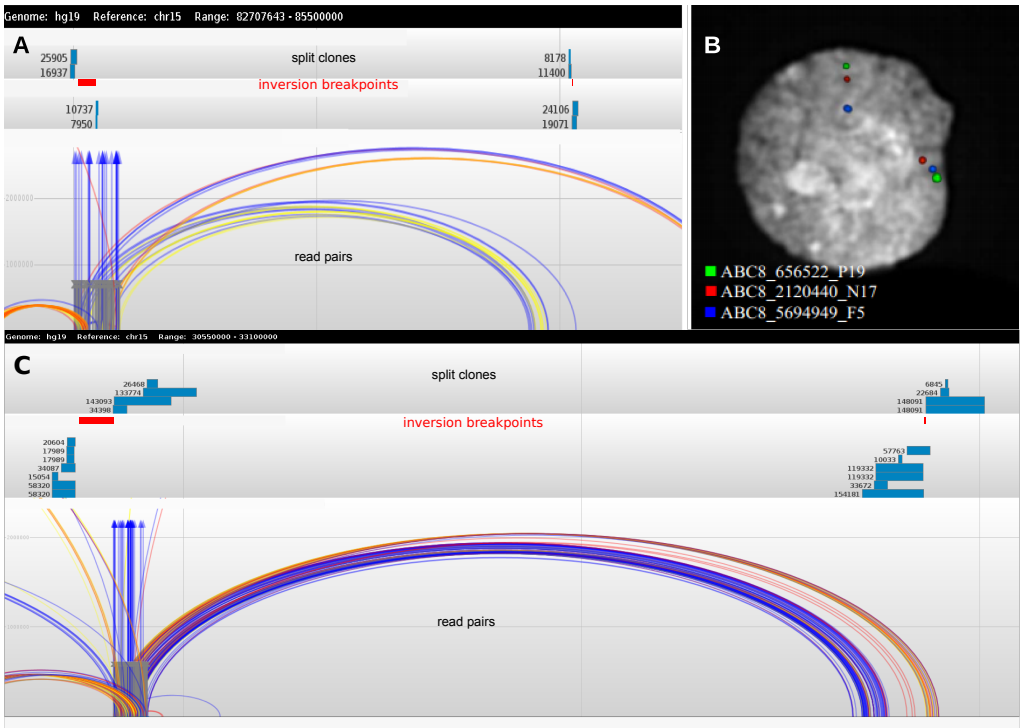

Fig. 4. Inversions discovered by dipSeq in the NA12878 genome. (A) Novel inversion found at chr15:30,433,406-32,898,559 (inner coordinates). We show the locations of split clones and the supporting read pairs using the SAVANT browser (Fiume *et al.*, 2012). (B) Experimental validation of the novel inversion discovered using interphase FISH (green-red-blue: direct, green-blue-red:inverted). (C) SAVANT browser view of the previously known inversion at chr15:83,089,659-84,865,500. SAVANT read pair colors are as follows. Light blue: concordant, red: discordant by length, dark blue: one end inverted, yellow: everted (tandem duplication), gray: one end unmapped.

Table 3. Duplications implanted on chromosome 1.

No.	Target Locus (Mbp)	Genotype (target)	Source Locus (Mbp)	Genotype (source)	Length (Mbp)	Site	Type
1	77	Hom	75-77	Hom	2	Inv3	Direct
2	81	Hom	83-84	Hom	2	Inv3	Direct
3	95	Het (P)	92-94	Het (M)	2	Inv4	Direct
4	97	Hom	98-99	Het (M)	1	Inv4	Direct
5	109	Hom	106.5-107.5	Het (M)	1	Inv5	Direct
6	174	Het (M)	175-177	Het (M)	2	Inv6	Direct
7	200	Hom	Inv7.start-Inv7.end	Hom	3	-	Inverted
8	221	Het (M)	217.8-219	Het (M)	1.2	-	Inverted
9*	223	Het (P)	217.8-219	Het (P)	1.2	-	Inverted

After implanting inversions (Table 1), we copied genomic segments from the one or both of the homologs (source genotype and locus) to the target homolog and loci. Duplications 1-6 were in direct orientation, and 7-9 were inverted. Duplication #7 shares the same breakpoints with Inv7. *The duplication was inserted twice.

might not be suitable to directly apply to the low-coverage data set we used. Instead, it will be better to simultaneously use WGS data generated from the genome of the same individual. Since the PCS method also requires WGS data for haplotype phasing, it can be expected to generate matching PCS-WGS data sets from the same genomes.

dipSeq can also be extended to characterize other forms of large structural variation, including deletions, insertions, direct and inverted duplications. Each of these types of SV present themselves with different split clone signatures that we summarize in Supplementary Figure 10. We also note that, determining the location of a segmental duplication event is yet a largely unsolved problem, even when long reads are used (Chaisson *et al.*, 2014). It may also be possible to discover translocations using split clones, however, chance of finding incorrect split clones will also increase, causing a reduction in the performance of maximal quasi clique approximation.

bioRxiv preprint doi: https://doi.org/10.1101/015156; this version posted February 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 4. Summary of inversions predicted in the genome of NA12878 using dipSeq.

