1 Structural variants

As described in (Volik et al., 2003; Tuzun et al., 2005; Korbel et al., 2007; Kidd et al., 2008; Lee et al., 2008), by examining the mapping span and orientation of paired-end read sequences, one can detect insertion, deletion, inversion, and translocation events in a test genome. Recently, it was shown that tandem repeat expansions can also be detected by end sequence profiling (ESP) (Cooper et al., 2008). In this section, we revisit the definitions of structural variants, and the properties of mate pair mappings that support each kind of variant in Figure 1.

The fundamental part of the structural variation projects is the use of paired-end read sequences from clone inserts that follow a tight length distribution, such as fosmids ($\sim$ 40Kb) and BACs ($\sim$ 150Kb). Similar techniques are used with the NGS technologies, however the insert length differ in various platforms: $\sim 200bp$ in Illumina, $600 - 3,000$bp in SOLiD, and 3Kbp in 454. Both length and orientation discordancies between the left and right ends of each clone insert on the reference genome identify the underlying structural variation event at that site (Figure 1).

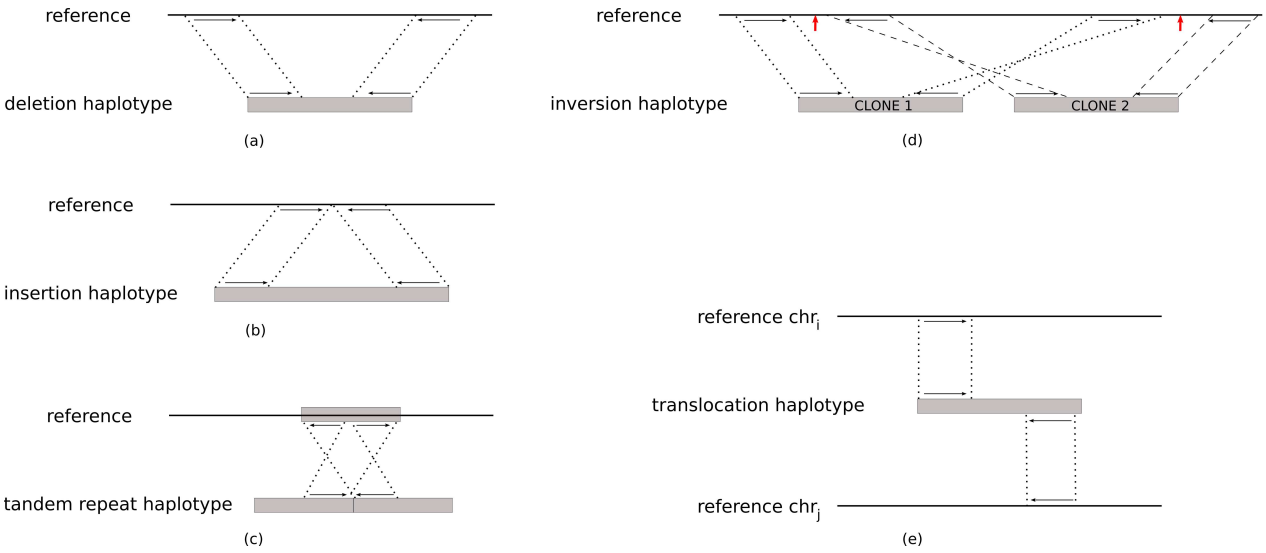


Figure 1: Types of structural variation that can be detected with paired-end sequences: mapped span of paired-end reads appear larger than the expected insert length if there is a (a) deletion, and smaller in an (b) insertion haplotype. Disagreement between the mapping orientations and the sequencing library specifications might either report a (c) tandem repeat, or an (d) inversion. Also, not that in the case of inversions (d), $CLONE_1$ and $CLONE_2$ predict two different inversion breakpoints (shown with arrows), but by examining the map locations and orientations, one can deduce that both clones predict the same inversion, and both breakpoints of the inversion event can be recovered. If the paired-end reads align confidently on different chromosomes, a (e) translocation event is reported. In this figure, we assumed the expected end-sequence orientation properties in capillary based sequencing and Illumina platforms.

