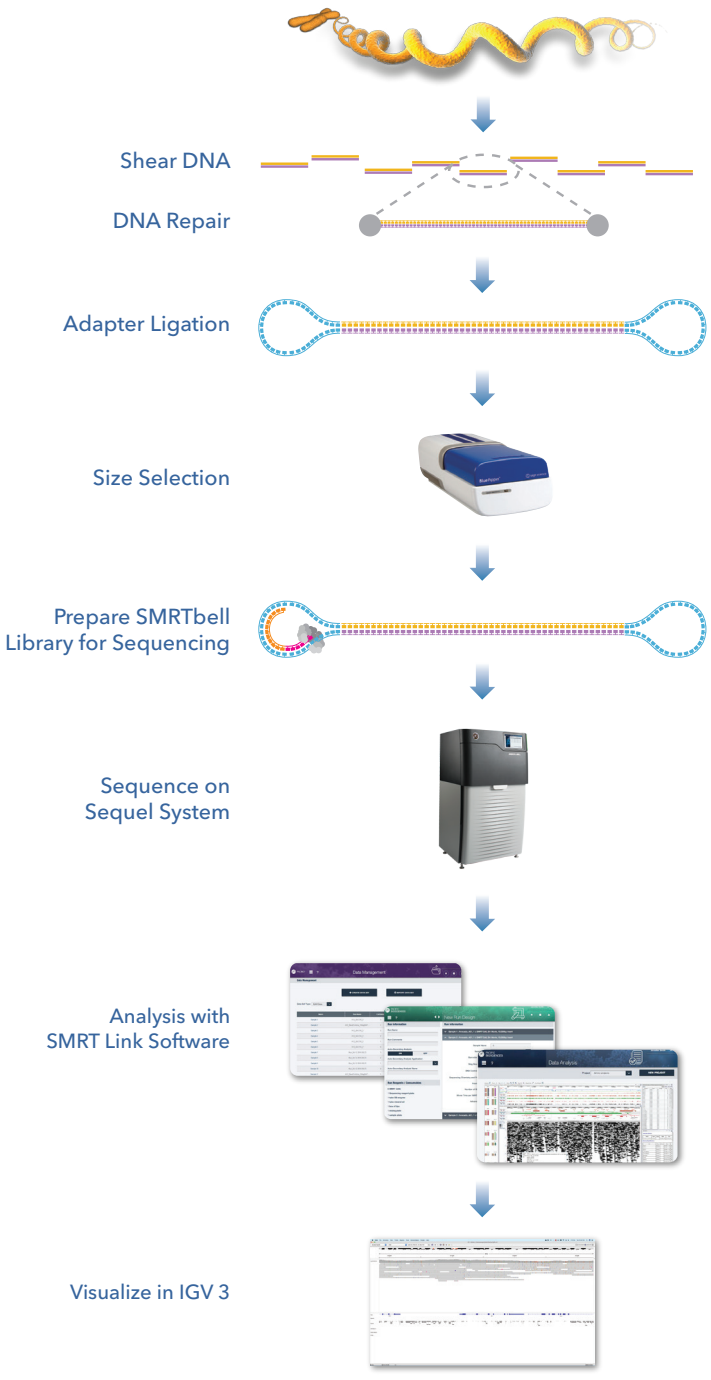
LOW-COVERAGE, LONG-READ WHOLE GENOME SEQUENCING FOR STRUCTURAL VARIATION
BEST PRACTICES

HUMAN BIOMEDICAL RESEARCH

With the Sequel® System, you can affordably and sensitively characterize structural variation (SV) of all types ranging in size from tens to thousands of basepairs. Low-coverage, long-read whole genome sequencing (WGS) data provides rapid discovery of common SVs for population genetics studies and resolves rare SVs unique to an individual, with a very low false-discovery rate.

