P71

A machine learning approach incorporating germline information improves genotyping of CRISPR-Cas9 gene editing events at single cell resolution

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

Vor Biopharma, Cambridge, MA, USA

VOR

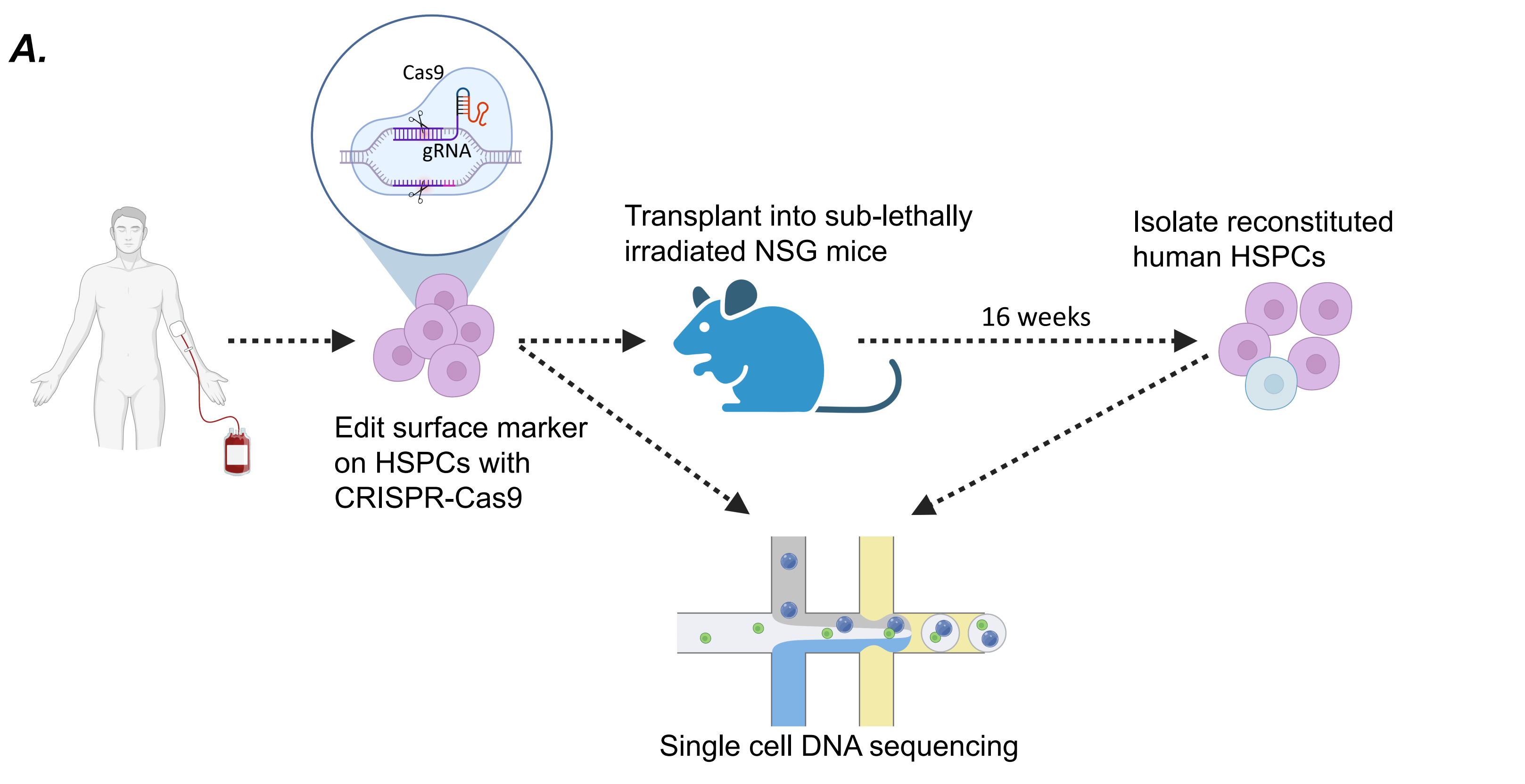
Introduction

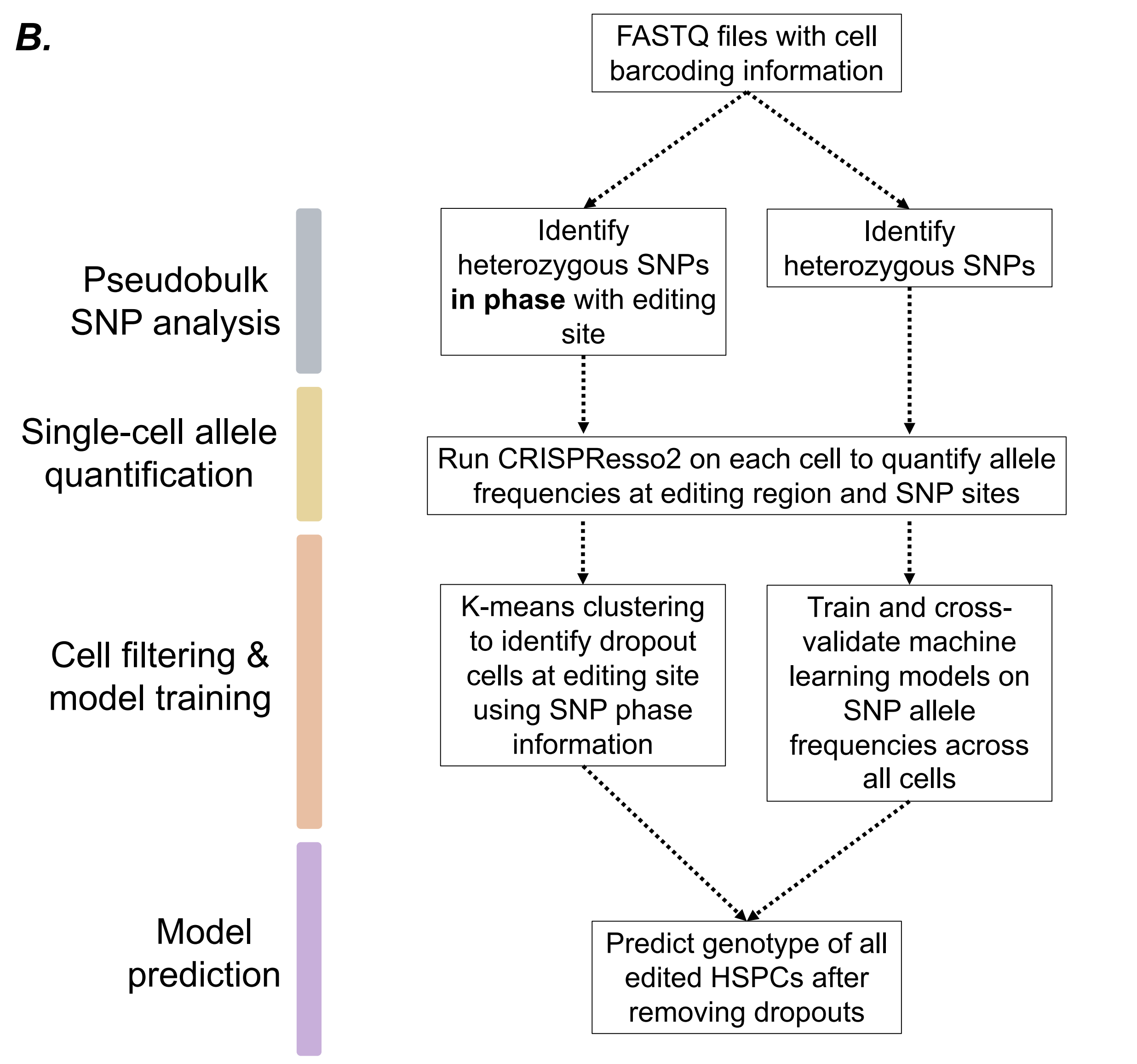
CRISPR-Cas9-based gene editing is a powerful approach to improve our ability to treat specific diseases with an unmet medical need. Developing robust cell therapies with genome engineering requires rigorous assessment of allelism at single cell resolution, especially when multiple targets are considered. Recently, droplet-based targeted single cell DNA sequencing (scDNAseq) has been used to genotype selected loci across thousands of cells enabling high-throughput assessment of gene editing efficiency. However, several technical issues must be accounted for including low sequencing depth and PCR amplification bias due to low input DNA in each droplet. These artifacts skew allele read frequencies in the readout which can confound accurate genotyping.

