In Silico Simulation of Low Allele Fraction Gene Rearrangement Detection with Deep Targeted DNA Sequencing

Onur Sakarya[1], Hyunsung John Kim[1], Roger Jiang[1], Tom Chien[1], Payal Shah[1], Hui Xu[1], Chenlu Hou[1], Byoungsok Jung[1], Xiaoyu Chen[2], Han-Yu Chuang[2], and Catalin Barbacioru[1]
[1] GRAIL Inc., Redwood City, CA, 94063
[2] Illumina Inc. San Diego, CA, 92122

Background:
Gene rearrangements are prominent somatic mutations driving cancers. Recent studies identified an increasing number of recurrent gene rearrangements in solid tumors. For example, more than 5% of patients with non-small cell lung cancer (NSCLC) harbor a rearrangement of ALK, ROS1, or RET genes each with multiple partners. Catalogue of Somatic Mutations in Cancer (COSMIC) database provides a curated list of gene rearrangements (*1*). Most of COSMIC gene rearrangement cases are based on RNA sequencing and report the fused exon coordinates of partner genes. However, at DNA level, most breakpoints happen on introns more of than on exons. Furthermore, most introns are in proximity of homologous, low complexity and repeat sequences. Thus, it is more challenging to detect gene rearrangements at DNA level.

We tackled the problem of estimating sensitivity of targeted DNA gene rearrangement detection as a function of breakpoint location within expected introns. We generated random breakpoint templates and simulated artificial reads from these templates in a titration setting. Simulated reads were titrated at low levels to background canonical intron reads to imitate circulating cell-free (cfDNA) setting. We processed the reads through our rearrangement calling pipeline to evaluate the sensitivity and specificity. We also tested our method on real sequencing data from titrated cell lines, a cell line mix and cfDNA from metastatic NSCLC patients whose tissue biopsy confirmed certain oncogenic gene rearrangements.

Results:
We simulated rearrangement breakpoints from 215 COSMIC rearrangements spanning 360,495 base pairs in 87 introns of 28 genes. We required a breakpoint to be at least 500bp apart from an existing breakpoint that was already in the simulation pool. Rearrangements were titrated from 0.2 to 5% within 3000 fragments covering each breakpoint. We repeated the simulation 100 times for each titration level, each time permuting the order of rearranged genes. For each fragment, we simulated 150bp paired-end reads from 167bp long (σ=50) fragments using HiSeq 2500 error profile with ART 2.3.7 (*1*). A custom pyflow (*2*) pipeline was used to map reads with bwa 0.7.10-r789 (*3*) and call breakpoints with Manta-0.29.3 software (*4*). Manta is a two step structural variant detection algorithm based on construction of a breakpoint graph followed by local assembly of individual regions, contig alignment, scoring and calling.

Sensitivity of rearrangement detection was above 99% at 1 to 5% allele fraction (AF), 98% at 0.5% AF, and 73% at 0.2% AF. In general, precise location of the breakpoint was more difficult to detect at lower AF due to lower number of reads going into the assembly process. We

required three paired-end reads as the threshold evidence to initiate the assembly process, which gave an approximate location in the absence of single reads spanning the breakpoint. Improvements to preciseness of the calls were demonstrated with improvements to assembly process. False discovery rate was 0.3% overall for all titration levels.

As real test cases, we deep-sequenced 8 plasma cfDNA samples with known tumor gene rearrangements (EML4>ALK, KIF5B>ALK and CD74>ROS1). We called the associated breakpoint from plasma cfDNA in all cases. Detected breakpoint AF ranged from 0.4 to 12%. There were no false positive breakpoints detected. Two of the cases were biological replicates, i.e. two tubes of whole blood from the same patient. Translocations usually create two reciprocal breakpoints and in one of the replicate cases, both samples had reciprocal calls. In the other replicate case, one sample had a reciprocal call and its replicate sample did not, suggesting reciprocal events may exist in the absence of a reciprocal call. We also sequenced and called gene rearrangements from individual titrated cell lines HCC78 (SLC34A2> ROS1) and H2228 (EML4>ALK) at different input titration levels and Horizon HD753 Structural Variant mix (CCDC6>RET and SLC34A2>ROS1) at AF in the range of 2 to 4%.

Conclusions:

We simulated majority of gene rearrangement breakpoints documented in COSMIC and demonstrated performance of a structural variant calling pipeline at cfDNA setting to achieve high sensitivity and low false discovery rate. We further investigated the performance of oncogenic rearrangement calls from patient plasma cfDNA samples and their localization and reciprocity.

References:

1. Forbes et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Research (2015) 43 (D1): D805-D811.
2. Huang, Weichun et al.ART: A next-Generation Sequencing Read Simulator. Bioinformatics (2012) 28 (4): 593–594.
3. Pyflow – a lightweight parallel task engine. https://github.com/Illumina/pyflow
4. Li, Heng, and Richard Durbin. Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform. *Bioinformatics* (2009) 25 (14): 1754–1760.
5. Chen, Xiaoyu et al. Manta: Rapid detection of structural variants and indels for clinical sequencing applications. Bioinformatics (2016) 32 (8): 1220-1222.