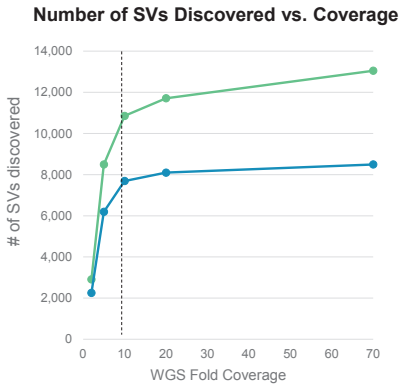
FROM gDNA TO MOST COMPREHENSIVE SV CALL SET



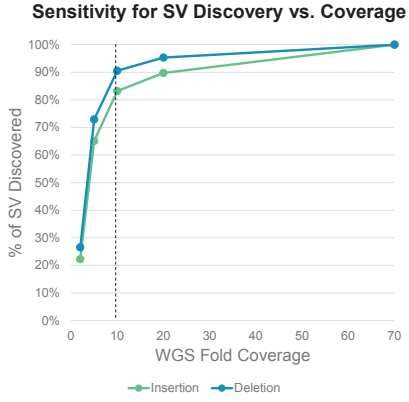
SAMPLE PREPARATION RECOMMENDATIONS

- Use recommended high-quality *unamplified* genomic DNA input (>5 µg)
- Prepare 20 kb long-insert SMRTbell™ library[1]
- Enrich for longest inserts with size selection[1]
- Automated library-prep solutions available[2]
- Sequence to desired coverage based on study needs:
  - 5 to 10-fold: population genetics studies - sensitivity limited per individual, but high for variants shared in the population
  - 10-fold: rare undiagnosed disease studies- sensitivity high per individual allowing discovery of pathogenic SVs
  - >20-fold: *de novo* SV detection in trios - nearly saturated sensitivity
  - ~5 Gb per Sequel SMRT Cell 1M

SV DISCOVERY POWER AT VARIOUS COVERAGE LEVELS



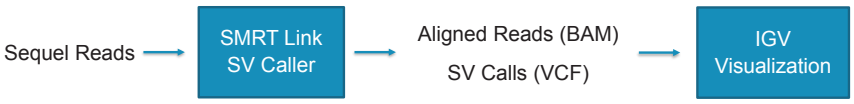
A diploid human (HG00733) was sequenced to 70-fold coverage on the Sequel System. The reads were randomly sampled to various coverage levels, and the SV calls at each coverage were evaluated against the calls at full 70-fold coverage.



Sensitivity increases sharply with coverage until about 10-fold, where it begins to level off. At 10-fold coverage, 10,854 insertions and 7,692 deletions are called (83% and 90.5% sensitivityrespectively)[3].

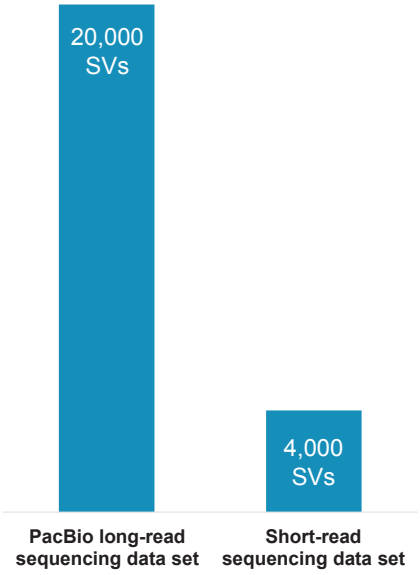
DATA ANALYSIS SOLUTIONS WITH SMRT® ANALYSIS, IGV 3 AND DEVNET

- Detect the broadest range of SV types including deletions, insertions, duplications, inversions, tandem repeats and translocations[4]
- Discover more SVs with the highest sensitivity - up to 20,000 SVs per genome[3,5]
- Resolve breakpoints to sequence level[6]
- Limit costly validation efforts with a low false-discovery rate of 5-10%[7]
- Output data in standard file formats – BAM and VCF – for seamless integration with downstream analysis tools
- Confirm SV calls visually with IGV 3[8] and GenomeRibbon[9]



