Structural Variant Detection with Low-Coverage PacBio Sequencing

Sarah B. Kingan and Aaron M. Wenger
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

Structural variants (genomic differences ≥50 base pairs) contribute to the evolution of organisms traits and human disease.

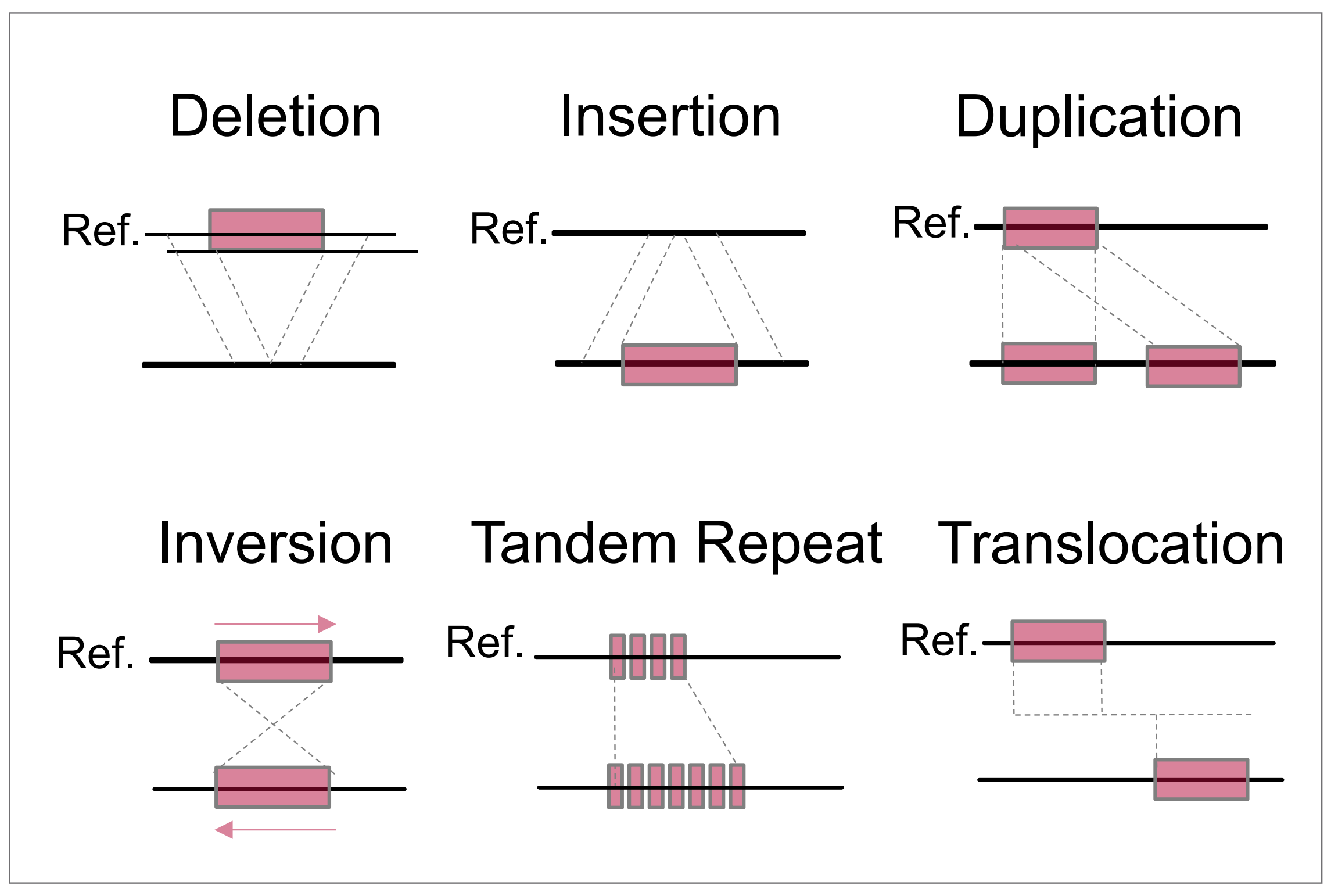
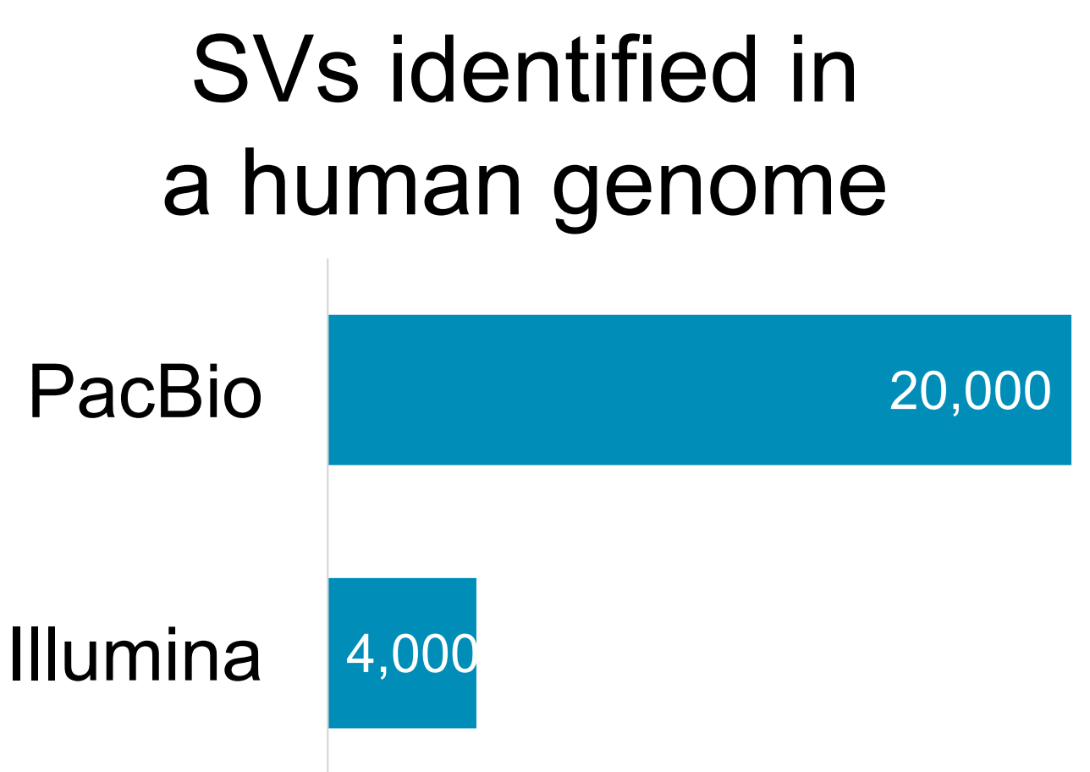


