Investigative single cell genetics: Determining the probability of a genotype after observing single-cell data enables robust database searches across mixture complexities

LFTDI

Qhawe A. Bhembe[a] M.S.; Nidhi C. Sheth[a] M.S.; Leah O'Donnell[c] M.S.; Ken R. Duffy[b] Ph.D.; Desmond S. Lun[a] Ph.D.; Catherine M. Grgicak[a] Ph.D.

[a]Rutgers University, Camden, US; [b]Northeastern University, Boston, US; [c]Maynooth University, Ireland

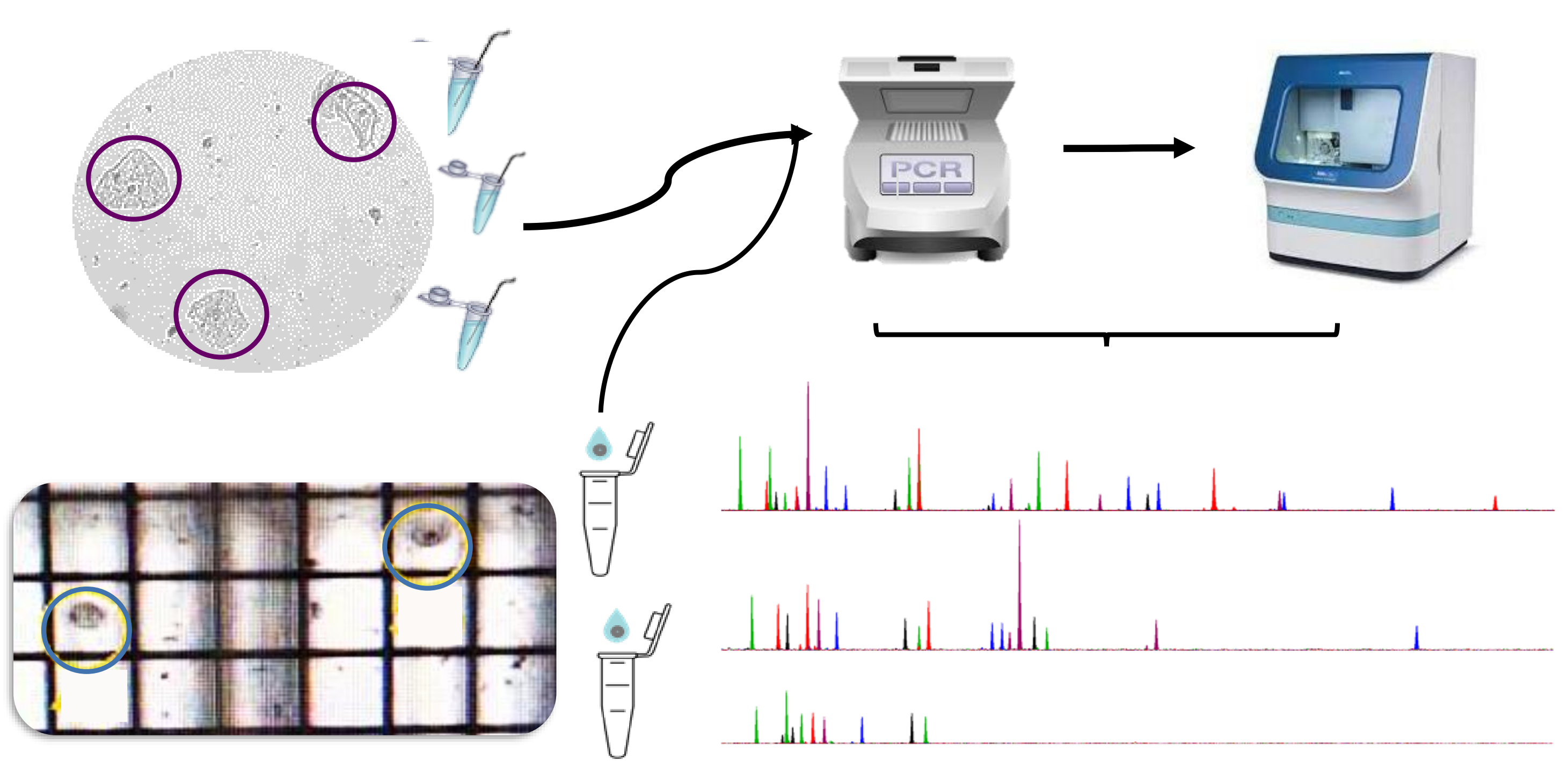
RUTGERS UNIVERSITY | CAMDEN

Highlights

