GUMM: A purpose-built computational workflow for single cell genotyping of gene editing experiments

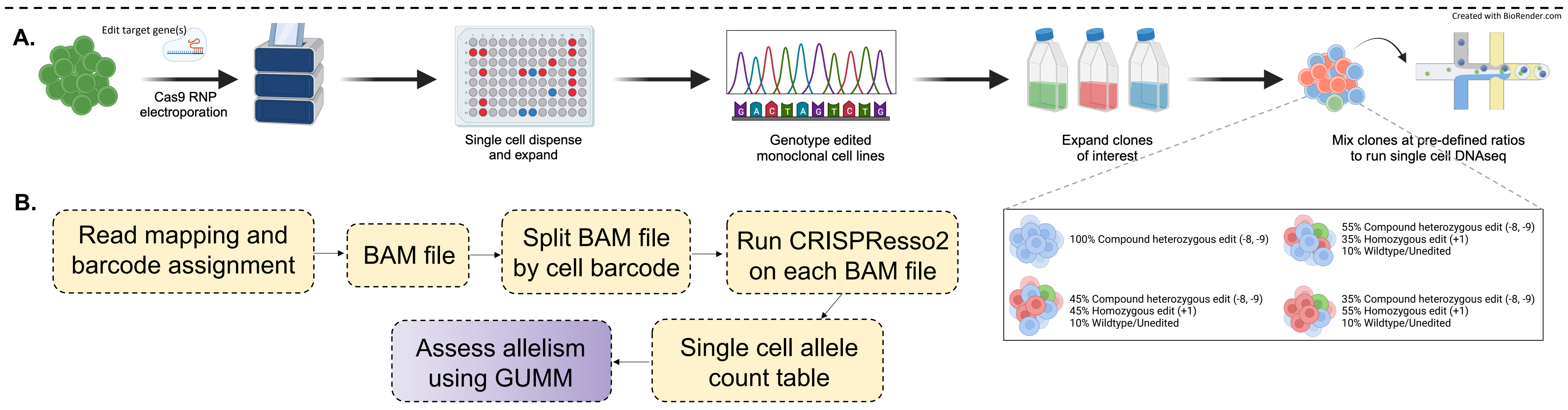
Matthew Ung*, Ruijia Wang*, Gabriella Angelini, Juliana Xavier-Ferrucio, Michelle Lin, Tirtha Chakraborty, **Gary Huanying Ge**

Vor Bio, Cambridge, MA, USA, *Authors contributed equally

Introduction

Engineering cell therapies requires accurate assessment of gene modified allelism because editing patterns can vary across cells and cause phenotypic heterogeneity in a sample. This can delay development of complex cell therapies involving the use of multigenic editing. Recently, droplet-based targeted single cell DNA sequencing (scDNAseq) has been used to genotype select loci across thousands of cells enabling high-throughput assessment of gene editing efficiency. Here, we developed a novel computational workflow called GUMM (Genotyping Using Mixture Models) that systematically infers single cell allelism at select loci from scDNAseq data by fitting a series of Gaussian mixture models (GMMs) to allele read counts generated by CRISPResso2; GUMM is uniquely well-suited for analyzing CRISPR-Cas9 gene editing experiments where cells in the sample are genetically homogenous and differ only at the intended editing site(s). GUMM outputs a probabilistic prediction of cell genotype and addresses technical artifacts including low coverage at the editing site, PCR amplification imbalance, multiplets, and sequencing error. Moreover, we constructed a gene editing "ground truth" scDNAseq atlas to deeply characterize these technical artifacts and leveraged it to optimize GUMM.

Methods


Created with BioRender.com
A.
Edit target gene(s)
Cas9 RNP electroporation
Single cell dispense and expand
Genotype edited monoclonal cell lines
Expand clones of interest
Mix clones at pre-defined ratios to run single cell DNAseq
100% Compound heterozygous edit (-8, -9)
55% Compound heterozygous edit (-8, -9)
35% Homozygous edit (+1)
10% Wildtype/Unedited
45% Compound heterozygous edit (-8, -9)
45% Homozygous edit (+1)
10% Wildtype/Unedited
35% Compound heterozygous edit (-8, -9)
55% Homozygous edit (+1)
10% Wildtype/Unedited
B.

