P203

A single cell DNA sequencing resource and computational approach to quantify CRISPR-Cas9 gene editing allelism

Matthew H. Ung[1],*, Ruijia Wang[1],*, Gabriella Angelini[1], Juliana Xavier-Ferrucio[1], Michelle Lin[1], Tirtha Chakraborty[1], Gary Ge[1]

[1]Vor Bio, Cambridge, MA, USA, *Authors contributed equally

Introduction

CRISPR-Cas9 gene editing is a powerful approach to improve our ability to treat specific diseases with an unmet medical need. Engineering cell therapies requires accurate assessment of allelism as editing patterns can vary across cells and cause phenotypic heterogeneity in a sample. Bulk sequencing is the current standard for assessing editing frequency but is not always sufficient for quantifying the diversity of bi- and mono-allelic knockout events in a cell population. This limitation can delay development of more complex cell therapies involving 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 at unprecedented resolution. However, to systematically analyze these data we must address technical artifacts that could arise including low coverage over the editing site, PCR amplification bias, and multiplets; all of which confound accurate genotyping and quantification of edited and unedited cells in a sample. In this study, we introduce a "ground truth" single cell gene editing data resource (>20,000 cells) to explore these artifacts in a controlled setting and develop computational solutions to circumvent issues that may arise when applying this technology to gene editing.

Methods

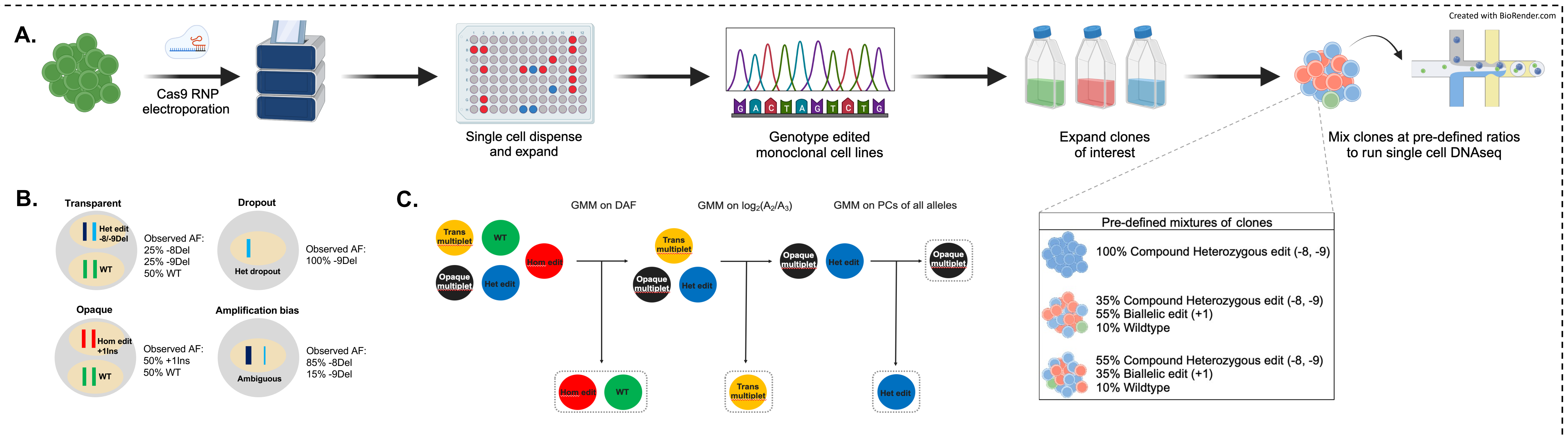
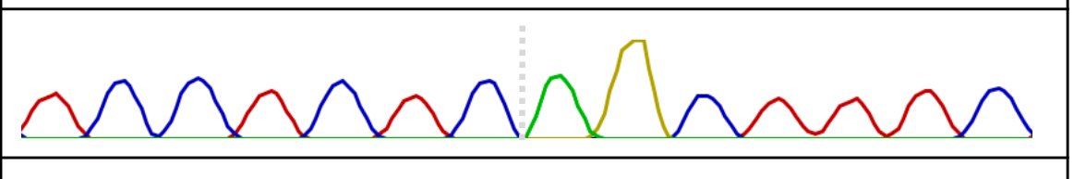
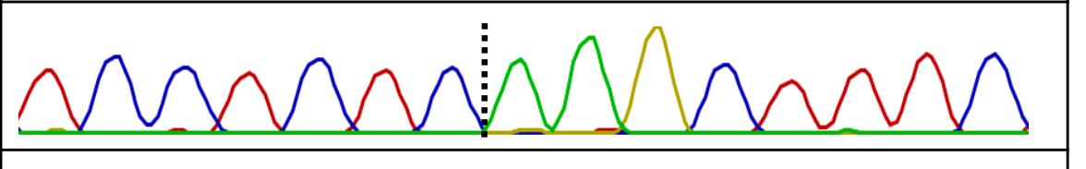
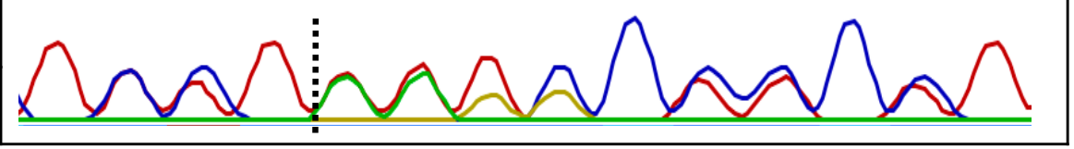

