GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly

Authors: Daniel L Cameron, Anthony T Papenfuss

Background

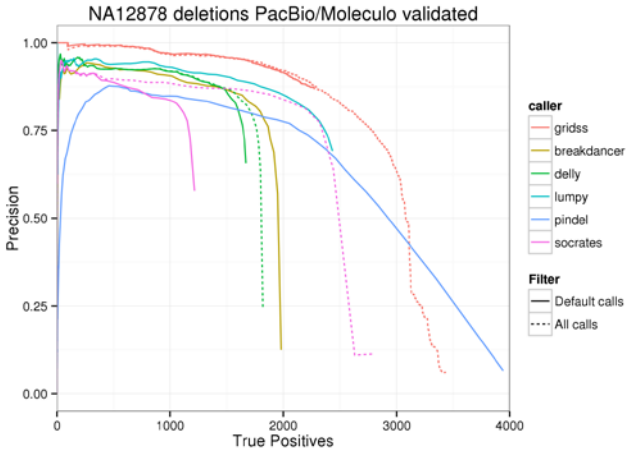
Many methods exist to identify structural variants (SVs) using high-throughput sequencing data with most methods using one or more of four approaches: read depth analysis (RD), discordantly-aligned read pair clustering (DP), split reads identification (SR), and assembly. RD approaches (e.g. CNVnator, Control-FREEC) are limited in their event size detection and cannot resolve breakpoint partners. DP approaches (e.g. BreakDancer, GASVPro) can be used to infer the presence of SVs but cannot in general identify exact breakpoint locations since the breakpoints occur in the unsequenced part of the fragments whereas SR approaches (eg CREST, Socrates) can obtain single nucleotide resolution by identifying breakpoint-spanning reads. Assembly-based methods perform either *de novo* assembly (e.g. cortex_var), targeted assembly based on previously identified candidates (e.g. SVMerge, TIGRA), or perform windowed assembly to detect small events (e.g. DISCOVAR, SOAPindel). These approaches are not mutually exclusive with some software incorporating two (e.g. DELLY) or three (e.g. LUMPY) of these approaches.

Here we describe GRIDSS, the Genome Rearrangement IDentification Software Suite, composed of an assembler, and a variant caller which combines assembly, split read and read pair evidence to identify structural variants. Our novel assembly approach performs genome-wide breakend assembly (that is, independent assembly of each side of each breakpoint) by using a genome-wide positional de Bruijn graph. Soft-clipped reads, split read, discordant read pairs, and read pairs with only one read mapped are assembled into the positional de Bruijn graph with the mapping locations of each read encoded as positional constraints within the graph itself. Post-assembly, we use the same realignment approach used to identify split reads from soft-clipped reads to identify the breakpoint supported by each breakend contig. Once identified, we used a probabilistic model to call variants from supporting assembly contigs, split reads, and discordant read pairs.

Results

To benchmark GRIDSS, we compared against BreakDancer, DELLY, LUMPY, Pindel, and Socrates on both simulated data and well-characterised cell lines. We simulated deletions, insertions, inversion, tandem duplication, and genomic fusions on 2x100bp sequencing at

varying levels of coverage. Above 8x coverage, GRIDSS sensitivity exceeds that of the other callers for events larger than 100bp, with Pindel showing highest sensitivity for small events. To compare performance on realistic data, we evaluated the callers on the Illumina Platinum Genomics 50x WGS NA12878 data (See Figure).



GRIDSS is able to almost halve the false discovery rate compared to other callers, with the highest scoring GRIDSS calls having a FDR close to zero. GRIDSS's execution time of 236 minutes is comparable to the 52, 82, 211, 489, and 2,184 minutes of SOCRATES, BreakDancer, LUMPY, DELLY, and Pindel respectively.

We have applied GRIDSS in multiple cancer contexts. Firstly, we have used GRIDSS to identify patient-specific somatic breakpoints. Secondly, we have used the single nucleotide precision of GRIDSS to identify complex compound rearrangements misclassified as simple events by a DP-based caller in 64 variants (5%) in tumour neochromosomes. Thirdly, we are using the multi-sample capability of GRIDSS to reconstruct somatic phylogenetic trees in both mouse xenograft and patient tumours.

Conclusion

GRIDSS achieves high sensitivity and specificity on simulated, cell line and patient tumour data. On NA12878 cell line data, GRIDSS halves the false discovery rate compared to other recent methods.

Our novel incorporation of assembly, split read and read pair evidence in the variant calling process is made possible by our approach of independently assembling each breakend. By using a genome-wide positional de Bruijn graph, we are able to perform untargeted assembly an order of magnitude faster than existing approaches. GRIDSS can perform combined variant discovery on multiple related samples and population data. GRIDSS is freely available at https://github.com/PapenfussLab/gridss.