Figure 1. Common types of structural variation.

Background

Most structural variants (SVs) are too small to detect with array comparative genomic hybridization but too large to reliably discover with short-read DNA sequencing. Recent studies in human genomes show that PacBio SMRT Sequencing sensitively detects structural variants[1].



"Fivefold increase in sensitivity [when compared to short-read sequence data]... from the improved mappability of long-read sequence data to repeat-rich regions (especially STRs and variable number tandem repeats), GC-rich DNA, and low-complexity DNA."[1]

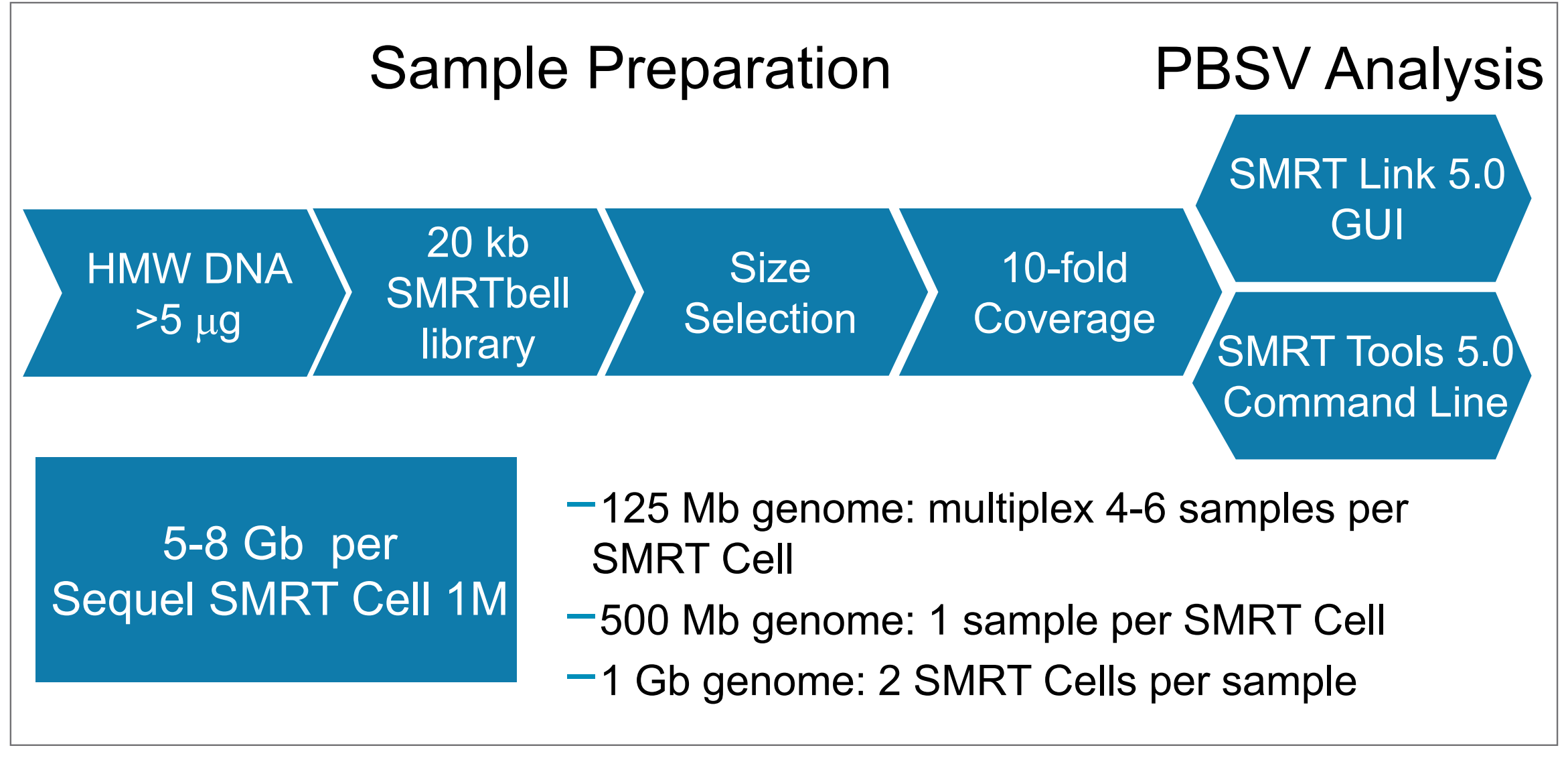
Figure 2. PacBio long reads have 5-fold increased sensitivity for structural variants compared to Illumina short reads.

While *de novo* assembly is the ideal method to identify variants in a genome, it requires high depth of coverage. Structural variant discovery using 10-fold coverage in humans analyzed with `pbsv` shows similar sensitivity for detecting variants.

Dataset	AK1[2]	CHM[3]	NA12878[4]
Analysis Method	*de novo* assembly	*de novo* assembly	pbsv
Fold Coverage	101-fold	41-fold	10-fold
Deletions (>50 bp)	7,358	6,111	8,209
Insertions (>50 bp)	10,077	9,638	11,350

Table 1. Structural variants in PacBio *de novo* human genome assemblies and low-coverage structural variants analysis.

Sample Preparation



Sample Preparation

PBSV Analysis

HMW DNA >5 µg → 20 kb SMRTbell library → Size Selection → 10-fold Coverage → SMRT Link 5.0 GUI / SMRT Tools 5.0 Command Line

5-8 Gb per Sequel SMRT Cell 1M

- 125 Mb genome: multiplex 4-6 samples per SMRT Cell
- 500 Mb genome: 1 sample per SMRT Cell
- 1 Gb genome: 2 SMRT Cells per sample

Figure 7. Recommendations for sample preparation and number of Sequel SMRT Cells for different genome sizes.