Read mapping and barcode assignment → BAM file → Split BAM file by cell barcode → Run CRISPResso2 on each BAM file → Single cell allele count table → Assess allelism using GUMM

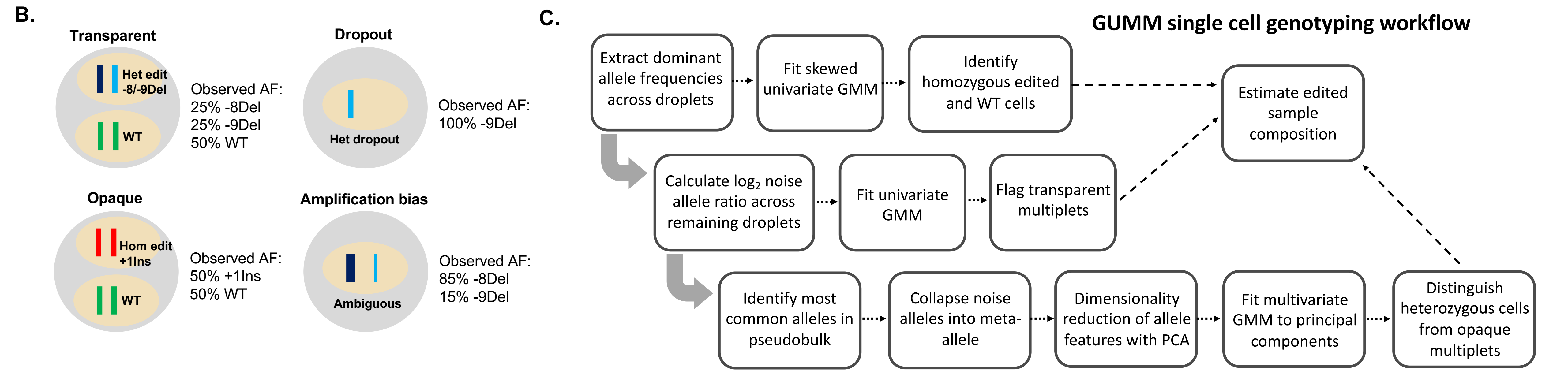
Figure 1) Construction of a "ground truth" gene editing scDNAseq resource and analysis workflow

A. We used CRISPR-Cas9 to create edited HL-60 cells and expanded clones with distinct homozygous and compound heterozygous indel profiles. Clones were mixed at pre-defined ratios to create artificial cocktails that mimic the potential editing diversity of a CRISPR-Cas9 experiment. Three unique cocktails were sequenced to generate an atlas containing single cell readout for more than 20,000 cells. **B.** Overview of computational pipeline used to analyze single cell DNA sequencing (scDNAseq) data from artificial cocktails. Pipeline consists of read mapping and barcode deconvolution, allele quantification with CRISPResso2[1], and artifact-aware genotyping with GUMM (Genotyping Using Mixture Models).

Results

A.

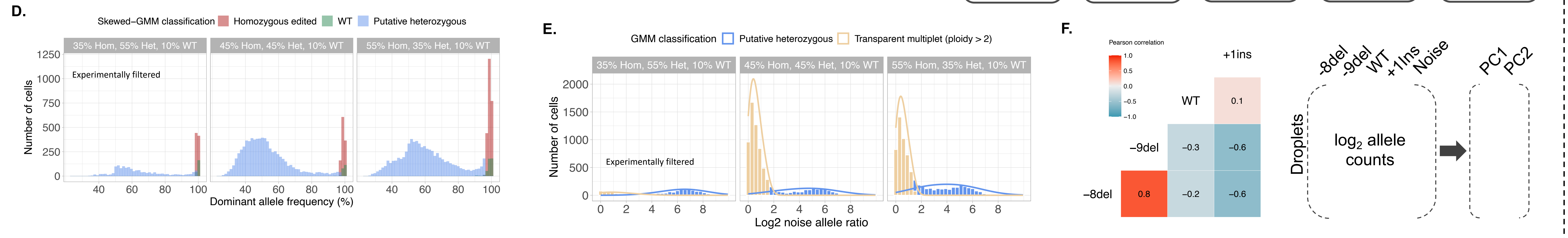
		Target gene alleles			
Cell barcode	Genotype	+1Ins	WT	-8Del	-9Del
AAGCAATCGACTCCAACT-1	WT/Unedited	0	**1059**	0	4
AACAACTGGAGCCTCGCA-1	Homozygous edit	**824**	0	0	0
AACAATGCACGTTAATGG-1	Heterozygous edit	0	0	**551**	**462**
CACATATCAACCTAGCAC-1	Transparent multiplet	0	**205**	**192**	**197**
AGGAAGGTGCACCAAGAT-1	Opaque multiplet	0	**98**	0	**290**