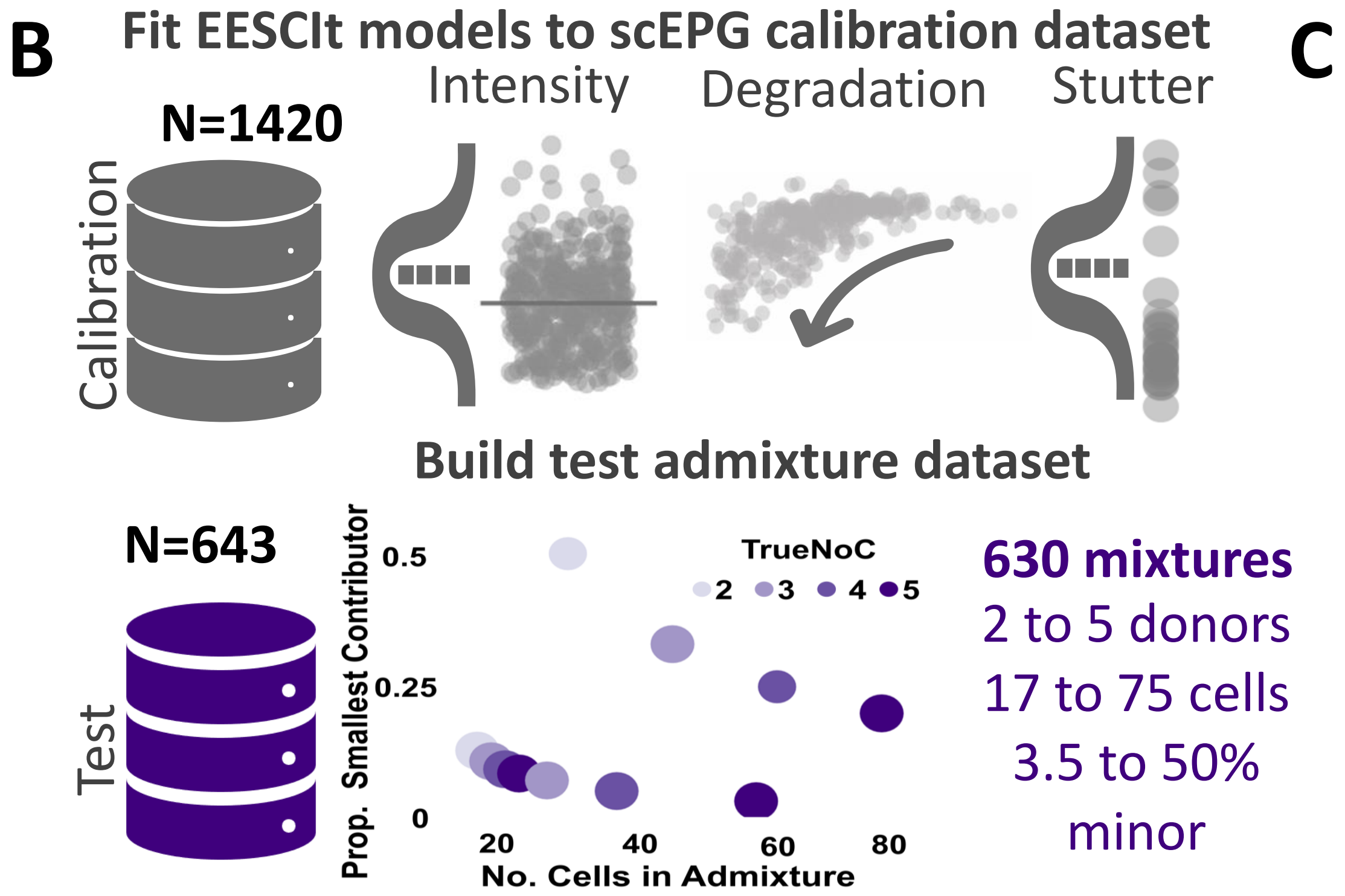
- We introduce a Bayesian approach that interprets forensic single-cell data in a suspect-agnostic manner, by
  - grouping single-cell electropherograms (scEPGs) via model-based clustering (MBC), based on scEPG similarity
  - using EESCIt™ – *Evidentiary Evaluation of Single Cells* – which determines the probability of observing the cluster of scEPGs given all possible genotypes
  - then determining the probability of a genotype, g, given the cluster, C, of scEPGs:

