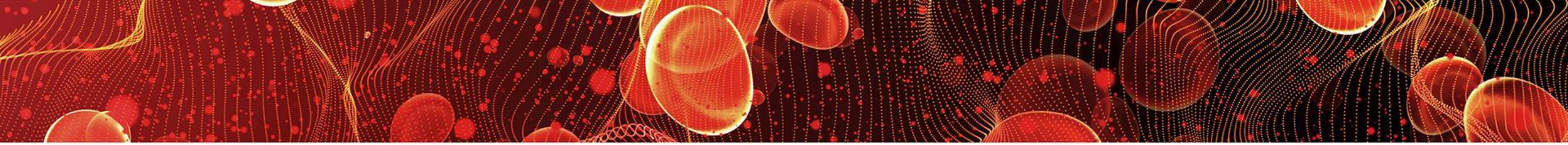


mission bio

Accurate and sensitive subclone identification in scDNAseq datasets

Saurabh Gulati[1], Shu Wang[1], Saurabh Parikh[1], Manimozhi Manivannan[1]

[1]Mission Bio, South San Francisco, CA, USA

Conflict of interest: S.G., S.W., S.P., M.M. are employees and share holders of Mission Bio, Inc.

Single cell workflow for accurate subclone identification

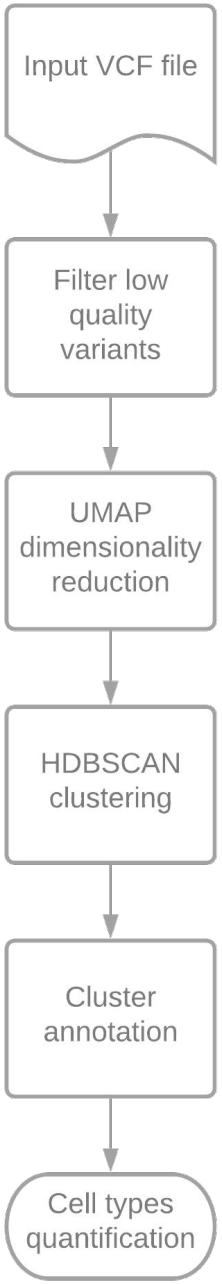
Single-cell sequencing has the potential to provide unique insights on the cellular and genetic composition, drivers, and signatures of cancer at unparalleled sensitivity. A challenge to reach such sensitivities is to have a subclone identification method that can accurately identify small populations of cells which have a distinct genotype. Various chemistry based errors lead to the creation of many small subpopulations of cells which make identification of subclones challenging. To address this issue, we present a cell type identification method that can identify cell populations upto 1% spike-in sensitivity.

Method

The method begins by filtering the VCF file to remove low quality data based on thresholds of allele frequency, read depth, genotyping quality etc. These filters aid in removing a significant amount of noise caused due to chemistry based errors.

The result is a matrix of cells vs variants with the values being either the allele frequencies for each cell/variant combination. This matrix is then reduced to 2 dimensions using UMAP, followed by HDBSCAN clustering.

The observed clusters are annotated by comparing their genotype signature with known genotypes of true clones.

