Abstract #5515

Structural variant detection with long-read sequencing reveals driver and passenger mutations in a melanoma cell line

Aaron Wenger[1], Meredith Ashby[1], Marcel Nelen[2], Wigard Kloosterman[3]
1. PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025; 2. Radboud University Medical Center, 6500 HB Nijmegen, Netherlands; 3. University Medical Center Utrecht, 3584 CG Utrecht, Netherlands

Structural variation and long-read sequencing

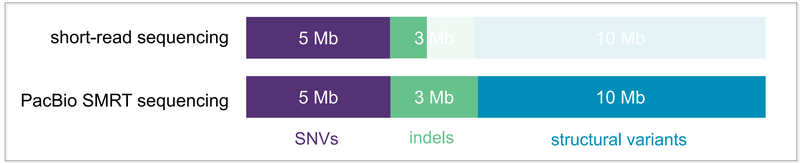


Figure 1. Variation in a typical germline human genome[1]. Most of the base pairs that differ between two human genomes are in indels 1-49 base pairs and in structural variants (SVs), differences ≥50 base pairs. Short-read sequencing has limited sensitivity for indels and SVs, while PacBio long-read sequencing comprehensively detects variants of all sizes.

Long Reads average up to 30 kb
High Consensus Accuracy random errors produce QV50 consensus
Uniform, Unbiased Coverage no GC% or sequence complexity bias
Epigenetic Characterization simultaneous detection of DNA methylation
Single-Molecule Resolution directly measure individual DNA molecules

Figure 2. Advantages of sequencing on the PacBio Sequel Systems.

Genome in a Bottle germline structural variation benchmark

The Genome in a Bottle Consortium[2] has developed a benchmark set of insertion and deletion structural variants (http://tinyurl.com/GIABSV06) in a human male, HG002/NA24385. Comparing technologies against this benchmark, PacBio has the highest precision and recall across the structural variant size range, and particularly for insertions.

Technology	Precision	Recall
PacBio	96.13%	95.99%
Oxford Nanopore	83.23%	87.46%
Illumina	85.35%	55.88%
10X Genomics	83.79%	39.83%

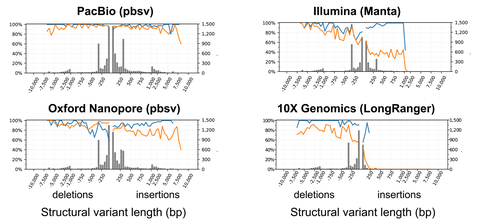
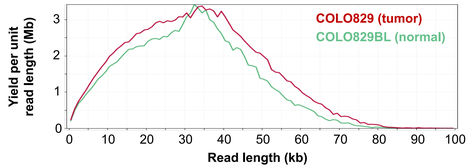


Figure 3. Variant calling performance against the GIAB HG002 v0.6 benchmark. Histograms indicate the number of variants and lines show the precision (blue) and recall (orange) at each variant size for call sets from different technologies.

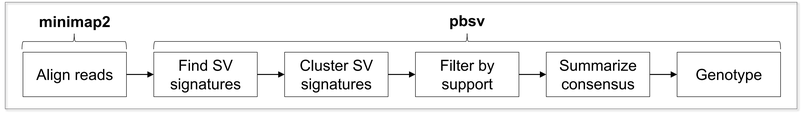
Long-read sequencing of COLO829 tumor and normal on PacBio Sequel System

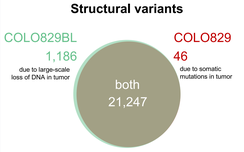
A highly-mutated melanoma cell line (COLO829) and matched normal lymphoblastoid line (COLO829BL) have been proposed as a benchmark for somatic variant detection, as the cell lines are readily available from ATCC[3]. Prior high-coverage Illumina sequencing identified >35,000 somatic point mutations and 446 somatic indel variants present in COLO829 but not COLO829BL. To evaluate germline and somatic structural variants in COLO829 and COLO829BL, we sequenced the cell lines to around 50-fold coverage on the PacBio Sequel System.