$$P(G=g|C)=\frac{P(C|G=g)P(G=g)}{\sum P(C|G=g_i)P(G=g_i)}$$

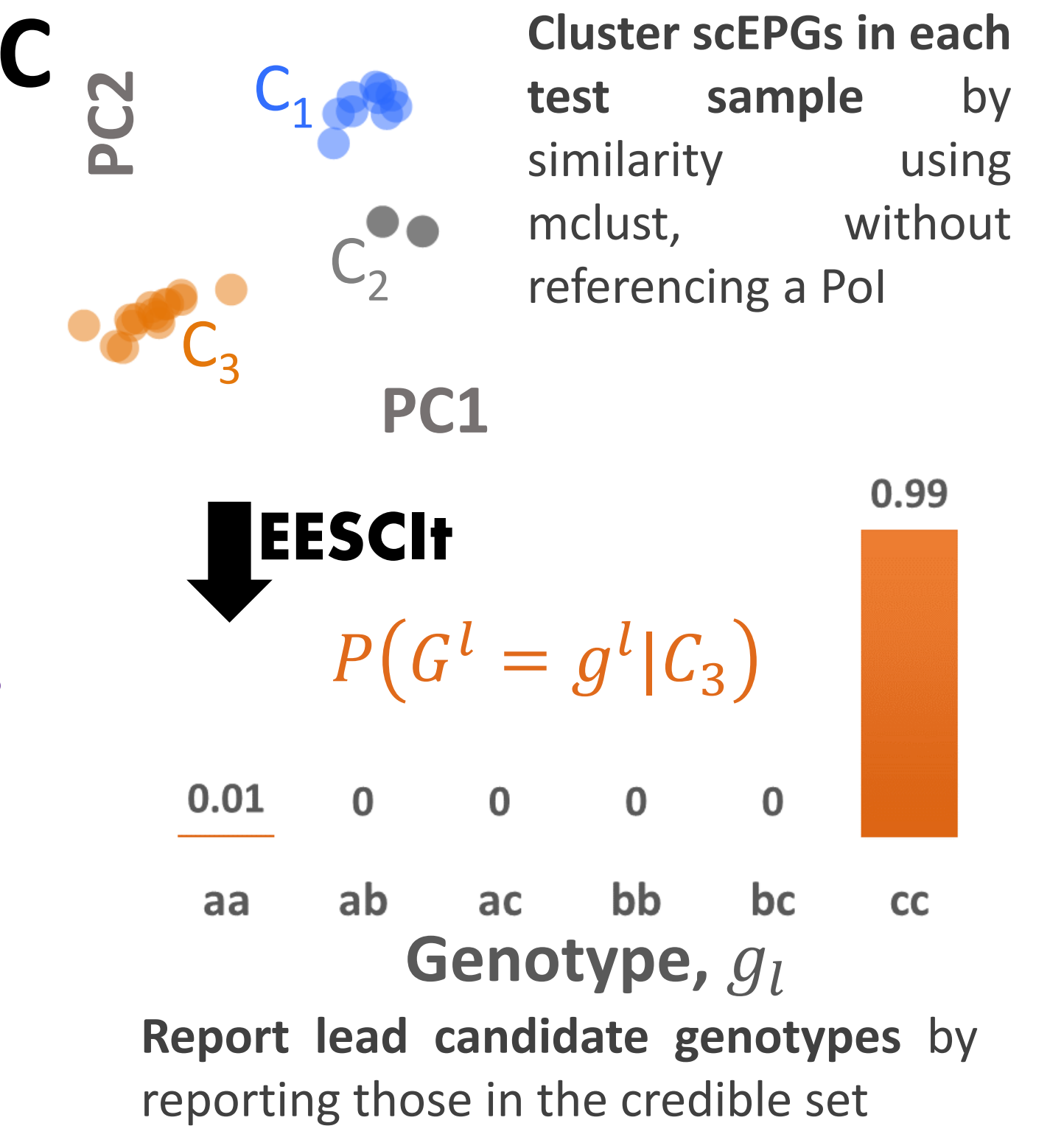
A Single-cell forensic workflow



Individual cells undergo isolation, followed by DNA extraction and amplification, resulting in one scEPG for each cell collected.