B.
Transparent
Het edit -8/-9Del
WT
Observed AF: 25% -8Del 25% -9Del 50% WT
Dropout
Het dropout
Observed AF: 100% -9Del
Opaque
Hom edit +1Ins
WT
Observed AF: 50% +1Ins 50% WT
Amplification bias
Ambiguous
Observed AF: 85% -8Del 15% -9Del
C.
GUMM single cell genotyping workflow

Extract dominant allele frequencies across droplets → Fit skewed univariate GMM → Identify homozygous edited and WT cells → Estimate edited sample composition

Calculate $\log_2$ noise allele ratio across remaining droplets → Fit univariate GMM → Flag transparent multiplets → Estimate edited sample composition

Identify most common alleles in pseudobulk → Collapse noise alleles into meta-allele → Dimensionality reduction of allele features with PCA → Fit multivariate GMM to principal components → Distinguish heterozygous cells from opaque multiplets → Estimate edited sample composition



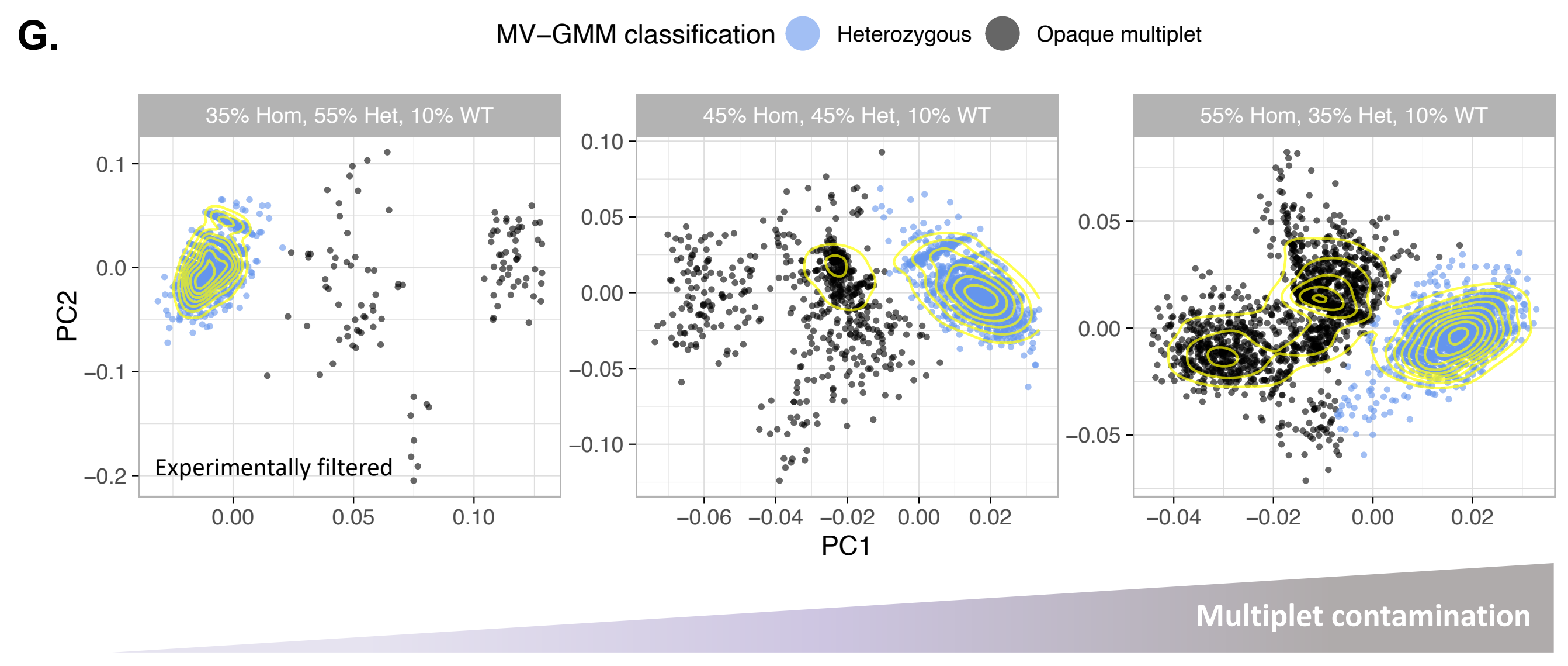


Figure 2) Artifact-aware single cell genotyping of artificial cocktails using GUMM

A. CRISPResso2 output of allele counts for six distinct droplet types observed in "ground truth" data resource. Homozygous droplets are identified by a +1 insertion and heterozygous cells are characterized by a -8/-9 compound deletion **B.** Illustration of four potential artifacts observed in scDNAseq data that can confound allelism assessment in gene editing experiments. Examples of allele frequencies (AF) corresponding to these scenarios are provided **C.