Analysis: Map Reads, Chain Alignments, Call Variants

The analysis workflow to identify structural variants from low-coverage PacBio sequencing is: 1) map reads to the reference, 2) chain alignments, and 3) cluster indels to call variants.

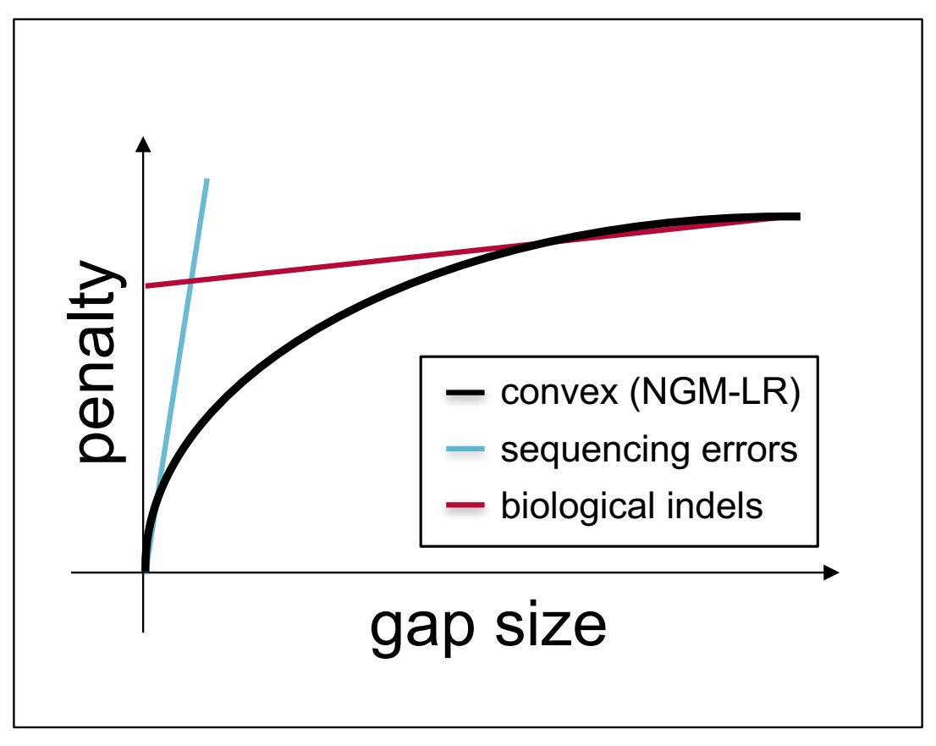


Figure 3. NGM-LR is a read mapper designed for PacBio reads[5]. NGM-LR uses a convex gap penalty to model two sources of alignment gaps: biological indels and sequencing errors.

