BaseJumper® genome analysis workflows with custom DeepVariant showcase best performance in variant detection across classifications, chemistries and sample types

Isai Salas-González[1], Viren Amin[1], Srisanth Balam[1], Pi-Chuan Chang[2], Andrew Carroll[2], Victor Weigman[1]
[1]BioSkryb Genomics, Durham, NC, USA, [2]Google Inc., Mountain View, CA, USA

BioSkryb GENOMICS

ASHG ANNUAL MEETING 2025 BOSTON, MA • OCTOBER 14-18

DeepVariant's custom PTA model with SHF-QC workflow lowers false positives in single-cell WGS to bulk-tissue levels, preserves >85% detection sensitivity, enables unmatched somatic variant calling without bulk references, processes hundreds of cells per day, and enhances biological insight via variant prioritization.

Background & Methods

Motivation: Accurate variant identification is crucial in single-cell genomics to unravel disease evolution. Yet, biases from whole-genome amplification introduce noise and hinder interpretation.

Prior work: Existing variant refinement methods are often computationally inefficient and rely on matched bulk normals.

Goal: Develop a scalable workflow that removes false positives from BioSkryb single-cell variant calls, preserves variant detection sensitivity, and improves data interpretability.

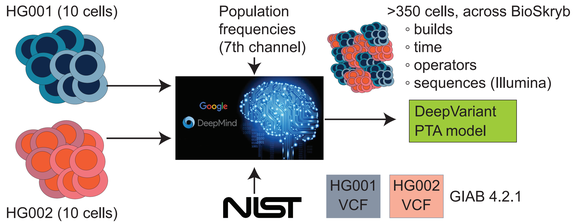

Figure 1. Building the DeepVariant Primary Template Amplification (PTA) model.
The model was trained with Google's DeepMind on GIAB reference cell lines (HG001, HG002) processed with ResolveDNA[RM] and ResolveOME™ chemistries and sequenced with Illumina sequencing technology, including >350 cells with variability across builds, times, operators and sequencing depth.

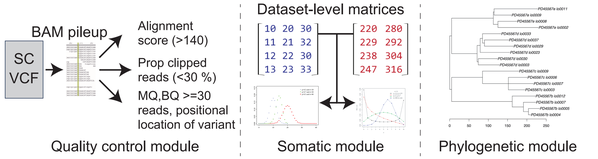

Figure 2. Building the Somatic Heuristic Filtering (SHF-QC) workflow.

- The QC module processes single cells, generating BAM pileups and filtering variants by alignment quality, read clipping, and variant position across HQ supporting reads.
- The somatic module aggregates read data across cells based on read support and total depth in position and then computes binomial and beta-binomial distributions to remove germline and low-input artifacts.
- The phylogenetic module performs phylogenetic reconstruction using Sequoia and places variants across trees branches using TreeMut.

References

1. Chen, Nae-Chyun et al. Improving variant calling using population data and deep learning. BMC Bioinformatics, 24, 197 (2023).
2. Coorens, T.H.H et al. The somatic mutation landscape of normal gastric epithelium. Nature, 640, 418-426 (2025).

Results: DeepVariant + SHF-QC benchmarking

Three HG002 single cells were analyzed with ResolveDNA[RM] v2 and ResolveOME™ v2 alongside NIST WGS bulk. Variants were called at ~15× and ~20× using DNAScope and DeepVariant (PTA/Illumina models). HF-QC and SHF-QC were applied with BioSkryb optimized parameters; performance of variant calling was assessed with VCFEval vs GIAB v4.2.1 sets.

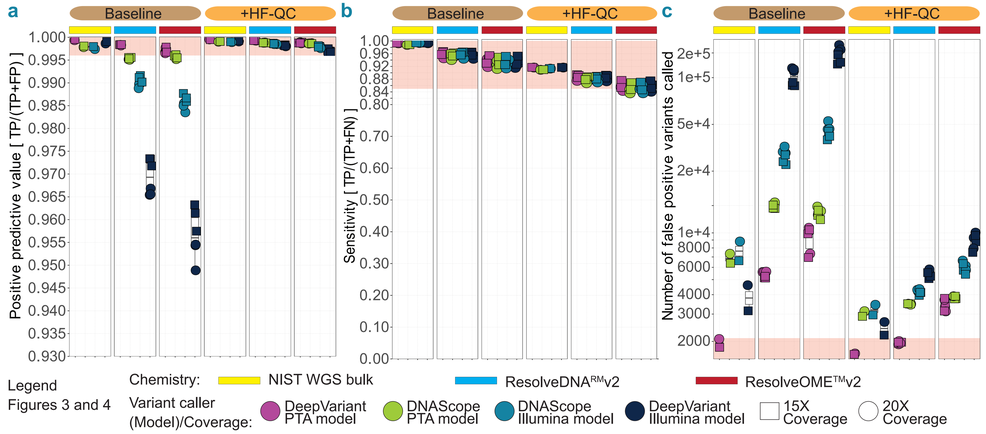

Figure 3. BioSkryb's DeepVariant+HF-QC workflow reduces false positives in single-cell data to bulk levels while maintaining >85% variant detection sensitivity.
a) PPV versus GIAB set for baseline variant models (DeepVariant PTA/Illumina, DNAScope PTA/Illumina) and baseline models +HF-QC; the +HF-QC pipeline markedly improves PPV irrespective of variant calling method (orange shading).
b) SNV detectionsSensitivity, of callers+models described in (a); +HF-QC retains >85% sensitivity (~3 million variants).
c) False-positive counts of callers+models; DeepVariant (PTA)+HF-QC reduces false positive calls to bulk levels (orange shading).