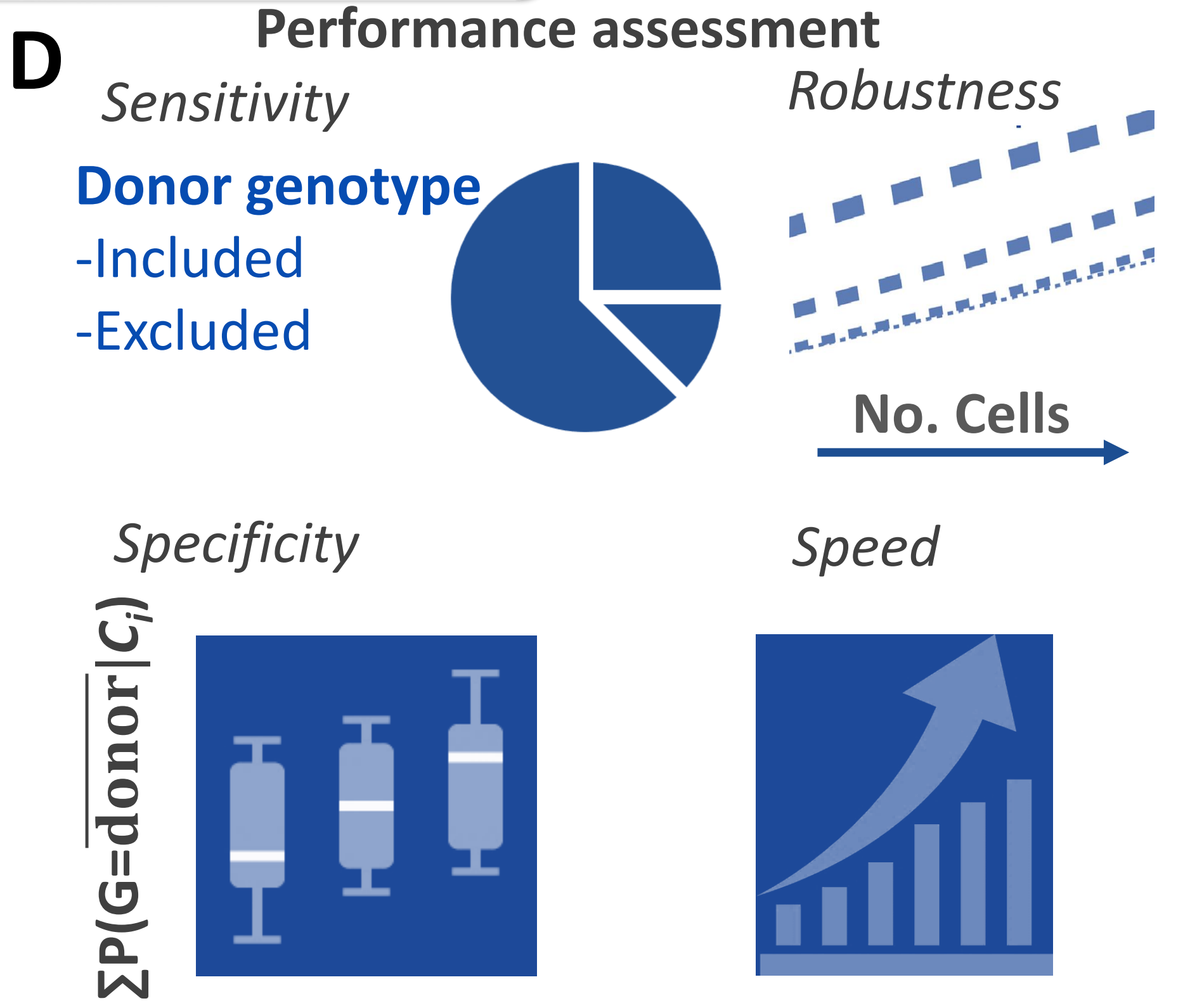
- scEPGs are divided into Calibration and Test sets, with the latter used to generate 630 admixtures comprising 2 to 5 donors and varying contributor ratios.