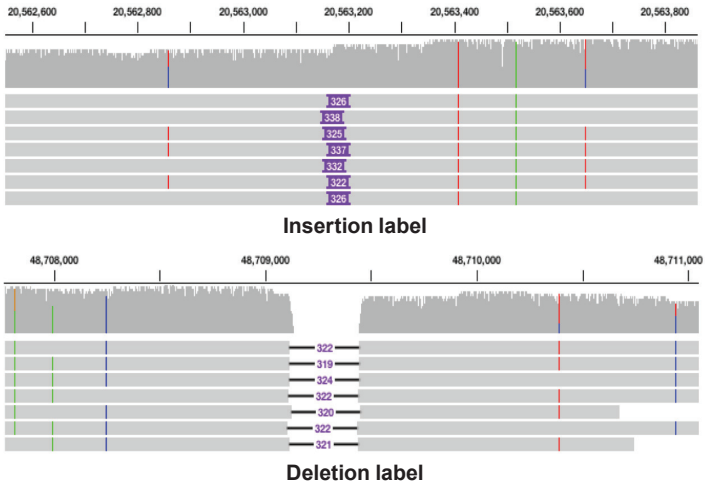
SMRT Link and IGV provide a complete workflow to call and visualize structural variants in PacBio reads.

LONG-READ SMRT SEQUENCING PROVIDES HIGHER SENSITIVITY FOR SV DISCOVERY



A comparison of SV sensitivity across short-read sequencing and PacBio long-read sequencing for SV discovery in published human genomes[7,10,11]. PacBio long-read sequencing uncovers many structural variants missed by short-read methods, providing up to five-fold higher sensitivity for true SV discovery[5,12].

VISUALLY VALIDATE SV CALLS BY EXPLORING READ EVIDENCE IN IGV 3



Visualization of long-read sequencing evidence in support of an insertion and deletion structural variant in IGV 3. Insertions are indicated with a purple box with the width of the box proportional to the size of the insertion, and the basepair size is written on the box. Deletions are indicated by a black line. The basepair size of the deletion is written on a white box at the center of the line[6].

KEY REFERENCES

1. Procedure and Checklist – 20 kb Template Preparation Using BluePippin™ Size-Selection System.
2. "SMRT Compatible Workflow Products, Template-Prep Automation." *PacBio Products and Services.* PacBio, 10 May 2017.
3. "The Human Genome Structural Variation Consortium". The 1000 Genomes Project. IGSR: The International Genome Sample Resource. PacBio Single molecule real time (SMRT) Sequel System high-coverage dataset for HG00733. March 2017.
4. Schatz, M. (2017, Feb) Personalized phased diploid genomes of the EN-TEx samples. *Presentation at Advances in Genome Biology and Technology*, Hollywood, FL.
5. Wenger, A. "Identifying structural variants in NA12878 from low-fold coverage sequencing on the PacBio Sequel System." Web blog post. *PacBio Blog*. PacBio, 19 Oct 2016, Web. 10 May 2017.
6. Wenger, A. "IGV 3 Improves Support for PacBio Long Reads." Web blog post. *PacBio Blog*. PacBio, 29 Mar 2017, Web. 10 May 2017.
7. Huddleston, J. et al. (2016) Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*. (5), 677-685.
8. Robinson JT, et al. (2011). Integrative genomics viewer. *Nature Biotechnology*. 29(1), 24-26.
9. Nattestad M. et al., (2016) Ribbon: Visualizing complex genome alignments and structural variation. *bioRxiv*. https://doi.org/10.1101/082123.
10. Seo, J. et al. (2016) *De novo* assembly and phasing of a Korean human genome. *Nature*. 538, 243-247.
11. Shi, L. et al. (2016) Long-read sequencing and de novo assembly of a Chinese genome. *Nature Communications*. 7, 12065.
12. Sudmant, P. et al. (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*. 526, 75–81.

PN: BP104-053117