Created with BioRender.com
A.
Cas9 RNP electroporation
Single cell dispense and expand
Genotype edited monoclonal cell lines
Expand clones of interest
Mix clones at pre-defined ratios to run single cell DNAseq
B.
Transparent
Het edit -8/-9Del
WT
Observed AF: 25% -8Del 25% -9Del 50% WT
Dropout
Het dropout
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.
GMM on DAF
GMM on log2(A2/A3)
GMM on PCs of all alleles
Trans multiplet
WT
Opaque multiplet
Het edit
Hom edit
Pre-defined mixtures of clones
100% Compound Heterozygous edit (-8, -9)
35% Compound Heterozygous edit (-8, -9)
55% Biallelic edit (+1)
10% Wildtype
55% Compound Heterozygous edit (-8, -9)
35% Biallelic edit (+1)
10% Wildtype

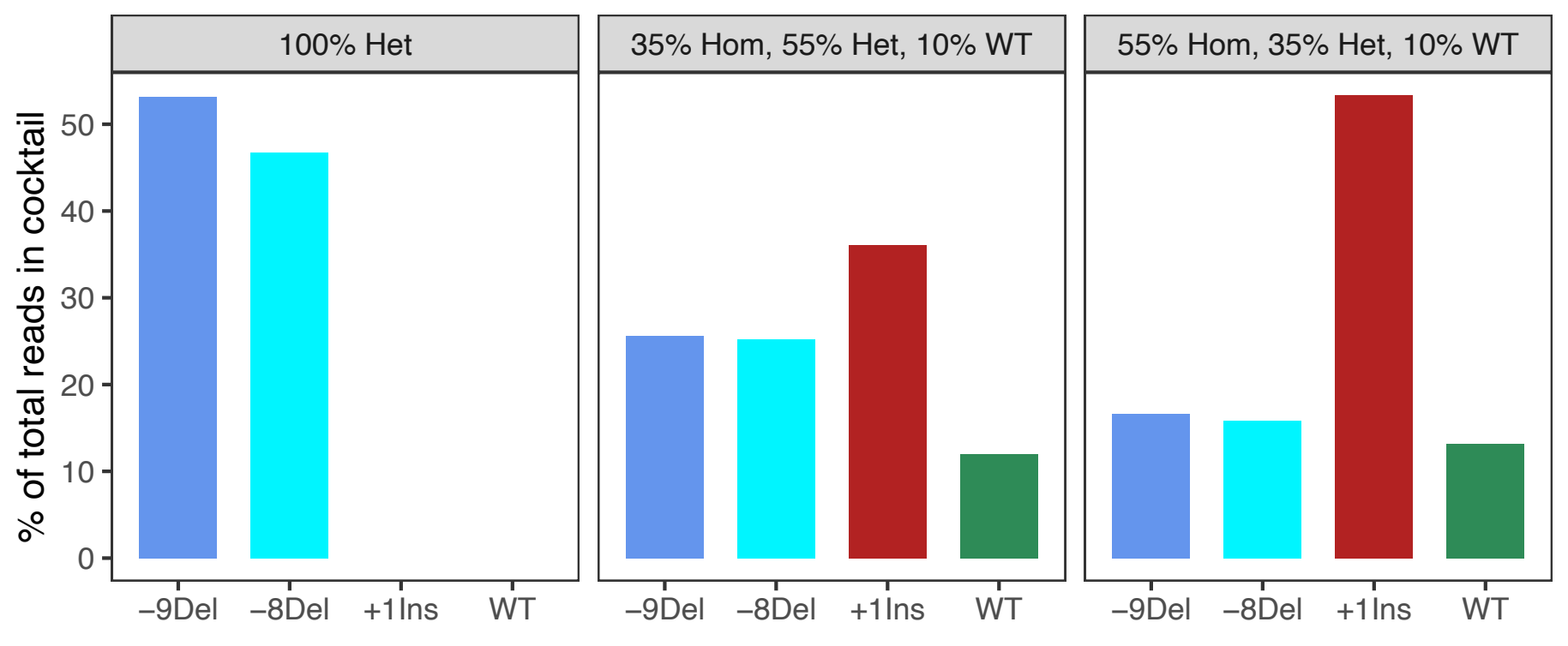
Figure 1) Construction of a gene editing scDNAseq resource and automated single cell genotyping approach A. We used CRISPR-Cas9 to create edited HL-60 cells and expanded clones with distinct homozygous and compound heterozygous indel profiles. Clones were mixed at pre-defined ratios to create artificial 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.** Diagram illustrating potential artifacts that can confound interpretation of gene editing. **C.** Overview of computational analysis 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 begins with allele read counts quantified by CRISPResso2[1] and automates the cell genotyping process to output cell-level genotype predictions and sample-level composition estimates.