Cluster scEPGs in each test sample by similarity using mclust, without referencing a PoI

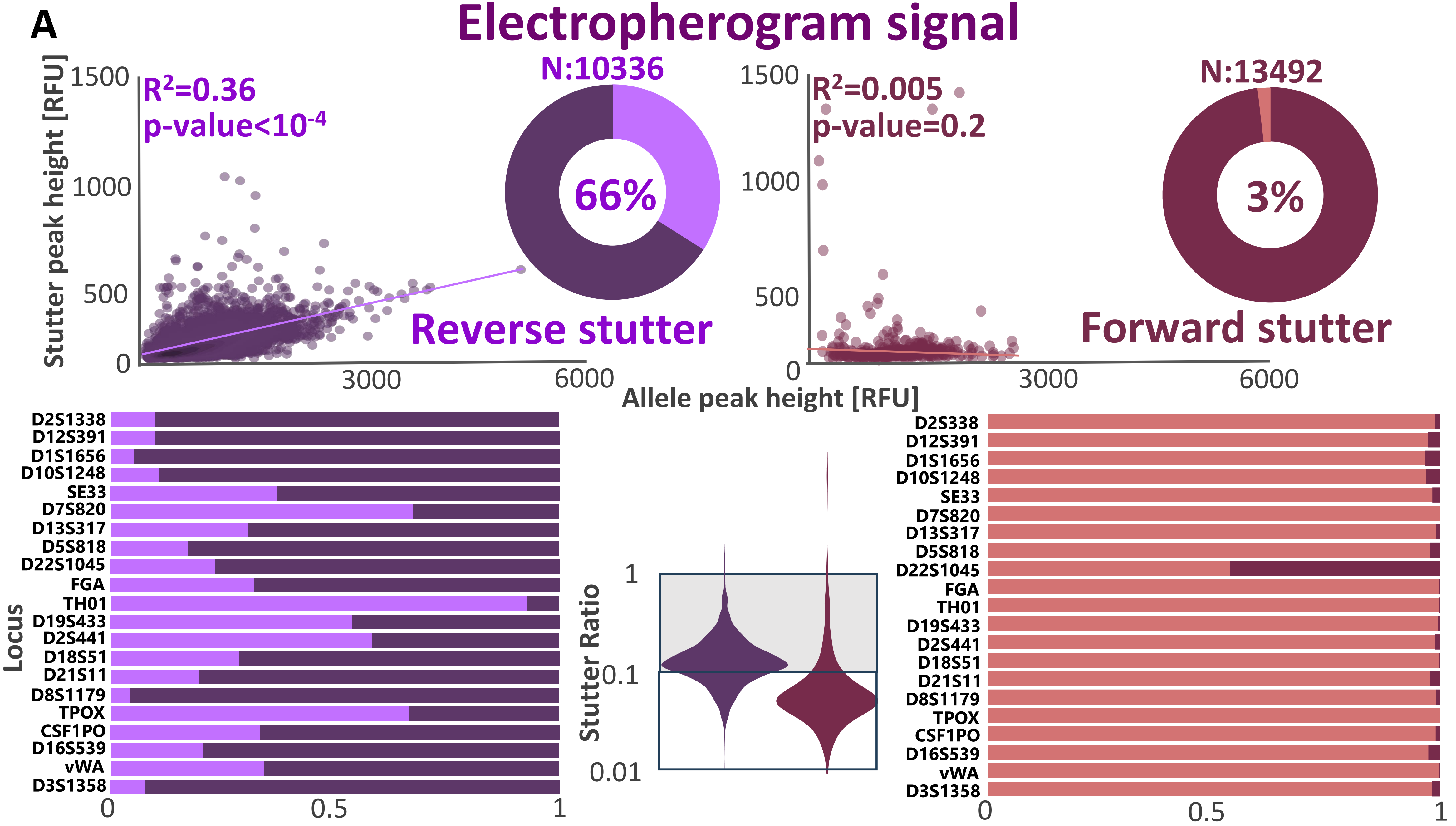
Report lead candidate genotypes by reporting those in the credible set

- Clustering enables genotype probability determinations per locus, given the scEPGs in a cluster.