InvFEST ID	chrom	start	end	size	status	dipSeq$_{BWA}$	dipSeq$_{mrFAST}$
-	chr2	87,588,105	90,486,279	2,898,174	not tested	No	Yes
-	chr7	72,617,615	74,198,333	1,580,718	not confirmed	No	Yes
HsInv0501	chr8	6,922,489	12,573,597	4,695,489	**confirmed (InvFEST)**	Yes	Yes
-	chr10	48,910,559	51,462,948	2,552,389	not tested	Yes	No
-	chr15	23,564,857	28,612,825	5,047,968	not confirmed	Yes	Yes
HsInv1049	chr15	30,370,112	32,899,708	2,032,351	**confirmed (InvFEST)**	Yes	Yes
-	chr15	83,089,462	84,865,500	1,776,038	**confirmed (novel)**	Yes	Yes
-	chr16	21,931,735	30,233,983	8,302,248	not confirmed	No	Yes
HsInv1048	chr17	34,725,850	36,295,000	1,512,198	**confirmed (InvFEST)**	Yes	Yes
-	chr22	18,766,919	21,560,727	2,793,808	not confirmed	No	Yes
-	chr22	20,697,917	21,505,192	807,275	not confirmed	Yes	No

dipSeq returns four coordinates for each inversion for the two breakpoint estimations. The coordinates above are the inner breakpoint predictions, and are from the GRCh37 reference genome. The InvFEST database reports inversions in NCBI build 36 coordinates, however, we converted the coordinates using the `liftOver` tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). We note that the novel inversion we predicted and confirmed is listed in the InvFEST database (ID: HsInv0547), as "unreliable prediction". dipSeq$_{BWA}$: predictions using BWA alignments, dipSeq$_{mrFAST}$: predictions using mrFAST alignments (edit distance ≤ 4) and all possible map locations for read pairs. We tested dipSeq using two other mapping parameters which returned slightly different results (Supplementary Table 3).

In summary, dipSeq is the first algorithm that can discover large genomic inversions using high throughput sequencing technologies. Our understanding of the phenotypic effects of inversions is still limited, and one of the reasons of this is the lack of reliable and cost effective methods to characterize such events. This is also true for other complex rearrangements such as duplications and translocations. Improvements in characterization of large complex rearrangements will help us better understand the biological mechanisms that lead to phenotypic difference, disease, and evolution.

ACKNOWLEDGMENTS

We would like to thank J. Kitzman, and B. Dumont for data access and their valuable input for the clone reconstruction algorithm, and J. Huddleston for computational assistance. We also thank M.C. Orhan for the discussions on probability calculations for clone reconstruction, B. Genç and M. Sarı for their help with the initial formulations of the problem, C. Arbib for discussions on how to formulate inversion events with estimated breakpoint intervals with SET-COVER, and M. Cáceres and S. Casillas for their help with the InvFEST database.

Funding: Funding for this project was provided by a Marie Curie Career Integration Grant (303772) and a TÜBİTAK grant (112E135) to C.A., an NIH grant (HG004120) to E.E.E., and a Firb-Programma "Futuro in Ricerca" grant (RBFR103CE3) to M.V. C.A. also acknowledges support from The Science Academy of Turkey, under the BAGEP program, and EMBO (IG-2521).

Conflict of Interest: E.E.E. is on the scientific advisory board for DNAnexus, Inc.

REFERENCES

Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X., and Shendure, J. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol*, **11**(12), R119.

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J. O., Baker, C., Malig, M., Mutlu, O., Sahinalp, S. C., Gibbs, R. A., and Eichler, E. E. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet*, **41**(10), 1061–1067.

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet*, **12**(5), 363–376.

Antonacci, F., Kidd, J. M., Marques-Bonet, T., Ventura, M., Siswara, P., Jiang, Z., and Eichler, E. E. (2009). Characterization of six human disease-associated inversion polymorphisms. *Hum Mol Genet*, **18**(14), 2555–2566.

Antonacci, F., Kidd, J. M., Marques-Bonet, T., Teague, B., Ventura, M., Girirajan, S., Alkan, C., Campbell, C. D., Vives, L., Malig, M., *et al.* (2010). A large and complex structural polymorphism at 16p12. 1 underlies microdeletion disease risk. *Nature genetics*, **42**(9), 745–750.

Antonacci, F., Dennis, M. Y., Huddleston, J., Sudmant, P. H., Steinberg, K. M., Rosenfeld, J. A., Miroballo, M., Graves, T. A., Vives, L., Malig, M., Denman, L., Raja, A., Stuart, A., Tang, J., Munson, B., Shaffer, L. G., Amemiya, C. T., Wilson, R. K., and Eichler, E. E. (2014). Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nat Genet*, **46**(12), 1293–1302.

Brunato, M., Hoos, H. H., and Battiti, R. (2008). On effectively finding maximal quasi-cliques in graphs. In *Learning and Intelligent Optimization*, pages 41–55. Springer.

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J., and Eichler, E. E. (2014). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*.

