Novel single-cell whole genome amplification long-read workflow highlights complex variation mosaicism

Joe Dahl[1], **Viren Amin**[1], Katie Kennedy[1], Don Trapolsi[1], Laura Williams[1], Durga Arvapalli[1], Isai Salas-Gonzalez[1], Jamie Remington[1], Khadija Clark[1], Victor Weigman[1]
[1]BioSkryb Genomics, Durham, NC, United States

Single-cell long-read sequencing with PTA delivers ~4 kb amplicons and 75-92% genome coverage per cell with high uniformity, >90% phasing, and >70% SV precision - enabling analysis of hard/repetitive regions, CNAs, and repeat expansions.

Background & Methods

- **Motivation.** Long reads resolve phasing, repeats, and complex SVs; but have underperformed in single-cell WGA due to significant bias, chimeras, and allele dropout.
- **Prior work.** Prior single-cell long-read WGA (e.g. SMOOTH-seq[1]/dMDA[2]) methods showed feasibility but with uneven coverage and artifact-driven SV calls.
- **Goal.** To create a WGA compatible method with standard long-read pipelines that keeps fragments long, chimeras low, and coverage uniform.

Primary Template-directed Amplification (PTA) formulation was tuned to produce multi-kb fragments (mean ~4 kb) from single cells for long-read sequencing.

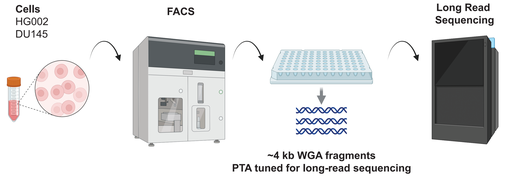

Figure 1. Workflow Overview. Single cells are isolated and FACS-sorted, followed by whole genome amplification (WGA) using Primary Template-directed Amplification (PTA) tuned for long-read sequencing. PTA amplicons are converted to long-read libraries and sequenced with PacBio HiFi for downstream analysis.

Study Design and Analysis

- **Samples:** 8 single cells: HG002 (n=4; GIAB reference, benchmarking) and DU145 (n=4; human prostate cancer cell line, AR-negative). Bulk HG002 PacBio data were obtained from the PacBio public dataset for comparison[3].
- **Sequencing:** PacBio Revio HiFi
- **Pipeline:** Alignment with pbmm2; SNV/indel calling with DeepVariant (PacBio model); phasing with HiPhase; SV calling with Sniffles2; CNV profiling with Ginkgo using custom PacBio reference build; STR analysis with TRGT. Variant benchmarking used TruVari and vcfeval.

References

1. Fan, X., Yang, C., Li, W. et al. SMOOTH-seq: single-cell genome sequencing of human cells on a third-generation sequencing platform. Genome Biol 22, 195 (2021).
2. Hård, J., Mold, J.E., Eisfeldt, J. et al. Long-read whole-genome analysis of human single cells. Nat Commun 14, 5164 (2023).
3. Pacific Biosciences. (2024). GIAB trio (HG002, HG003, HG004) — Revio HiFi WGS (SPRQ), 2024Q4 release [Public dataset]. PacBio Public Datasets. Retrieved March 15, 2025.

Results: Coverage, Uniformity, SNVs & Phasing

PTA tuned for long-read resulted in ~18–24M total reads per single-cell (HiFi reads ~6-8M) with ~4 kb read length, low chimera (~7–9%) and duplication (~4–7%) rates. Both DU145 and HG002 showed broad genome coverage. DU145 converted depth to breadth efficiently (≥1×: 91–92%; ≥5×: 84–86%), while HG002 showed wider variability (≥1×: 73–88%; ≥ 5×: 54–80%) from poor library conversion in a subset of cells. This chemistry outperforms prior WGA single-cell long-read chemistries narrowing the gap to bulk long-read sequencing performance.

Table 1. Summary of the sequencing and coverage metrics.

	Sample ID	Total Reads	Percent Chimeras	Percent Duplication	Mean Read Length (Kb)	≥ 1X	≥ 5X	≥ 30X	Median Coverage	Mean Coverage
PacBio Bulk data from long-read sequencing[3]	HG002 Bulk	5,997,240	2%	1%	16.6	94%	93%	35%	27X	26X
Single-cell data from PTA tuned for long-read sequencing	DU145_SC1	23,623,992	8%	6%	4	92%	86%	38%	23X	26X
	DU145_SC2	25,884,637	9%	6%	4	91%	86%	40%	25X	28X
	DU145_SC3	19,737,834	7%	8%	4.2	91%	84%	28%	19X	22X
	DU145_SC4	22,872,506	7%	5%	3.9	92%	86%	36%	23X	25X
	HG002_SC1	24,532,657	8%	4%	3.9	88%	80%	39%	23X	27X
	HG002_SC2	29,000,520	9%	7%	3.9	68%	55%	38%	11X	31X
	HG002_SC3	17,959,903	9%	10%	4.1	73%	54%	26%	7X	19X
	HG002_SC4	19,873,667	8%	4%	3.9	70%	59%	30%	12X	22X

Table 2. Comparing single-cell long-read methods.