Results

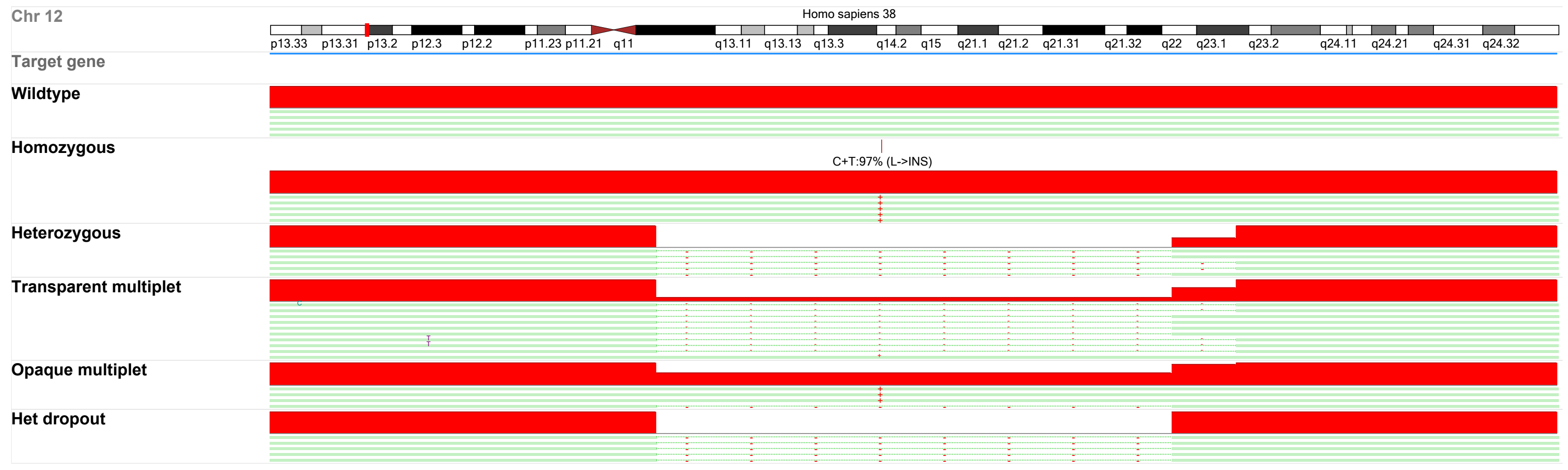
A.

HL-60 clone	**Sequence** (cut site indicated by |)	**Electropherogram** (cut site indicated by ┆)
Wildtype	ACCACAC|TGCAAAC	
Homozygous edit (+1 ins)	ACCACAC|**+T**TGCAAAC	
Compound heterozygous edit (-8, -9)	ACCA---|-----ACAATAGCC ACCACA-|--------ATAGCC	

B.



C.