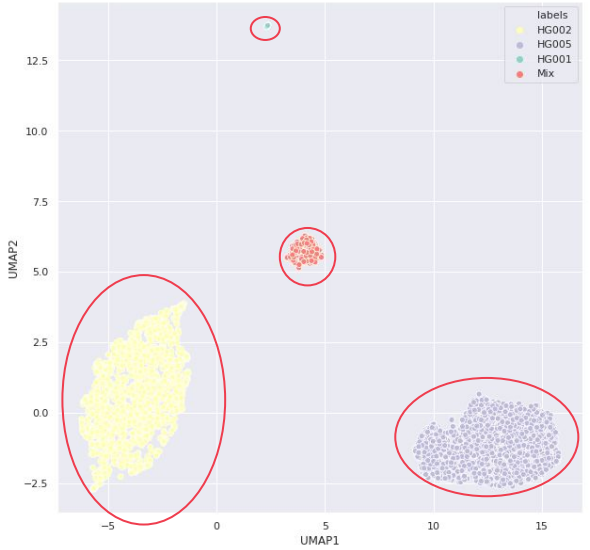


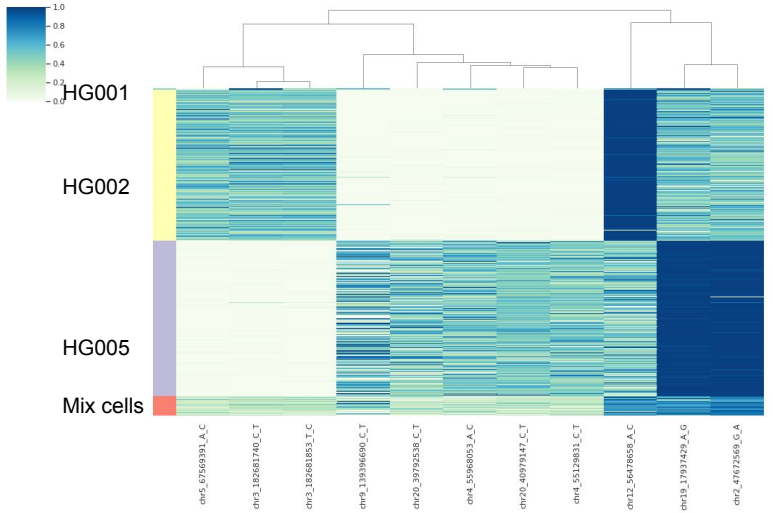
Results

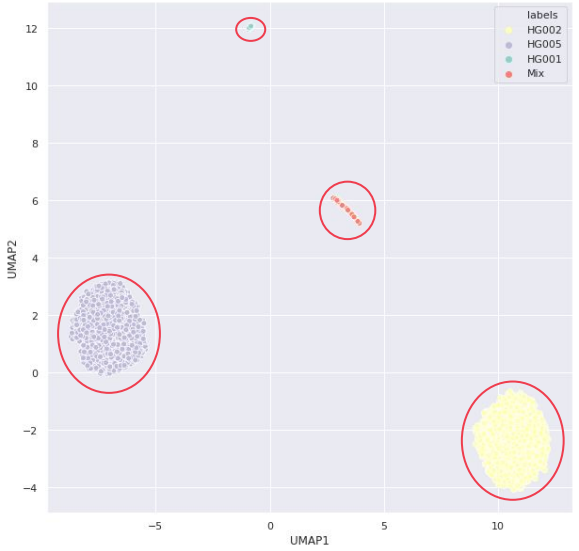
This method was tested using different targeted sequencing panels of sizes 100 to 1000 amplicons each, created for mixtures of 3 GIAB cell lines at 49.5/49.5/1% ratios. We were successfully able to identify all true cell line populations along with mix cells which were caused due to the cell mixing of any 2 cell lines.

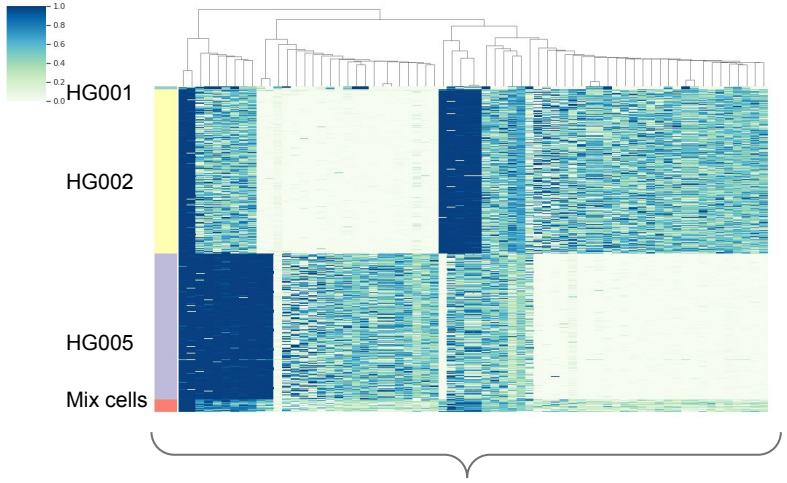
The heatmap shows the variants as columns and cells as rows with the values being the allele frequencies. We can clearly see how the variants differ in the cell lines.

This method was successfully able to identify all cell lines, including the 1% spike in cell line in these runs.









Panel size	Total cells	HG001	HG002	HG005	Mix cells
100plex (up panel)	3204	19 (0.59%)	1473 (45.97%)	1524 (47.57%)	188 (5.87%)
600 plex (bottom panel)	2214	21 (0.95%)	1117 (50.45%)	989 (44.67%)	86 (3.88%)

mission bio