Chemistry	Genome Coverage	Mean Read Length (Kb)	SNV Recall	SNV Precision
PacBio Bulk[3]	94%	16.6	98.2%	99.9%
PTA tuned for long-read	75-92%	3.9-4.2	52-80%	99.2%
dMDA[2]	~40%	2.8-3.6	17%	86%
SMOOTH-seq[1]	10.6-41.3%	6	-	-

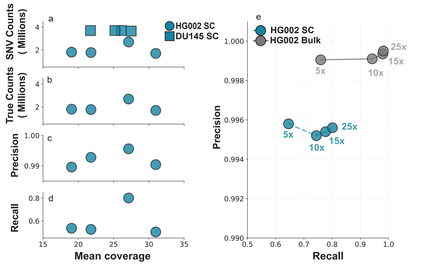

Figure 3. Benchmark of SNV calling in single cells using PTA tuned for long-read. Benchmarked against the HG002 high-confidence set. (a) per-cell total SNV calls, (b) true positive SNV calls, (c) precision, and (d) recall versus mean coverage for DU145 and HG002 single cells. (e) Subsampling analysis shows precision remains consistently high while recall scales with coverage and coverage uniformity.

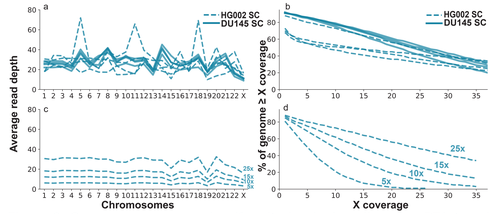

Figure 2. Genome coverage and uniformity in single cells using PTA tuned for long-read. (a) Mean read depth per chromosome for each sample. (b) Genome-wide distribution of per-base coverage. (c–d) Same metrics after subsampling from 5x-25x. Overall chemistry shows high genome coverage and uniformity across single-cells.

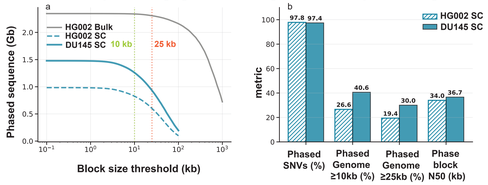

Figure 4. Phasing SNV performance in single cells using PTA tuned for long-read. a) Phasing contiguity curves showing cumulative phased sequence (Gb) vs phase block size threshold, with reference lines at 10 kb (local scale) and 25 kb (gene level). b) Bar plots compare SNV phasing rates, genome coverage in phase blocks at local and gene level, and phase block N50 value. PTA tuned for long-read provides sufficient haplotype continuity to co-phase nearby SNVs/indels, enabling haplotype-aware interpretation at single-cell resolution.

Results: Structural Variations, CNAs, & Repeat Expansions

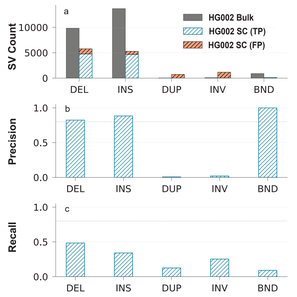

Figure 5. Benchmark of SV calling in single cells using PTA tuned for long-read. (a) SV counts by class (DEL, INS, DUP, INV, BND). Gray bars show the HG002 PacBio bulk HiFi callset; hatched bars show single-cell calls split into true positives (TP) and false positives (FP) after comparison to bulk. (b) Precision per class. (c) Recall per class. Calls are grouped by SV type and assessed against the bulk reference. Overall PTA tuned for long-read maintains high precision across SV classes.

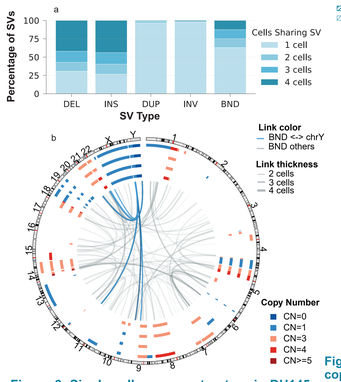

Figure 6. Single-cell genome structure in DU145 resolved using PTA tuned for long-read. (a) SV counts are shown based on recurrence across 4 cells. (b) CNV+BND circos plot per cell: rings denote CN state (1 MB bin-size); links mark translocations, with thickness indicating sharing (2/3/4 cells). Recurrent chrY fusions to chr9/11/22/21 coincide with broad Y loss, distinguishing ancestral vs subclonal events that bulk long-read or short-read cannot resolve.

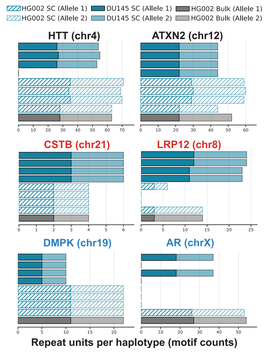

Figure 7. Single-cell STR dosage mirrors copy-number state in DU145. TRGT-estimated repeat units per allele are plotted on the x-axis for DU145 SC and HG002 SC + Bulk (diploid control). Each horizontal bar represents one cell or bulk; colored segments denote repeat units per allele. In DU145, STR dosage is concordant with the local copy-number state measured in the same cell. By coupling per-cell CNA/SV with STR measurements, single-cell long-read can provide copy-number-aware STR estimates per cell with breakpoint-level linkage that bulk long-read cannot capture.

Conclusions

- PTA tuned for long-read sequencing in single cells delivers high genome coverage and uniformity with low chimera/duplication rates.
- Phasing resolves >90% of heterozygous variants, enabling co-phasing of nearby SNVs/indels and haplotype-aware interpretation of SV and CNA breakpoints.
- Overall, the method narrows the gap to bulk long-read performance while retaining true single-cell resolution, enabling analyses of hard/repetitive regions and complex rearrangements.
- Future work will focus on higher throughput, improved uniformity and sensitivity, and joint haplotype-aware SV/indel phasing across cells.

Acknowledgments

We thank Eremid Genomic Services (BioSkryb Certified Service Provider) for long-read library preparation and sequencing support and acknowledge the whole BioSkryb Genomics' team for their continuous effort and work. If you are interested in learning more or would like to explore data further, scan the QR code or contact info@bioskryb.com