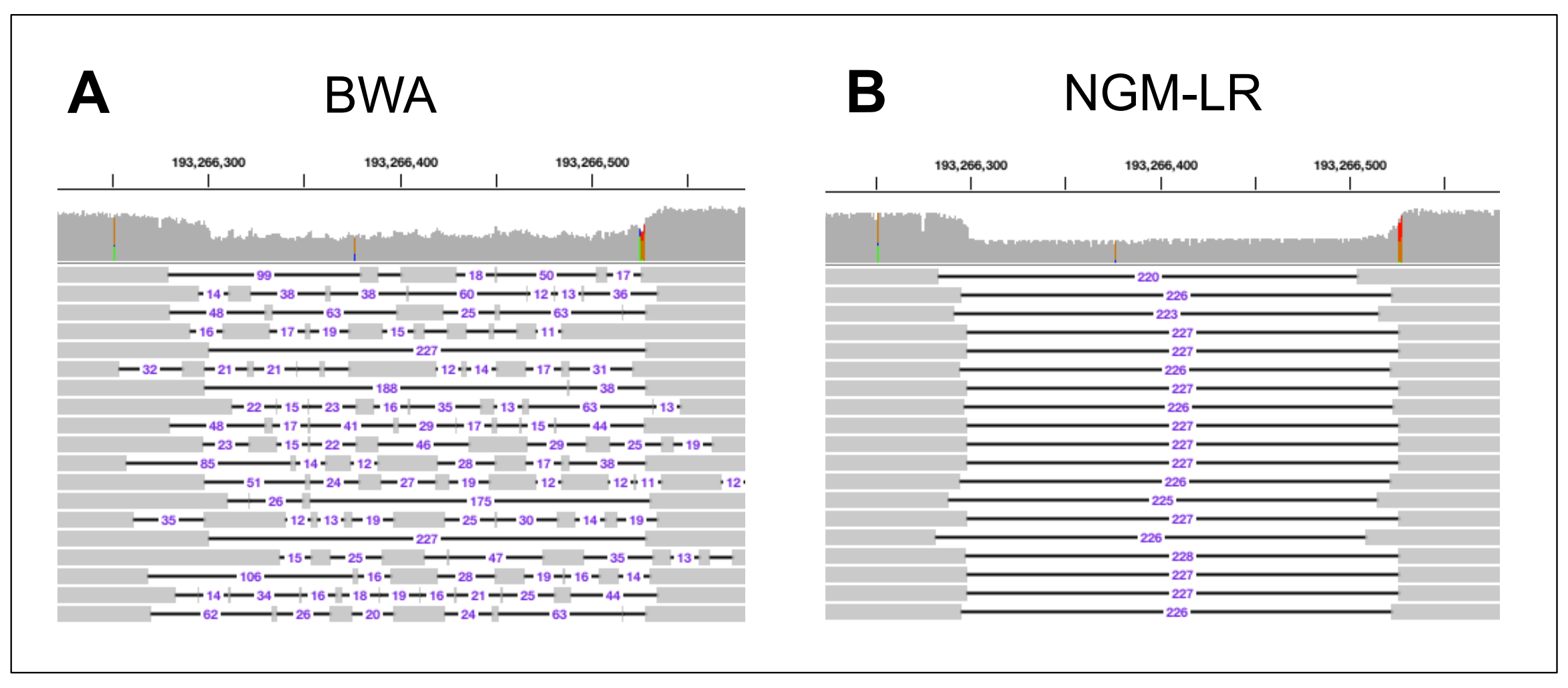


Figure 4. Comparison of read mapping spanning a deletion. (A) BWA, which uses a standard affine gap penalty, produces fragmented alignments at a deletion variant. (B) NGM-LR aligns the same PacBio reads with sharp boundaries at the deletion.