2 MPSV problem is NP-hard.

In what follows, we first will (roughly) show that the MPSV problem is NP-hard and then give an $O(logn)$-approximation algorithm for it.

Theorem 1. *MPSV problem is NP-hard.*

Proof. The reduction is from the set cover problem (Karp, 1972). Given a set $U = \{e_1, \ldots, e_n\}$ and $S = \{S_1, S_2, \cdots, S_k\}$, a collection of subsets of U, the set cover problem asks to find the minimum number of sets in S whose union include all $e_i \in U$. The reduction from an instance of a set cover problem to the MPSV problem is as follows:

1. Set $DisCor = U$, that is, for each e_i generate a paired-end read pe_i.
2. For each set S_i set an interval (L_{S_i}, R_{S_i}), which does not overlap with any other such interval.
3. Finally, set $Align(pe_i) = \{(L_{s_j}, R_{s_j}) | \forall S_j : e_i \in S_j\}$.

Clearly, the two problems are equivalent and a subset S' of S is a minimum size set cover of S iff the set of intervals corresponding to S' includes the minimum number of intervals to which each paired-end read pe_i can be mapped to. □

3 Insert size distribution of paired-end reads

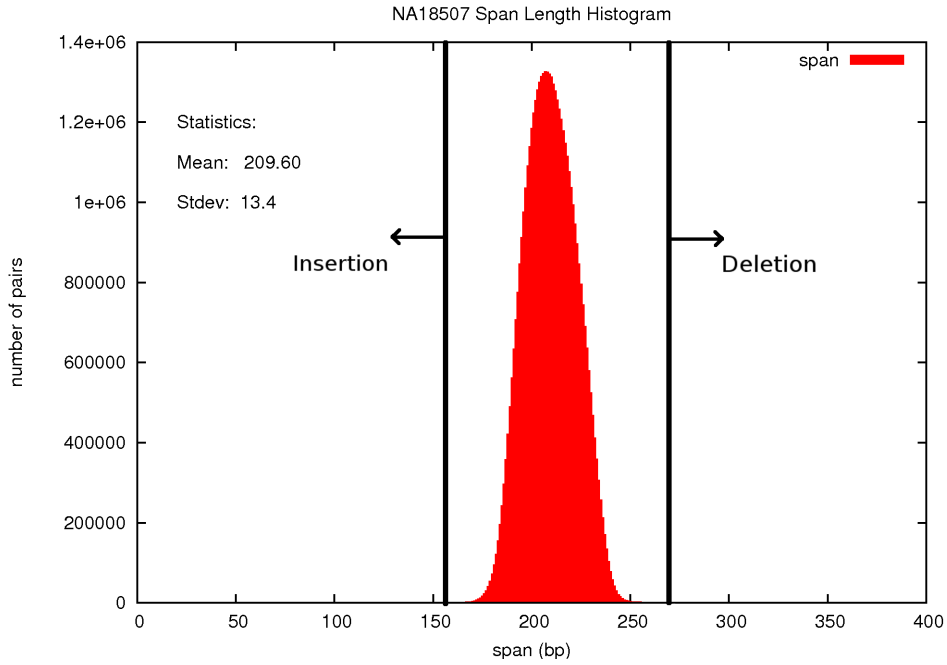
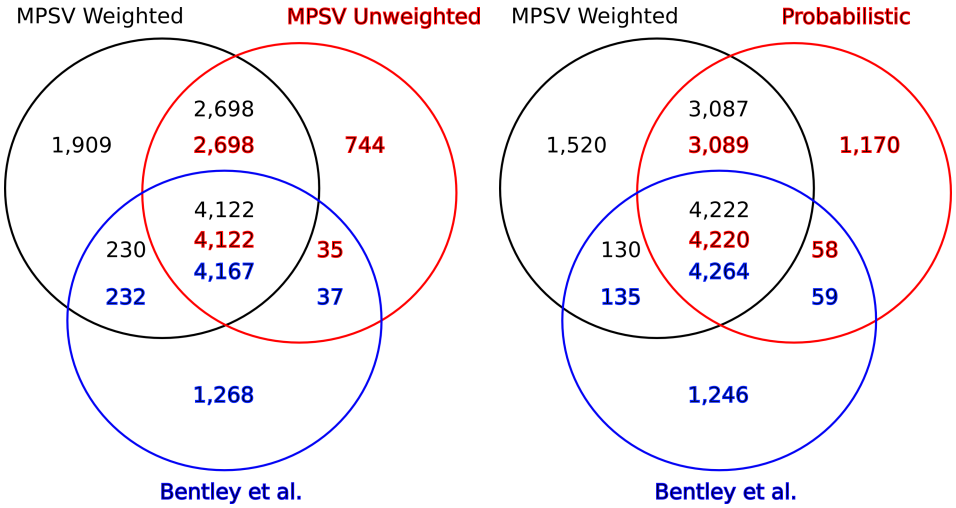