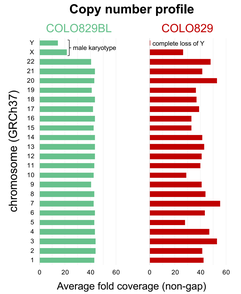
Metric	COLO829	COLO289BL
Yield	146.6 Gb	154.9 Gb
Reads	9,673,612	10,500,345
Read length N50	29,161 bp	26,950 bp

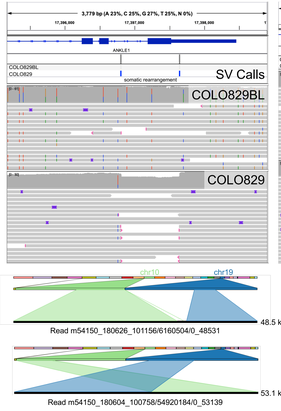


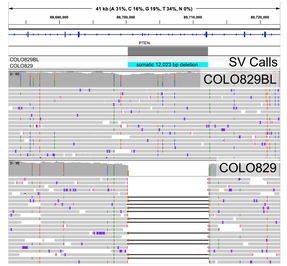
Germline and somatic structural variants in COLO829







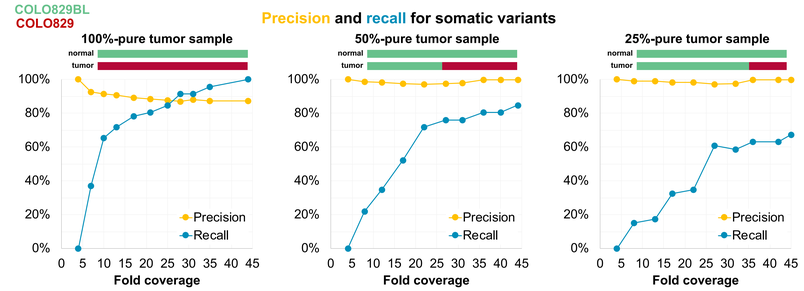




Somatic variants in cancer genes.
Center: Rearrangement involving the last exon of *ANKLE1*, a DNA damage repair nuclease linked to breast and ovarian cancer risk.
Right: Homozygous deletion that includes exons of *PTEN*, a tumor suppressor that is frequently mutated in cancers.

Long-read coverage requirements and tumor sample purity

To evaluate the coverage required to detect somatic variants at different tumor sample purity levels, a stringent (≥6 variant reads in tumor, 0 in normal) set of 46 somatic variants from the full-coverage, pure input sample were treated as a benchmark truth set. Tumor purity was titrated by mixing sequences from the tumor and normal samples. Coverage was titrated by subsampling (same in tumor and normal), and putative somatic variants were called as variants present in the tumor but not normal (≥2 variant reads in tumor, 0 in normal). Variant calls in the truth set were considered true positives; others were considered false positives. Precision=TP/(TP+FP), Recall=TP/(TP+FN).



Conclusions

- PacBio long-read sequencing has the highest precision and recall for structural variants of evaluated sequencing platforms.
- PacBio long-read sequencing of the cancer cell line COLO829 and its matched normal COLO829BL shows large-scale copy-number changes and tens of somatic structural variants, including a rearrangement of chr10 and chr19 and a 12 kb deletion that impacts *PTEN*.
- Even with permissive criteria for calling somatic structural variants, precision is high across a range of tumor purity and coverage. Recall is high at 20-fold coverage in samples that are at least half-tumor and nearly saturates by 30-fold coverage at all purity levels.

References

1. Chaisson MJ, et al. (2017). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv*. doi:10.1101/193144.
2. Zook JM, et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*. 3:160025.
3. Craig DW, et al. (2016). A somatic reference standard for cancer genome sequencing. *Sci Rep*. 6:24607.

Thank you to David Scherer, Kristin Robertshaw, and Pamela Bentley Mills for poster production support. Thank you to the Genome in a Bottle Consortium for the structural variant benchmark. Thank you to collaborators who are analyzing COLO829 with other platforms.