In this study we:

- ▷ Addressed these issues by developing a machine learning method that learns the extent of this skew from single nucleotide polymorphisms (SNPs) across all cells and amplicons
- ▷ Determined that SNPs can be initially identified through pseudobulk genotyping and in theory should be detectable in every cell because they occur in the germline
- ▷ Analyzed scDNAseq data generated from Cas9-edited human hematopoietic stem and progenitor cell (HSPC) samples before and after *in vivo* transplantation into mouse bone marrow
- ▷ Found that the model trained and cross-validated on observed heterozygous and homozygous SNPs across all cells was able to predict genotype with greater accuracy than GATK

Methods


A.
Cas9
gRNA
Edit surface marker on HSPCs with CRISPR-Cas9
Transplant into sub-lethally irradiated NSG mice
16 weeks
Isolate reconstituted human HSPCs
Single cell DNA sequencing


B.
Pseudobulk SNP analysis
Single-cell allele quantification
Cell filtering & model training
Model prediction

FASTQ files with cell barcoding information

Identify heterozygous SNPs **in phase** with editing site

Identify heterozygous SNPs

Run CRISPResso2 on each cell to quantify allele frequencies at editing region and SNP sites

K-means clustering to identify dropout cells at editing site using SNP phase information

Train and cross-validate machine learning models on SNP allele frequencies across all cells

Predict genotype of all edited HSPCs after removing dropouts

Fig. 1) Overview of experimental and computational pipeline **A.** Experimental design of HSPC gene editing and transplantation. Human primary HSPCs were edited on one gene and transplanted into mouse bone marrow. Droplet-based targeted single cell DNA sequencing was performed on both edited input HSPCs and HSPCs collected from mouse bone marrow 16 weeks after transplantation. Amplicon panel used covered editing site and several heterozygous SNPs **B.** Flowchart of computational method for single cell genotyping of CRISPR-Cas9 editing. Genotyping of single cells was performed through a 4-step process involving robust SNP identification, allele quantification with CRISPResso2, model training, and prediction.