** Diagram of the GUMM workflow that applies a series of Gaussian mixture models (GMM) to identify homozygous cells, true compound heterozygous cells (Het edit), transparent multiplets (triploid), and opaque multiplets (diploid). The workflow starts with allele read counts quantified by CRISPResso2 and automates the cell genotyping process to output cell-level genotype predictions and sample-level composition estimates. Multiplets are simultaneously flagged using ploidy information and principal component analysis (PCA) of allele counts. **D.** Distribution of dominant allele frequency across all droplets and identification of homozygous cells by GUMM. scDNAseq protocol for 35% Hom, 55% Het, 10% WT cocktail included an additional cell filtering step prior to droplet barcoding to decrease multiplet rate. **E.** Distribution of $\log_2$ noise allele ratio which was computed by taking the read count ratio between the second and third most frequent alleles for each droplet. GUMM fits a univariate GMM to this statistic to identify transparent multiplets. **F.** Correlogram showing allele correlation structure revealing heterozygous alleles with high co-occurrence. PCA was performed on allele counts to generate low dimensional representation of data. **G.** Scatterplots showing opaque multiplet identification by fitting GUMM to first two principal components. Dense cluster of droplets correspond to true heterozygous cells with -8/-9 compound deletion. GUMM is robust to multiplet contamination in data.