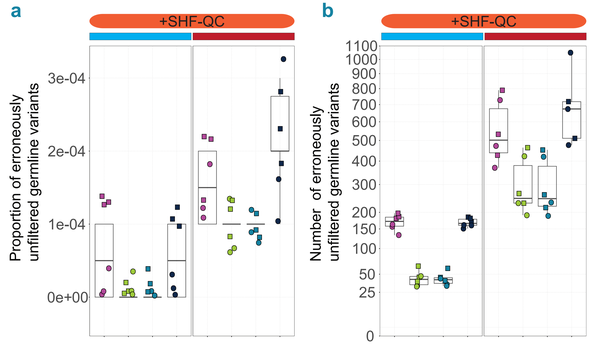

Figure 4. SHF-QC workflow effectively separates germline from somatic variants.
Proportion (a) and number (b) of the HG002 ground truth germline variants erroneously unfiltered by the SHF-QC's somatic module.

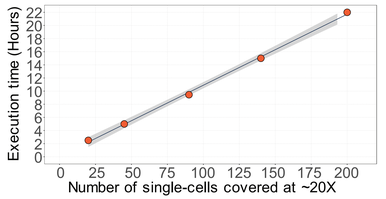

Figure 5. SHF-QC workflow is computationally scalable and cost effective.

Results: DeepVariant + SHF-QC deconvolutes cancer heterogeneity

Utilizing the DeepVariant + SHF-QC workflow, we called somatic mutations and inferred lineage tracing over 45 single cells coming from primary cancer data.

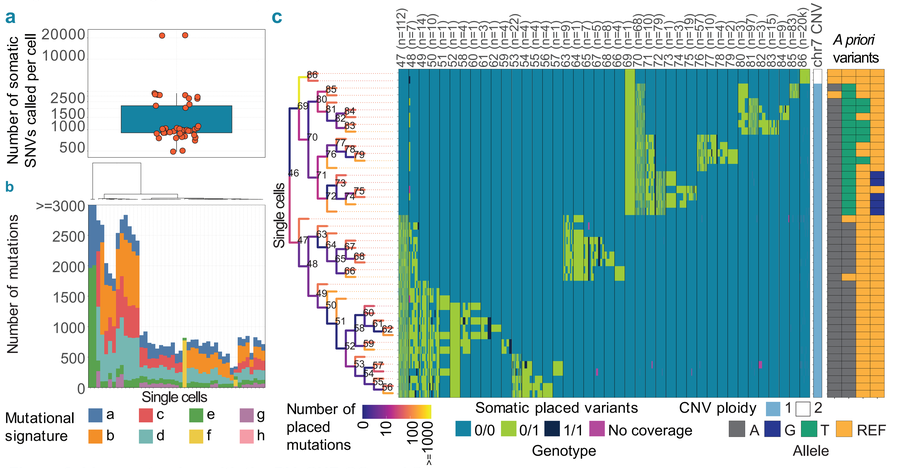

Figure 6. Lineage tracing with the DV+SHF-QC workflow.
a) Somatic SNVs detected per cell (45 single cells, dots).
b) Number of somatic mutations assigned to known mutational signatures (colors across cells).
c) Left: Phylogenetic tree from high-quality somatic calls; branch colors indicate marker mutations, node ids are displayed in black. Middle: Heatmap of marker mutations separating multicell branches. Right: CNV profiles (chr7, 1 Mb bins) show loss of heterozygosity (LOH) matches distinct branching in node 86 cells versus the remaining cells. Final heatmap shows *a priori* known marker variants consistent with the *de novo* inferred phylogeny topology.

Conclusions

- **Custom model:** We trained a DeepVariant model with Google DeepMind optimized for primary template amplification (PTA) Illumina data, outperforming prior variant callers.
- **Scalable workflow:** We built a Nextflow-based workflow enabling stringent variant QC, somatic calling without bulk normals, and lineage tracing across hundreds of high-pass WGS cells in one day.
- **Improved performance:** In conjunction, the DeepVariant PTA model + SHF-QC workflows exhibit reduced false positives to bulk-WGS levels while maintaining >85% sensitivity in SNV detection.
- **Availability**: The DeepVariant PTA model workflow is available through ResolveServices[SM]. The SHF+QC workflow can be downloaded via github (https://github.com/BioSkryb/bj-somatic-variantcalling).

Acknowledgements

We are thankful to the whole BioSkryb Genomics' team for their continuous effort and work.
If you are interested in learning more or like to explore data further, scan the QR code or contact