Results

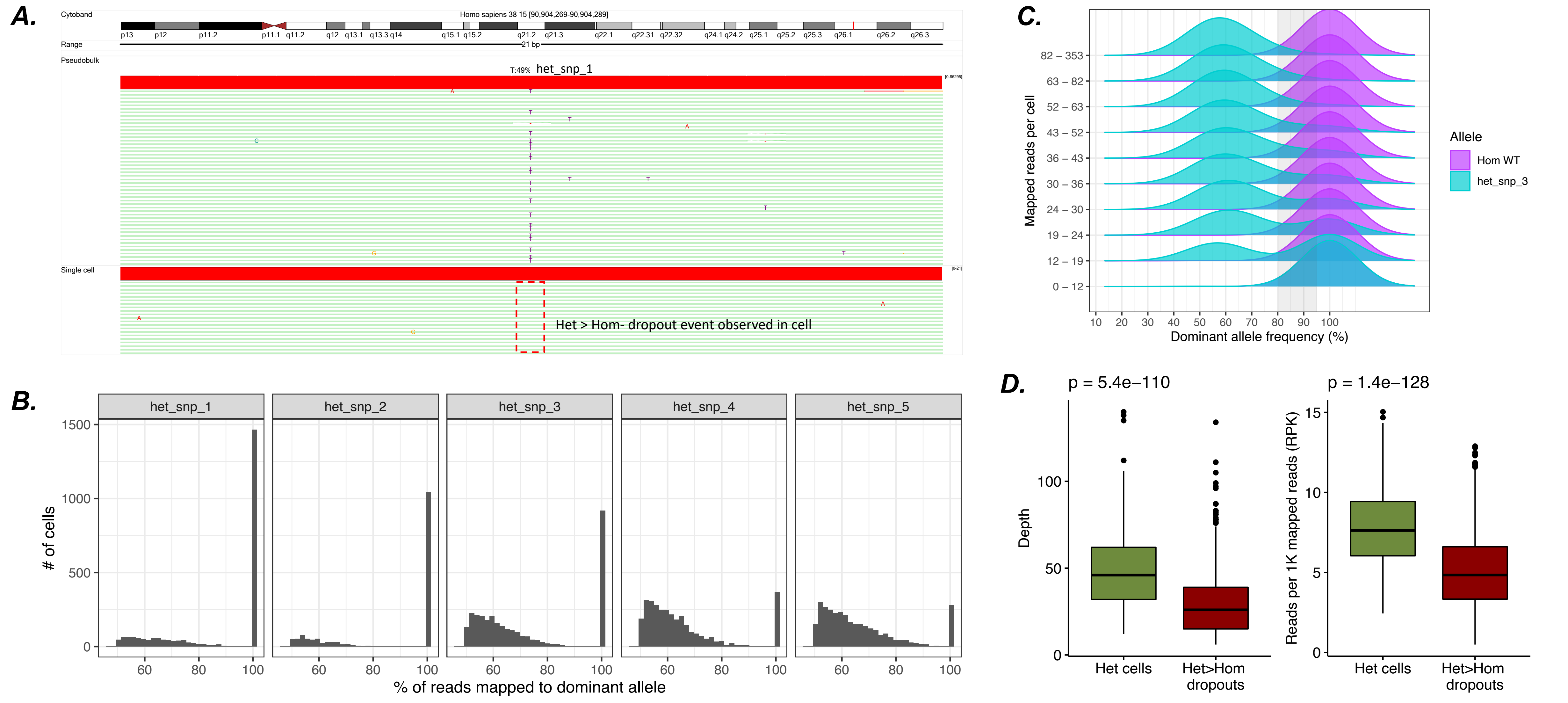