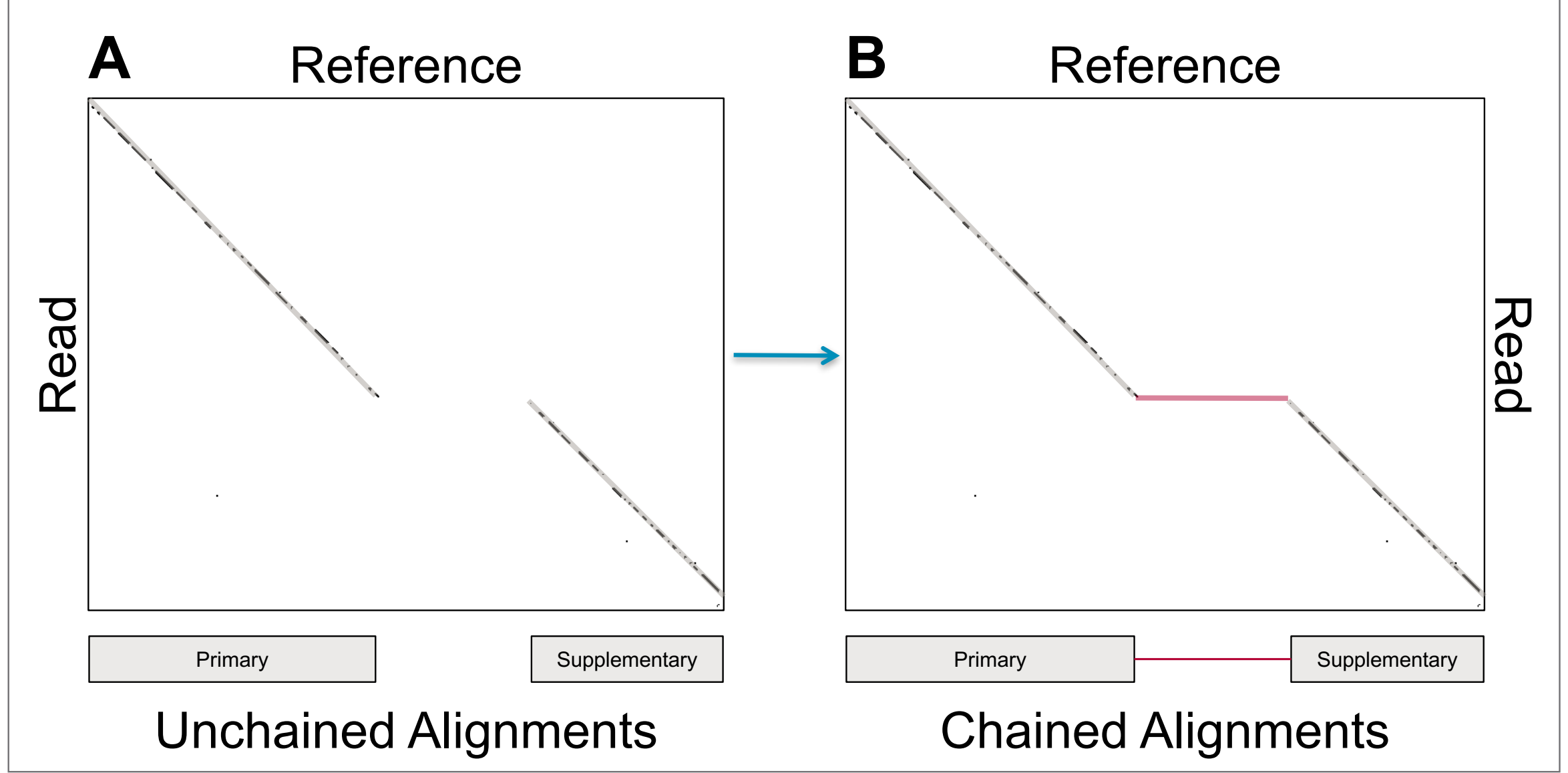


Figure 5. Chaining split alignments. Large gaps split NGM-LR alignments into primary and supplementary segments. Chaining connects collinear segments across large gaps. (A) A large deletion splits alignments of a read into two disjoint segments. (B) Chained alignments directly include a biological deletion, which simplifies visualization and variant calling.