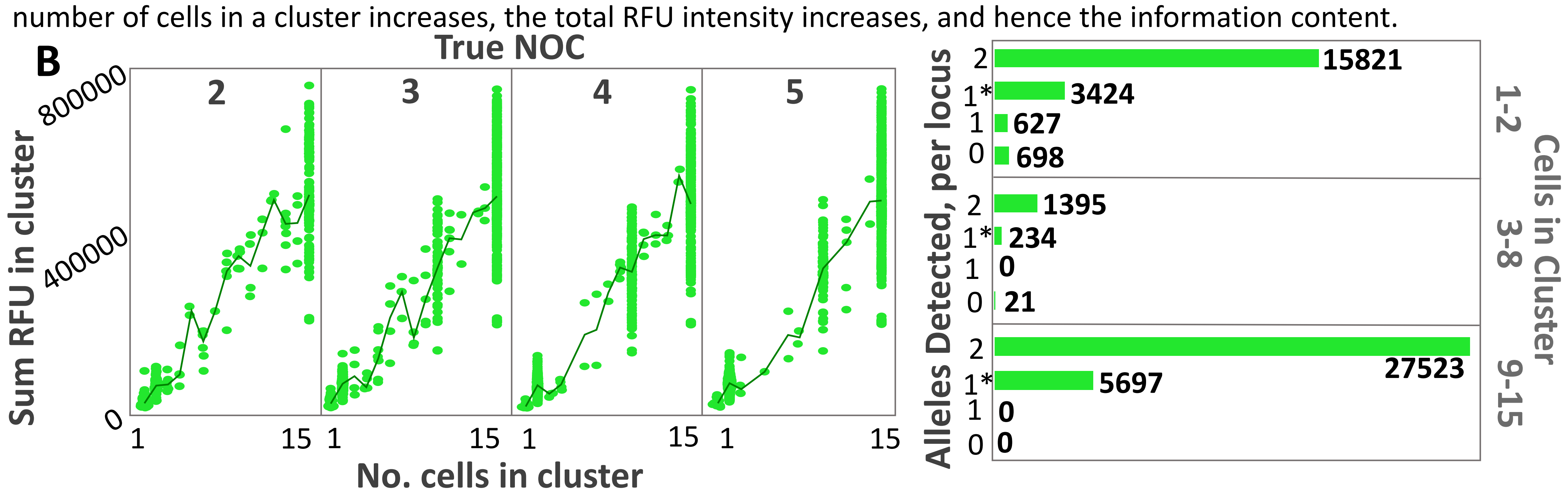


- Examine $P(G^l = g^l_{true}|C)$ & $P(G^l \neq g^l_{true}|C)$ across number of cells, total signal, and mixture ratios.