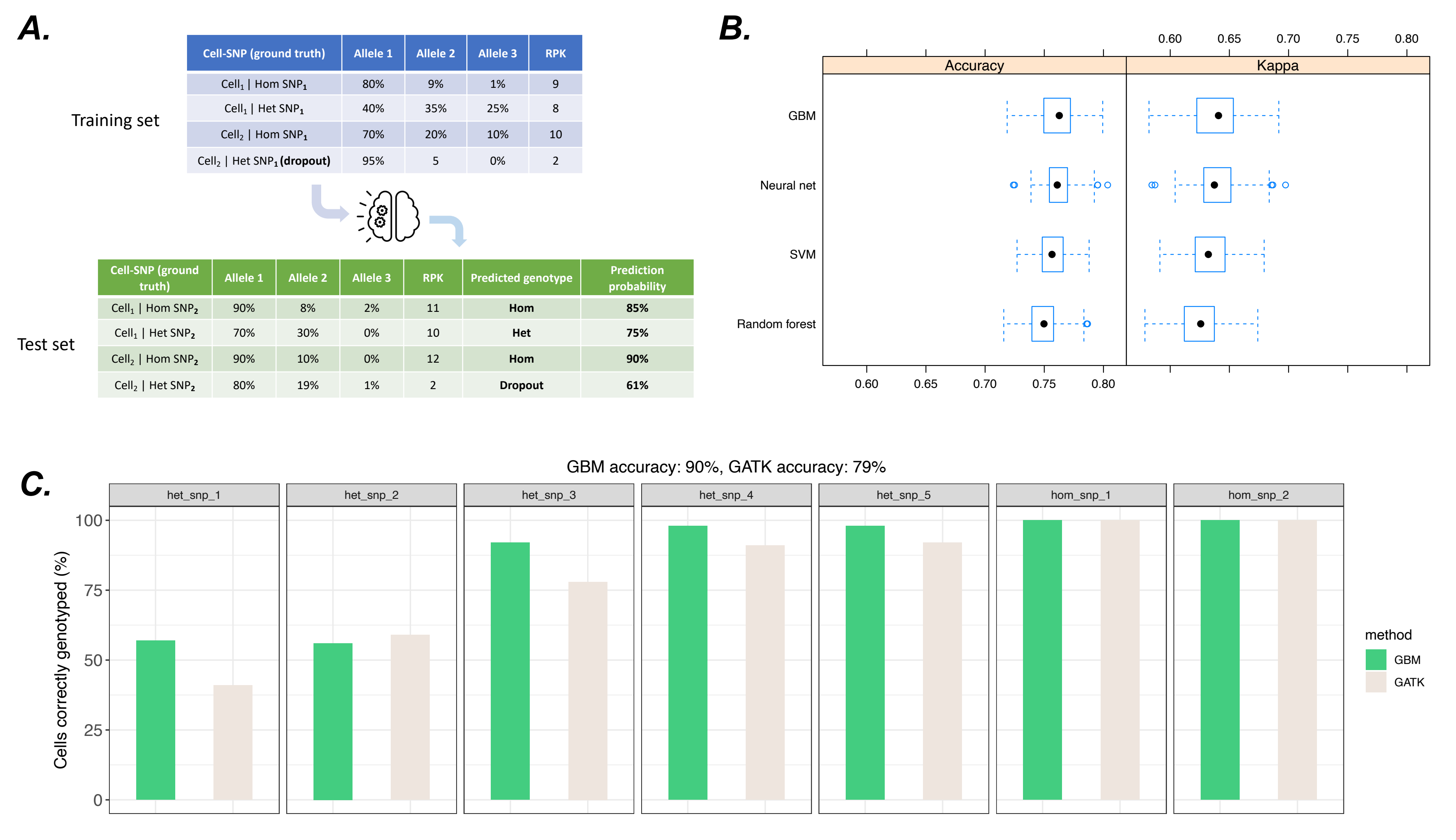