Cell barcode	Genotype	+1Ins	WT	-8Del	-9Del
AACAACCTAGGTAGCATC-1	WT/Unedited	0	**306**	0	0
TATCACCTGGGAATTCAC-1	Homozygous edit	**344**	**0**	0	0
AACAGCAGTCCTCCAATC-1	Heterozygous edit	0	0	**26**	**38**
AATTGGTGATACCGCGTT-1	Transparent multiplet	**54**	0	**26**	**28**
CCTCAGGTGCGTACATCT-1	Opaque multiplet	0	**34**	**72**	0
AAGGTCTGAACGCTATGT-1	Heterozygous dropout	0	0	0	**138**

Figure 2) Cell and allele composition in single cell readout of HL-60 clone cocktails A. Electropherogram and ICE[2] results from Sanger sequencing of monoclonal cell lines show distinct alleles of target gene exclusive to each clone. **B.** Pseudobulk quantification of clonal alleles in scDNAseq data from three artificial cocktails. **C.** Single cell read alignment profiles (genome browser) and allele read counts (table) of clone-specific genotypes observed in data. Homozygous cells have a +1 insertion and heterozygous cells have a -8 and -9 compound deletion at the editing site. Transparent multiplets display a triploid profile and opaque multiplets display a diploid profile of a non-existent genotype. Heterozygous dropouts are falsely "homozygous" for either the -8 or -9 deletion allele.

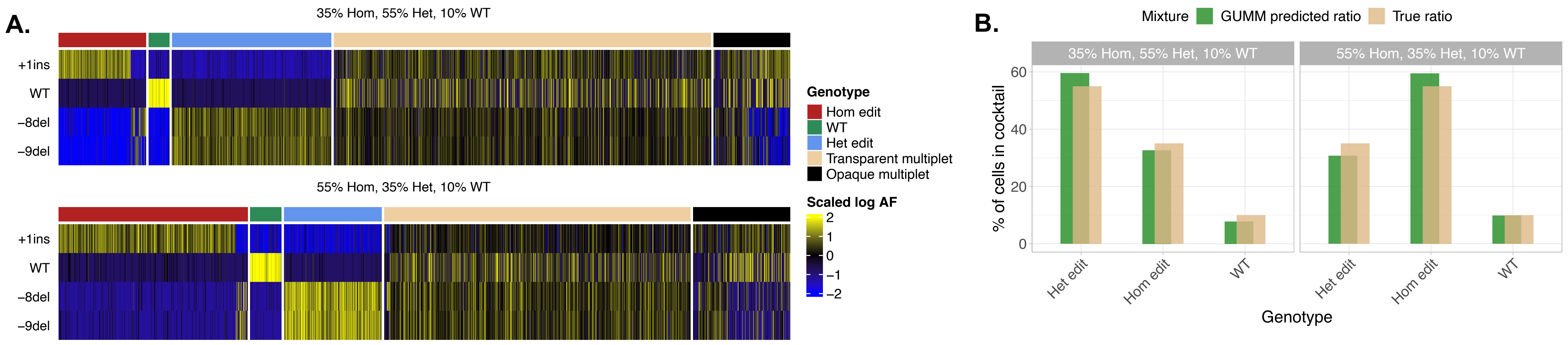


Figure 4) Automated genotyping enables accurate estimation of original clonal mixture composition A. Heatmap showing scaled frequencies for the four possible allele combinations in cocktail. Top annotation bar shows cell genotype predicted by Gaussian mixture modeling. Predicted genotypes correlate well with expected allele frequencies. Heterozygous dropouts cannot be detected by this method **B.** Estimated frequencies of homozygous, heterozygous, and WT cells in cocktail after genotype prediction with Gaussian mixture models. Composition estimate strongly correlates with true mixing ratios after removing or splitting multiplets.