Chen, K., Chen, L., Fan, X., Wallis, J., Ding, L., and Weinstock, G. (2014). TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome Res*, **24**(2), 310–317.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., Macarthur, D. G., Macdonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Consortium, W. T. C. C., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W., and Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, **464**(7289), 704–712.

Fiume, M., Smith, E. J., Brook, A., Strbenac, D., Turner, B., Mezlini, A. M., Robinson, M. D., Wodak, S. J., and Brudno, M. (2012). Savant genome browser 2: visualization and analysis for population-scale genomics. *Nucleic acids research*, **40**(W1), W615–W621.

Hormozdiari, F., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res*, **19**(7), 1270–1278.

bioRxiv preprint doi: https://doi.org/10.1101/015156; this version posted February 11, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nat Genet*, **36**(9), 949–951.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tzn, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., and Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**(7191), 56–64.

Kitzman, J. O., Mackenzie, A. P., Adey, A., Hiatt, J. B., Patwardhan, R. P., Sudmant, P. H., Ng, S. B., Alkan, C., Qiu, R., Eichler, E. E., and Shendure, J. (2011). Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol*, **29**(1), 59–63.

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

Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M., and Snyder, M. (2014). Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol*, **32**(3), 261–266.

Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). Lumpy: a probabilistic framework for structural variant discovery. *Genome Biology*, **15**(6), R84.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**(14), 1754–1760.

Martínez-Fundichely, A., Casillas, S., Egea, R., Ràmia, M., Barbadilla, A., Pantano, L., Puig, M., and Cáceres, M. (2014). Invfest, a database integrating information of polymorphic inversions in the human genome. *Nucleic acids research*, (42), D1027–32.

McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., Dallaire, S., Gabriel, S. B., Lee, C., Daly, M. J., Altshuler, D. M., and , I. H. C. (2006). Common deletion polymorphisms in the human genome. *Nat Genet*, **38**(1), 86–92.

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods*, **6**(11 Suppl), S13–S20.

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

Peters, B. A., Kermani, B. G., Sparks, A. B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y. T., Haas, J., Robasky, K., Zaranek, A. W., Lee, J.-H., Ball, M. P., Peterson, J. E., Perazich, H., Yeung, G., Liu, J., Chen, L., Kennemer, M. I., Pothuraju, K., Konvicka, K., Tsoupko-Sitnikov, M., Pant, K. P., Ebert, J. C., Nilsen, G. B., Baccash, J., Halpern, A. L., Church, G. M., and Drmanac, R. (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, **487**(7406), 190–195.

Qi, J. and Zhao, F. (2011). ingap-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic acids research*, **39**(suppl 2), W567–W575.

Quinlan, A. R., Clark, R. A., Sokolova, S., Leibowitz, M. L., Zhang, Y., Hurles, M. E., Mell, J. C., and Hall, I. M. (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res*, **20**(5), 623–635.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., and Korbel, J. O. (2012). Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**(18), i333–i339.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzlez, J. R., Gratacs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, **444**(7118), 444–454.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*, **305**(5683), 525–528.

Sharp, A. J., Cheng, Z., and Eichler, E. E. (2006). Structural variation of the human genome. *Annu Rev Genomics Hum Genet*, **7**, 407–442.

Smith, J. J., Stuart, A. B., Sauka-Spengler, T., Clifton, S. W., and Amemiya, C. T. (2010). Development and analysis of a germline BAC resource for the sea lamprey, a vertebrate that undergoes substantial chromatin diminution. *Chromosoma*, **119**(4), 381–389.

Steinberg, K. M., Antonacci, F., Sudmant, P. H., Kidd, J. M., Campbell, C. D., Vives, L., Malig, M., Scheinfeldt, L., Beggs, W., Ibrahim, M., *et al.* (2012). Structural diversity and african origin of the 17q21. 31 inversion polymorphism. *Nature genetics*, **44**(8), 872–880.

Talkowski, M. E., Ernst, C., Heilbut, A., Chiang, C., Hanscom, C., Lindgren, A., Kirby, A., Liu, S., Muddukrishna, B., Ohsumi, T. K., Shen, Y., Borowsky, M., Daly, M. J., Morton, C. C., and Gusella, J. F. (2011). Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *Am J Hum Genet*, **88**(4), 469–481.

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M. V., and Eichler, E. E. (2005). Fine-scale structural variation of the human genome. *Nat Genet*, **37**(7), 727–732.

Ventura, M., Catacchio, C. R., Alkan, C., Marques-Bonet, T., Sajjadian, S., Graves, T. A., Hormozdiari, F., Navarro, A., Malig, M., Baker, C., Lee, C., Turner, E. H., Chen, L., Kidd, J. M., Archidiacono, N., Shendure, J., Wilson, R. K., and Eichler, E. E. (2011). Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res*, **21**(10), 1640–1649.

Xin, H., Lee, D., Hormozdiari, F., Yedkar, S., Mutlu, O., and Alkan, C. (2013). Accelerating read mapping with fasthash. *BMC Genomics*, **14 Suppl 1**, S13.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**(21), 2865–2871.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res*, **19**(9), 1586–1592.