Fig. 2) Allele frequencies at single cell resolution harbor technical artifacts **A.** Alignment profiles of pseudobulk and single cell reads. Amplification bias can cause a Het to Hom transition in single cell readout. **B.** Allele frequency distribution of allele with most mapped reads (dominant allele) across cells. Five SNPs with varying dropout rates were identified from pseudobulk and used in downstream analysis. **C.** Amplification bias and genotype ambiguity is associated with low read depth where dominant allele frequency distributions deviate from the expected 50% and 100% frequencies for heterozygous and homozygous alleles, respectively. Gray shading corresponds to range of ambiguous frequencies observed in CRISPResso2 output. **D.** Normalized depth (reads per thousand mapped reads) associates more strongly with dropouts compared to absolute read depth.



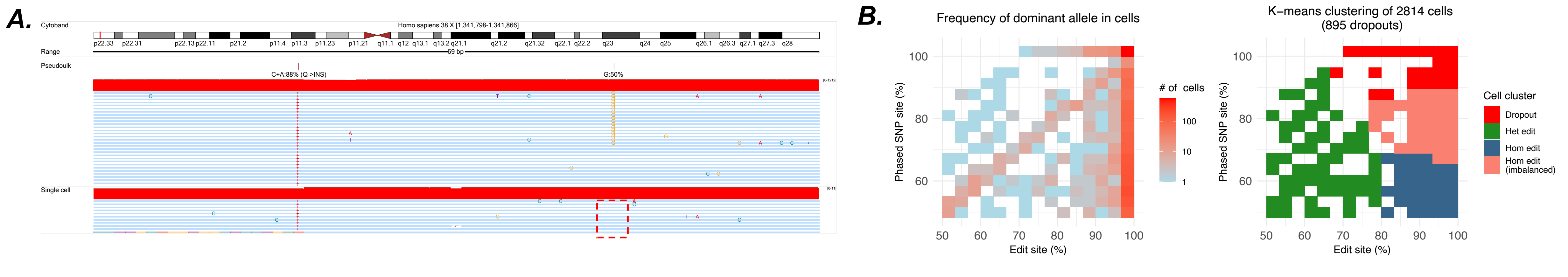
Training set

Cell-SNP (ground truth)	Allele 1	Allele 2	Allele 3	RPK
$Cell_1$ | Hom SNP_1	80%	9%	1%	9
$Cell_1$ | Het SNP_1	40%	35%	25%	8
$Cell_2$ | Hom SNP_1	70%	20%	10%	10
$Cell_2$ | Het SNP_1 **(dropout)**	95%	5	0%	2

Test set

Cell-SNP (ground truth)	Allele 1	Allele 2	Allele 3	RPK	Predicted genotype	Prediction probability
$Cell_1$ | Hom SNP_2	90%	8%	2%	11	**Hom**	**85%**
$Cell_1$ | Het SNP_2	70%	30%	0%	10	**Het**	**75%**
$Cell_2$ | Hom SNP_2	90%	10%	0%	12	**Hom**	**90%**
$Cell_2$ | Het SNP_2	80%	19%	1%	2	**Dropout**	**61%**

Fig. 3) Training machine learning models with germline information **A.** Construction and organization of tabular input data. Mapped read frequencies for the dominant allele, secondary allele, and noise allele were used along with normalized read depth as features. Each observation consists of a cell barcode (3219 cells) and SNP pair. **B.** Evaluation of different machine learning models trained on heterozygous, homozygous, and dropout labeled data. Repeated 10-fold cross validation with class down-sampling was used to compare prediction accuracy across models. **C.** Comparison of gradient boosted machine (GBM) model with GATK. Each SNP was left out as testing data and the remaining SNPs were used to train GBM model. GATK was implemented on a per-cell basis after read mapping. Observations predicted as "dropouts" by machine learning were removed prior to comparison. GATK does not incorporate dropout detection.