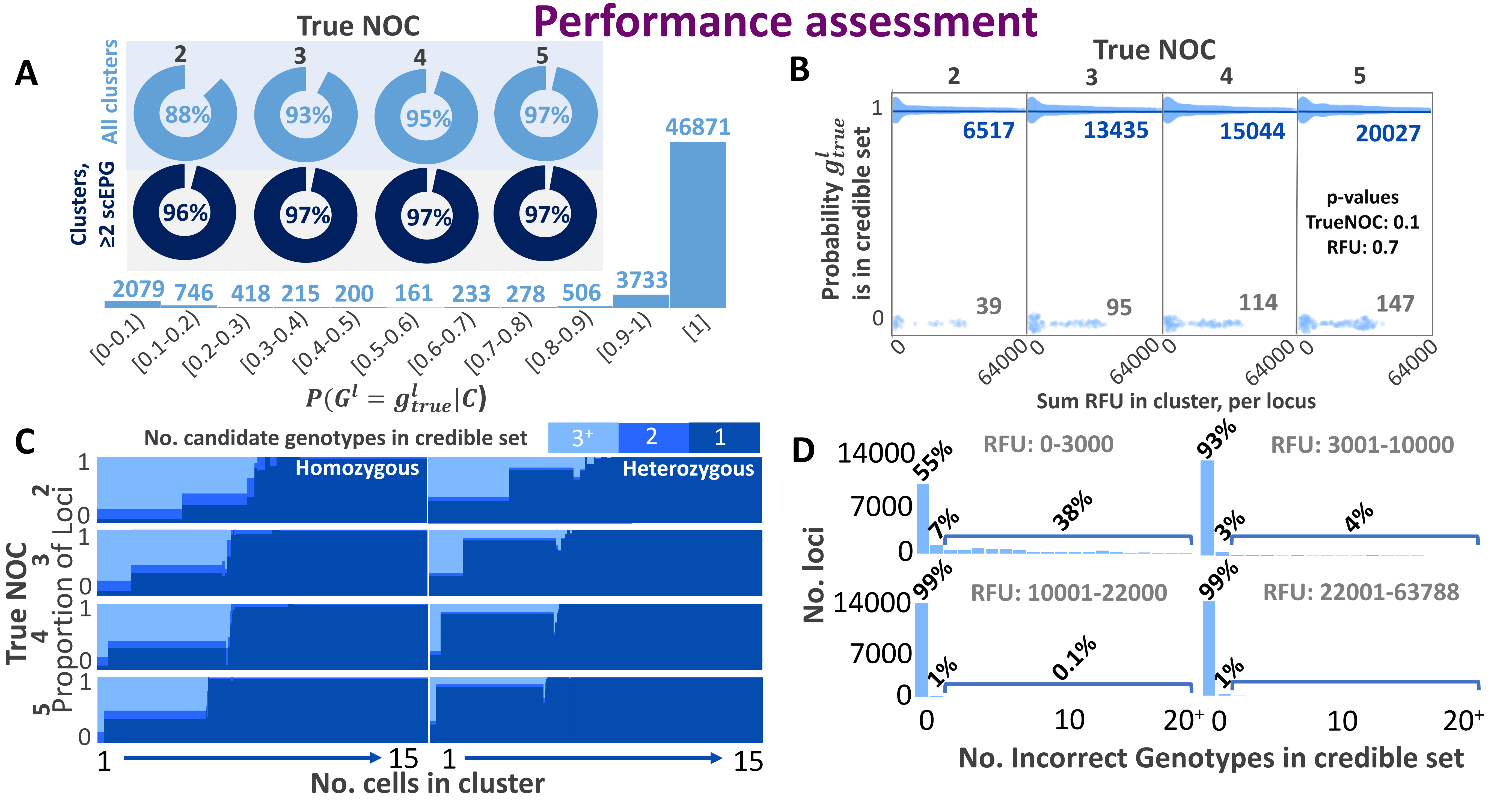
Electropherogram signal



(A) Pie charts display 66% and 3% of positions in reverse and forward stutter, respectively, with an RFU intensity greater than five, given that the allele position also has an RFU greater than five. The stacked plots separate these stutter proportions by locus, indicating a potential locus effect on reverse and forward stutter. The scatterplots present non-zero stutter peak heights (RFU) against allelic signal (RFU). The violin plot of the stutter ratio on a log scale. Average stutter ratio is 0.09 with 0.23 being the 95th percentile. **(B)** Scatter plots show the total RFU intensity across clusters with the number of cells in a cluster ranging from 1-15 cells, separated by NoC. The bar graph presents the number of detected alleles, grouped by the number of cells in a cluster, with an asterisk (*) denoting homozygous loci. As the number of cells in a cluster increases, the total RFU intensity increases, and hence the information content.



Performance assessment



(A) Histogram showing the posterior probability associated with the true genotype. Pie charts illustrate the proportion of clusters where the maximum probability aligns with the true genotype across NoC. **(B)** Scatterplots showing the probability that the list of candidate genotypes in the credible set contains the true genotype plotted against the sum of RFU intensity in a cluster per locus separated by NoC. **(C)** Mosaic plots showing the proportion of loci for which there were 1, 2, or 3+ genotypes included in the credible set separated by the number of cells in a cluster, which ranged from 1-15 and by whether the true genotype is homozygous or heterozygous. **(D)** Bar chart showing the number of incorrect genotypes included in the credible set separated by NoC and total intensity per locus.

Discussion: Our method consistently performs well and includes the true genotype in the set of candidate genotypes that explain the evidence regardless of NoC, number of cells in a cluster and peak heights. Low information content is the only factor that increased the number of candidate genotypes. Low information content is primarily associated with clusters consisting of only one cell.

Conclusion: Our findings demonstrate the legitimacy of single-cell data for investigative genetics. We observed consistent inclusion of the true genotype regardless of NoC, number of cells and peak height using a threshold of 0.998 to define the genotype set associated with that posterior probability.

Funding & Collaborators: This work was partially supported by NIJ2020-R2-CX-0032 and NIJ2018-DU-BX-K0185 awarded by the National Institute of Justice, Office of Justice Programs. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not reflect those of the Department of Justice.