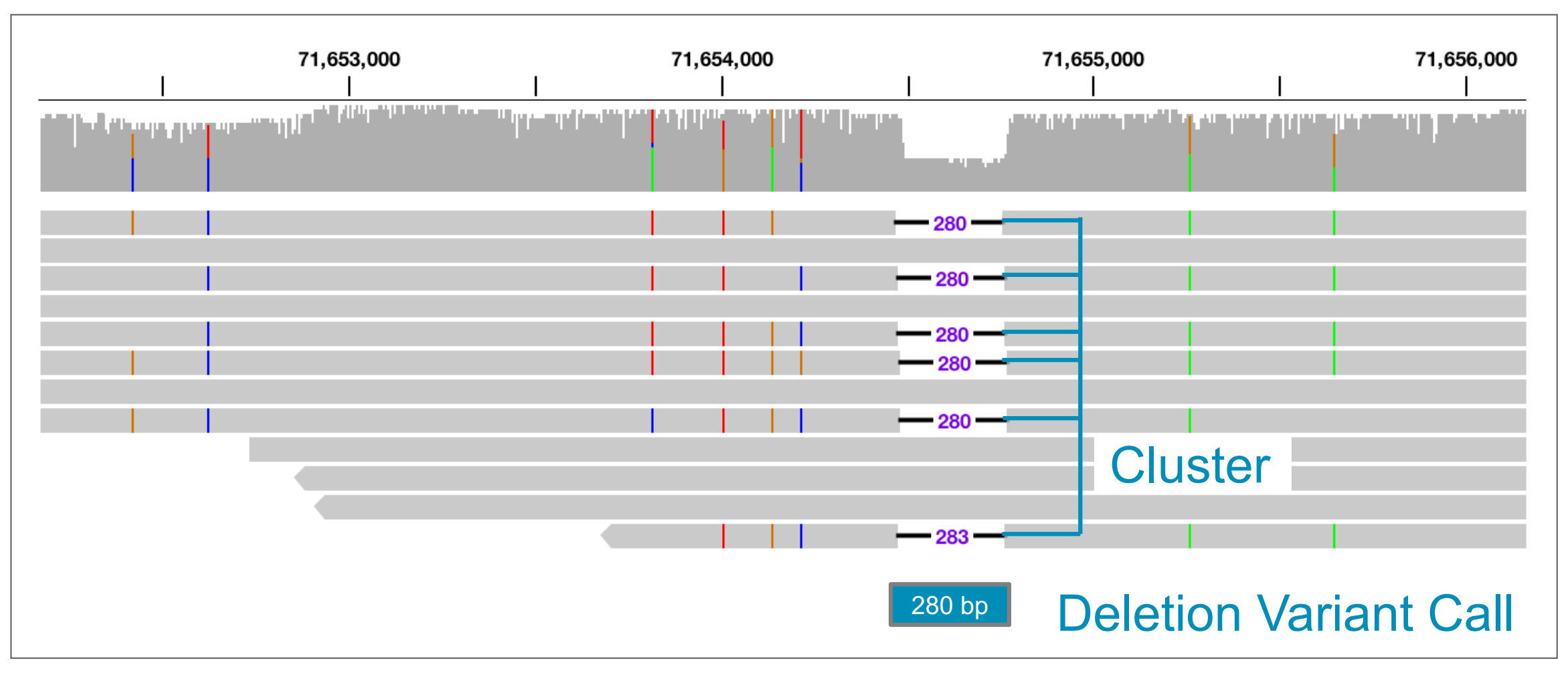


Figure 6. Variant calling. To call structural variants from low-coverage sequencing, identify large deletion or insertion events in chained alignments, cluster nearby events that have similar length and sequence, and summarize into a call. Visualized in IGV v3.0 beta.

Application of PBSV to *Drosophila*

PacBio RS II data[6] for *D. yakuba* mapped to v1.05 of Dyak genome with `pbsv align`. Variants called with `pbsv call` to generate BED and VCF formatted outputs.