Figure 2: Span length histogram of paired-end reads from NA18507 on human genome build 36. We call a clone insert concordant if its span is within $4 \times std$ of the mean length. For this library, the concordant size cut off values are 155bp and 266bp.

4 Three way comparison of predicted deletions



(a) (b)

Figure 3: Comparison of deletion calls from our VariationHunter-SC/MPSV (both weighted and unweighted versions), and VariationHunter-Pr/probabilistic algorithms and intervals from the original Illumina study (Bentley et al., 2008). Note that Bentley et al. also used long insert libraries (expected clone length = 2Kbp), which were not available for download when we performed our analysis. Four Venn diagrams are presented here: VariationHunter-SC/MPSV (both versions) and Bentley et al. with minimum (a) 50% reciprocal overlap ; and comparison of VariationHunter-SC/MPSV (weighted) and VariaionHunter-Pr/probabilistic methods with the original study with (b) 50% overlap.

5 Mapping statistics

# Sequences	$3,519,246,954$
# HQ Sequences	$2,261,838,984$
unique, e.d.=0	$1,512,419,495$ (66.87%)
unique, e.d.=1	$245,586,578$ (10.85%)
unique, e.d.=2	$60,194,526$ (2.66%)
repeat, e.d.=0	$250,118,990$ (11.07%)
repeat, e.d.=1	$66,094,390$ (2.93%)
repeat, e.d.=2	$35,978,574$ (1.6%)
no match	$91,446,431$ (4.01%)

Table 1: Mapping statistics of the Illumina short insert library from NA18507 (Bentley et al., 2008). We first removed lower quality end sequences prior to mapping stage. Approximately 95.9% of the remaining sequences were mapped to human genome build 36 using our in-house sequence mapper *mrFAST*. Although *mrFAST* provides all possible map locations within an edit distance of 2, we also report the properties (unique vs. repetitive and edit distance) of the best map locations in this table.

References

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., *et al.*, 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**(7218):53–59.

Cooper, G. M., Zerr, T., Kidd, J. M., Eichler, E. E., and Nickerson, D. A., 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet*, **40**(10):1199–1203.

Karp, R. M., 1972. Reducibility among combinatorial problems. *Complexity of Computer Computations*, :85–103.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., *et al.*, 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**(7191):56–64.

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., *et al.*, 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**(5849):420–426.

Lee, S., Cheran, E., and Brudno, M., 2008. A robust framework for detecting structural variations in a genome. *Bioinformatics*, **24**(13):i59–i67.

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., *et al.*, 2005. Fine-scale structural variation of the human genome. *Nat Genet*, **37**(7):727–32.

Volik, S., Zhao, S., Chin, K., Brebner, J. H., Herndon, D. R., Tao, Q., Kowbel, D., Huang, G., Lapuk, A., Kuo, W. L., *et al.*, 2003. End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc Natl Acad Sci U S A*, **100**(13):7696–701.