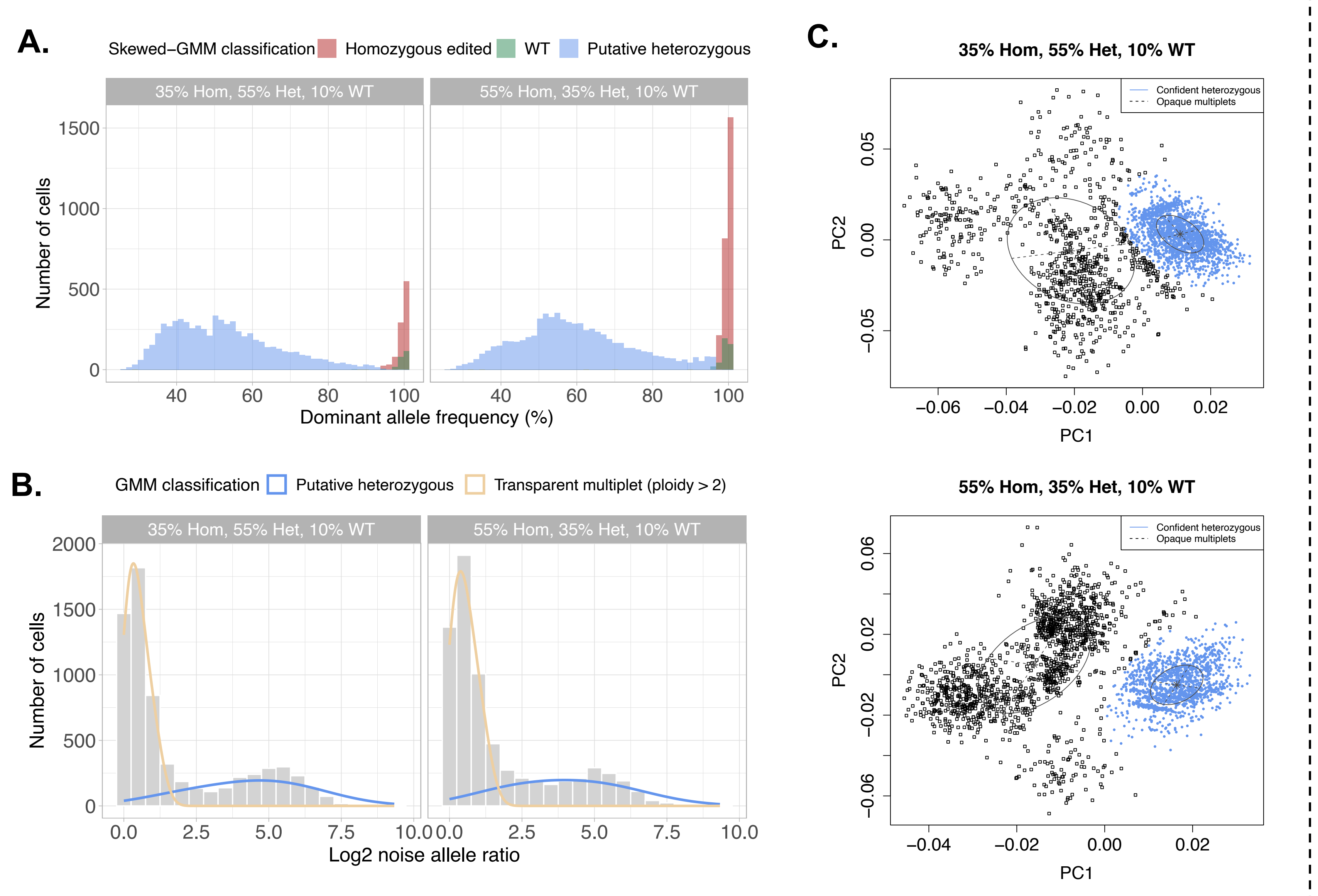


Figure 3) Application of Gaussian mixture models to automate single cell genotyping A. Distribution of the dominant allele frequency across all cells. A skewed GMM was used to identify homozygous edited or homozygous WT cells in a probabilistic fashion **B.** A GMM model was fit to the distribution of $\log_2$ ratios of the second and noise (third) alleles computed across non-homozygous cells. Cells with small ratios (left peak) are labeled as transparent multiplets **C.** A multivariate GMM model was used to discriminate true heterozygous cells from diploid cells with rare allele combinations.

Discussion

Producing robust and safe cell therapies through gene editing requires careful optimization and vetting of cell products. In this study, we apply scDNAseq to artificial mixtures of CRISPR-Cas9 edited HL-60 clones with distinct allele combinations at a single target gene. We explore potential technical artifacts in the data relevant to gene editing experiments and developed a novel computational workflow to automate the cell genotyping process and show that it is robust to data with high multiplet rates. Moving forward, additional scDNAseq data from HL-60 clones edited at multiple genes will be included in our atlas. This comprehensive data resource can be used to establish how future cell therapy products involving CRISPR-Cas9 editing are analyzed through sequencing. 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.

Citations

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. Conant, D., Hsiau, T., Rossi, N. *et al.* Inference of CRISPR Edits from Sanger Trace Data. *The CRISPR Journal* **5**, 123-130 (2022).
3. 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).
4. 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).

Abbreviations

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