An assembly approach utilizing next and third generation sequencing data for powerful structural variant detection

Xian Fan[1,2], Zechen Chong[2], Luay Nakhleh[1], Human Genome Structural Variation Consortium, Ken Chen[1,2*]

[1]Department of Computer Science, Rice University, Houston, Texas (USA)
[2]Department of Bioinformatics and Computational Biology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas (USA)

Background

Detection of structural variations (SV), including both large and small (<50 bp) ones, is important in understanding human genetic diseases. Conventional approaches that utilize next generation sequencing (NGS) technologies (such as Illumina) have limited detection power due to short read length. Third generation sequencing (TGS) technologies, such as the Pacific BioScience (Pacbio) single molecule sequencing technology, facilitate SV identification by generating much longer reads. However, the long reads produced by TGS often have high sequencing error rates (~15%), which leads to 1) imprecise alignment to the reference, and 2) challenges in detecting small SVs.

It is therefore reasonable to develop computational approaches that combine the advantages of the NGS and TGS data in order to further improve the detection of SVs. However, while very few algorithms have been developed for this task thus far, none jointly utilizes both types of reads for discovery on the sequence level, nor targets novel insertions and small INDELs.

Result

We developed a hybrid assembly-based approach that utilizes both Illumina and Pacbio reads to discover SVs. The approach starts with an Illumina BAM file and Pacbio raw subreads. It extracts Illumina reads that cannot be well aligned to the reference, and aligns these reads to all Pacbio reads, aiming to extract Pacbio reads that span SVs. This process requires an aligner with high sensitivity in spite of high error rate in Pacbio reads and short length in Illumina reads. We utilized a customized version of BLASR for this purpose that achieved >90% success rate (percentage of the Illumina reads that have at least one high quality alignment to Pacbio reads). The pairwise alignments between Illumina and Pacbio reads form a bipartite graph, in which nodes represent the reads (Illumina and Pacbio reads are the two partite sets), and edges correspond to matches by alignment. We cluster the graph into connected components using a near linear graph-theoretic union-find algorithm. Each connected component contains a set of reads (including both Illumina and Pacbio) that have shared homology and likely originate from the same SV. We apply Celera Assembler to assemble the Pacbio reads in each connected component and produce contigs representing reconstructed alternative alleles. We align the contigs to the reference and identify putative SV breakpoints. Finally,

Illumina reads in the corresponding connected component are aligned to the assembled contigs to confirm the existence of the breakpoints. This method allows us to detect SVs of a wide range of sizes (11bp to > 10kbp), particularly INDELs in Short Tandem Repeats (STR) and large novel insertions.

To evaluate this approach, we ran it on the Pacbio and Illumina data generated from a haploid hydatidiform mole (CHM1) genome. An SV call set (A) was previously generated from the Pacbio data by a reference-alignment guided local assembly approach by Chaisson et al. We also generated SV call sets using Delly (B) and Lumpy (C) from Illumina data only. Our algorithm utilized 0.7% Illumina and 9% Pacbio reads and identified 3,268 large deletions (>50bp), 5,651 large insertions (>50bp), 13,223 small deletions (<=50bp), and 14,715 small insertions (<=50bp). 72% of large deletions identified by our method were also identified by at least one other method, which indicates a high specificity of our method. Additionally we detected 826 unique calls, which overlap well with known SVs in database of genomic variants (DGV, 87%) and STRs (65%), indicating a high sensitivity and specificity of our approach. Our method also identified 14 large (>500 bp) novel insertions (relative to build37) missed by Chaisson et al. but validated by build38. To evaluate small INDELs, we compared with Pindel and GATK. 70% deletions and 76% insertions in our call set were identified by at least one other method. The 3,933 novel deletions and 3,347 novel insertions we identified overlapped well with dbSNP (87% for deletion and 86% for insertion) and STR annotations (89% for deletion and 67% for insertion), indicating a high sensitivity and specificity of our approach.

We further applied our method to the three trios (YRI, PUR and CHS) in the 1000 Genomes Project (or Human Genome Structural Variation Consortium). Both Illumina and Pacbio reads were available for these trios. On average, we called ~26,000 SVs per sample and 20% of our calls are novel with respect to 4 other methods that analyze either only Illumina data (e.g., Delly, Pindel, Manta) or only Pacbio data.

Conclusion

We developed a novel method for SV detection through joint utilization of both NGS and TGS coverage at raw read level. Results obtained from analyzing a single haploid and 3 trio human samples indicate that our method can utilize the advantages of two platforms (accuracy of the Illumina reads and the length of the Pacbio reads), and generate high accuracy SVs with novel calls. In particular, our method can detect SVs of a wide size range, from 11bp to >10kbp, and is particularly effective at detecting large novel insertions not present on the reference, and small INDELs, which are challenging to other methods.