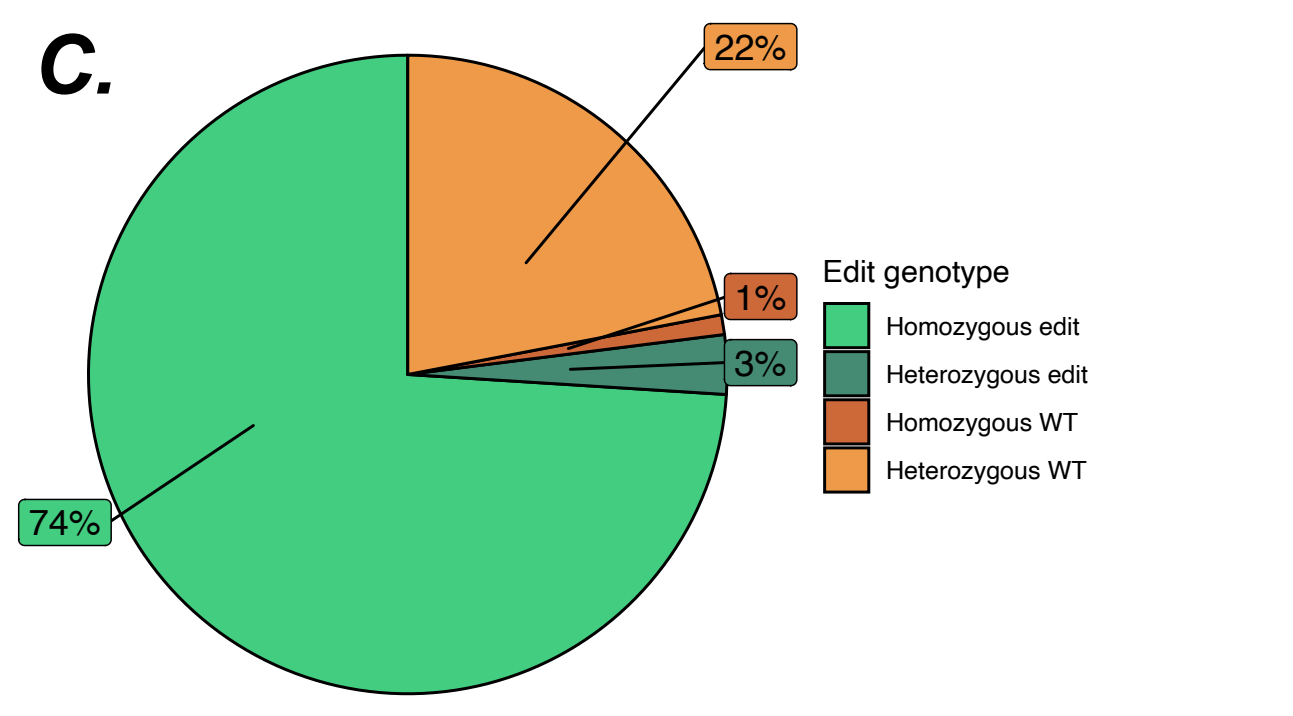


Fig. 4) Identifying true dropout cells at edit site using phase information from adjacent heterozygous SNPs
A. Pseudobulk and single cell read alignment profiles from scDNAseq readout of edited HSPCs transplanted into mouse bone marrow for 16 weeks. Mapped reads show heterozygous SNP phased with CRISPR-Cas9 edit site. Single cell profile shows Het to Hom transition at SNP site indicating a dropout event. **B.** Binned scatterplot showing joint allele frequencies between edit site and phased SNP across cells. Non-uniformity of cell density suggests the presence of distinct clusters corresponding to heterozygous, homozygous, and dropout cells. K-means clustering of edit site and SNP allele frequencies automates dropout detection. **C.** Percent breakdown of editing genotype predicted by gradient boosted machine model trained on allele frequencies from 7 SNPs after removing dropouts.

Discussion

CRISPR-Cas9 technology has enabled the field of cell therapy to advance rapidly due to its on-target precision and relatively low off-target consequences. However, producing robust and safe cell therapies through gene editing requires careful optimization and vetting of cell products. This is especially true with multi-target editing, which has received recent attention in the cell therapy community. This will likely introduce greater complexity into the data analysis procedure and affect downstream interpretation. To ensure that a high percentage of cells do indeed receive a biallelic edit becomes a key question that cannot be addressed through bulk sequencing. By leveraging both state-of-the-art commercially available scDNAseq technology and purpose-built analytical methodologies we can study gene editing efficiency at single cell resolution providing unprecedented insight into the biology of cell therapy treatments. In this study, we explore technical artifacts associated with scDNAseq and develop a novel *in silico* methodology to address them. We tested the method on data derived from an experiment that mirrors the pre-clinical development process of a CRISPR-Cas9-based cell therapy. We find that our approach outperforms methods designed for bulk sequencing data by incorporating germline and artifact information that is embedded in the readout. Model accuracy may also improve as more experimental data becomes available for training. Lastly, our method can be extended to data from other gene editing technologies including base and prime editing.

References:

Clement, K., Rees, H., Canver, M.C. et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis. Nat Biotechnology. 37, 224-226 (2019).
Van der Auwera GA & O'Connor BD. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition).* O'Reilly Media.
Kuhn, M. (2008). Caret package. Journal of Statistical Software, 28(5)

Abbreviations:

SNP: Single Nucleotide Polymorphism
WT: Wild Type
Hom: Homozygous
Het: Heterozygous

HSPC: Hematopoietic Stem and Progenitor Cell
scDNAseq: Single Cell DNA Sequencing
PCR: Polymerase Chain Reaction
GATK: Genome Analysis Toolkit

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats
Cas9: CRISPR-associated protein 9
gRNA: guide RNA