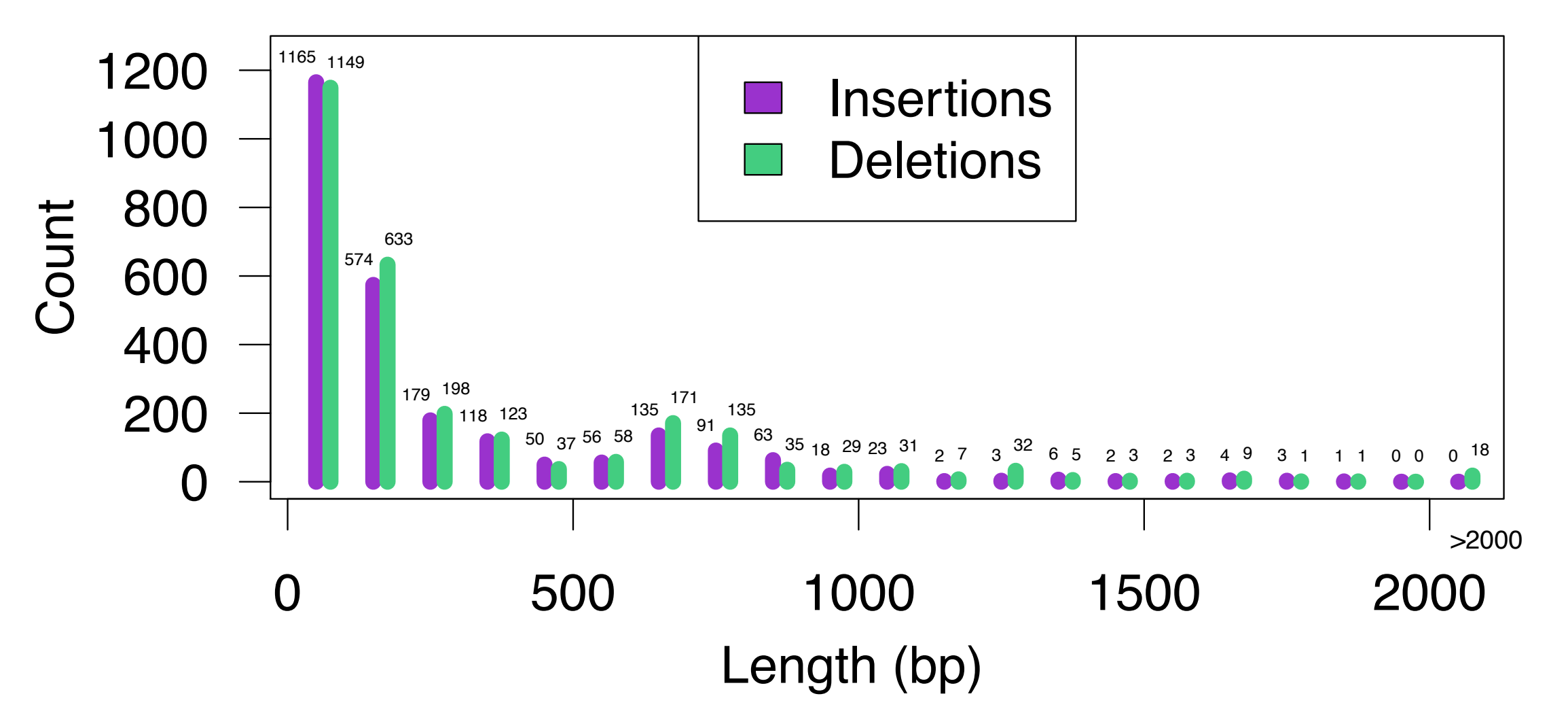


Figure 8. Length Histograms of Insertions and Deletions in female of Cameroon strain CY21B3 (SRR1200825). Repeat Masker analysis identifies abundant 600-800bp SVs as Helitron repeats DNARep1_Dyak and DNARep1_DM.

More X Insertions, Fewer Coding SVs

	INSERTIONS		DELETIONS	
	Observed	Expected	Observed	Expected
X	***434***	***369***	393	374
Autosomes	***1572***	***1638***	1638	1657
Coding	***81***	***374***	***52***	***380***
Noncoding	***1949***	***1656***	***2009***	***1681***

Table 2. Counts of SVs on the X *vs* Autosomes and Coding *vs* Noncoding sequence. pbsv BED file and Dyak v1.05 GFF file analyzed with `bedtools intersect`[7]. 2X2 table values in ***bold italic***: P<0.05, X^2 test with Bonferonni correction. Expectations based on length of major chromosome arms or CDS versus non-CDS region lengths.

Conclusion

- `pbsv align` uses the NGM-LR read mapper and alignment chaining to accurately map PacBio reads to a reference.
- `pbsv call` produces standard VCF and BED formats for custom downstream analysis.
- With yield of Sequel SMRT Cells 1M of 5-8 Gb, it is cost effective to screen populations for SVs using the `pbsv` workflow.
- Applications to humans and *Drosophila* identify thousands of insertions that cannot be detected with short-read technologies.

References

1. Huddleston J, et al. (2016). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research.* doi:10.1101/gr.214007.116.
2. Seo JS, et al. (2016). *De novo* assembly and phasing of a Korean human genome. *Nature.* 538(7624), 243-247.
3. Chaisson MJ, et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*. 517(7536), 608-611.
4. Wenger AM, et al. (2016) http://www.pacb.com/blog/identifying-structural-variants-na12878-low-fold-coverage-sequencing-pacbio-sequel-system/
5. Rescheneder P, et al. (2017). https://github.com/philres/ngmlr
6. Rogers RL, et al. (2014). Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans.* *Mol Biol Evol.* 2014 31(7):1750-66.
7. Quinlan AR and Hall IM. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6), 841-842.

Acknowledgements

The authors would like to thank Yuan Li, Luke Hickey, and Paul Peluso at PacBio.

Conflict of Interest Statement

S.B.K. and A.M.W. are employees and shareholders of Pacific Biosciences, a company commercializing DNA sequencing technologies.