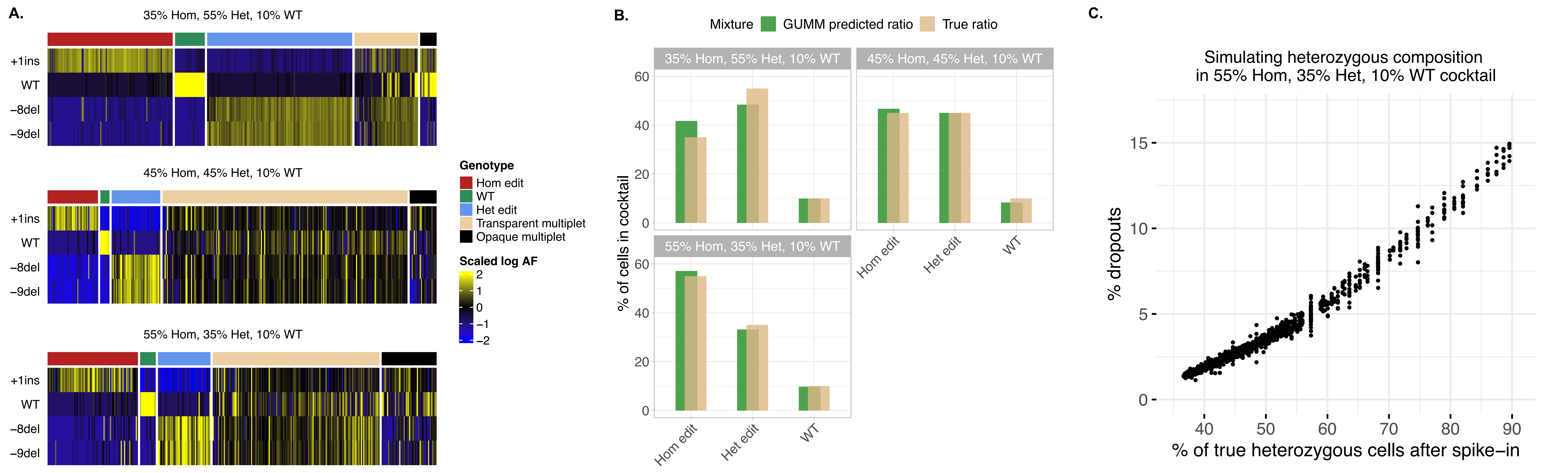


Figure 3) GUMM accurately predicts cell genotype and estimates original cocktail ratios

A. Heatmaps showing log-transformed frequencies of four possible alleles at edit site across droplets for all three artificial cocktails. Annotation bar indicates genotype or multiplet category predicted by GUMM. Allele frequency patterns of predicted genotypes are consistent with "ground truth" **B.** Bar plots comparing estimated and true cocktail genotype compositions after removing multiplets. The 35% Hom, 55% Het, 10% WT cocktail showed greater deviation from ground truth compared to other cocktails due to higher dropout rate which correlates with greater Het composition. **C.** Simulating heterozygous composition of 55% Hom, 35% Het, 10% WT cocktail by *in silico* spike-in. Cells were randomly selected from the pure 100% heterozygous sample data and computationally added to artificial cocktail data at increasing frequency. This was performed concurrently with cocktail down-sampling to increase the maximum heterozygous rate in the cocktail. Each random sampling event was repeated 5 times to ensure robustness. An association between cocktail heterozygous composition and dropout rate was observed in the simulation.

Conclusion

Producing robust and safe cell therapies through gene editing requires careful optimization and vetting of cell products. In this study, we developed a computational workflow called GUMM to rapidly genotype scDNAseq data from gene editing experiments. We applied our method to artificial mixtures of CRISPR-Cas9 edited HL-60 clones with distinct allele combinations at a single target gene. Our workflow accurately genotyped individual cells based purely on various transformations of the allele frequency readout produced by CRISPResso2. It remained robust to data containing technical artifacts including amplification bias and multiplet contamination. Our study provides both a rich data resource and novel bioinformatic solution for researchers in the gene editing community looking to characterize complex genotypes in engineered cell populations.

References

1. Clement, K., Rees, H., Canver, M.C. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat Biotechnology* **37**, 224–226 (2019).
2. Prates, M. O., Cabral, C. R. B. & Lachos, V. H. mixsmsn : Fitting Finite Mixture of Scale Mixture of Skew-Normal Distributions. *J. Stat. Soft.* **54**, (2013).
3. Scrucca, L., Fop, M., Murphy, T. B. and Raftery, A. E. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models *The R Journal* **8**, 289-317 (2016).
4. ten Hacken, E., Clement, K., Li, S. *et al.* High throughput single-cell detection of multiplex CRISPR-edited gene modifications. *Genome Biology* **21**, 266 (2020).

Abbreviations

WT: Wild type
PCR: Polymerase Chain Reaction
Hom: Homozygous
Het: Heterozygous
PCA: Principal Component Analysis
GMM: Gaussian Mixture Model
GUMM: Genotyping Using Mixture Models
AF: Allele Frequency
Del: Deletion
Ins: Insertion