Grouped Single Cell Electropherograms Enable Precise DNA Interpretation: Relevancy & Legitimacy of Single Cells to Forensics

Catherine M. Grgicak[1], Madison M. Mulcahy[1], Leah O'Donnell[2], Nidhi Sheth[1], Desmond S. Lun[1], Ken R. Duffy[2]

[1]Rutgers University Camden, [2]National University of Ireland Maynooth

NIJ | National Institute of Justice — STRENGTHEN SCIENCE. ADVANCE JUSTICE.

Forensic Technology CENTER OF EXCELLENCE — A program of the National Institute of Justice

RTI INTERNATIONAL

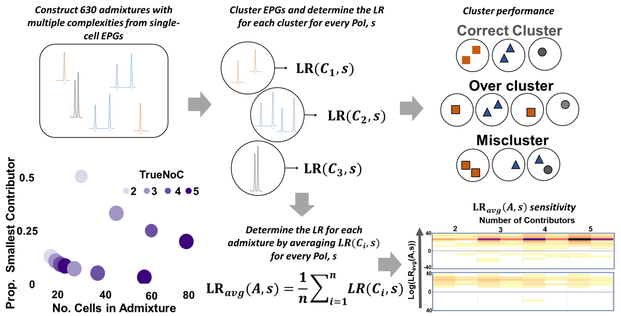
Highlights

- We present a likelihood ratio (LR) framework for the forensic interpretation of single cell data
- Successful reduction to practice is demonstrated by:
  - grouping single-cell electropherograms (scEPGs) via model-based clustering (MBC), which makes no reference to a person of interest (PoI), and showing high clustering sensitivities regardless of mixture complexity
  - using EESCIt™ – standing for *Evidentiary Evaluation of Single Cells* – and calculating the LR for each cluster, needing only the simplest propositions; that the number of contributors is one and no other persons contributed
  - presenting a new determination that the:

average LR across clusters for a given PoI is the single-cell analogue of the LR for bulk mixtures

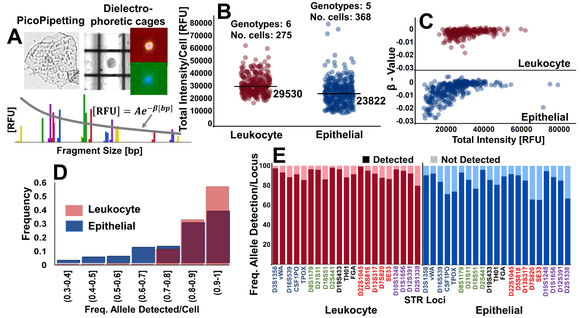
  - showing that the average single cell LR is robust, maintaining extremely high sensitivities and specificities, regardless of the complexity of the profile or number of PoI

Graphical Abstract



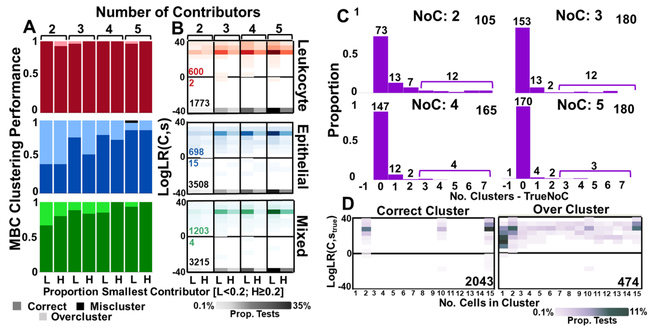
Admixture construction, MBC and EESCIt testing using single-cell data. Eleven types of admixtures were generated by combinatorically mixing scEPGs of known genotype to create admixtures containing 2- to 5- donors with 17 to 75 cells, and proportions of the least concentrated contributor ranging from 3.5% to 50%. Constructing combinatorial admixtures gives the ground-truth genotype of each cell enabling performance evaluation. We tested performance on two fronts: By determining the number of correct, over clustered and misclustered samples and confirming $\log \mathrm{LR}(C,s)$ was positive when $s = s_{\mathrm{true}}$ and negative when $s = s_{\mathrm{false}}$; and by assessing log LR for the entire admixture, which we denote as $\log_{avg} \mathrm{LR}(A,s)$.

Exploring Single Cell Qualities



First, we explore signal qualities associated with each single cell collected. (A) Image of a nucleated, unstained, epithelial cell, which is transferred using micromanipulation – i.e., pico-pipetting – to a 0.2 mL well of a 96-well plate. Also depicted are brightfield, PE and DAPI images of white blood cells taken by the DEPArray™ NxT. If cells were well separated from others, of the correct size, and color density, the cells were collected, extracted, and amplified using direct-to-PCR methods. Amplification and fragment analysis followed, with single cell electropherograms (scEPGs) being the result. Those scEPGs with steep changes to peak heights across molecular weights indicate the longer molecular weight fragments are not amplifying as well as the shorter ones, suggesting damaged DNA. The more severe the sloping the more negative β. (B) Scatterplots of the total intensity of a scEPG separated by cell-type. Also shown are the number of genotypes and cells represented in the test data. (C) Scatter plot depicting β-value versus the total scEPG intensity [RFU] for each scEPG, separated by cell type. (D) Histograms of the frequency of the proportion of alleles detected per-cell for heterozygous alleles across all scEPGs. (E) Frequency of allele detection by locus, ordered by color and size.

Clustering Results and LR(*C*,*s*)



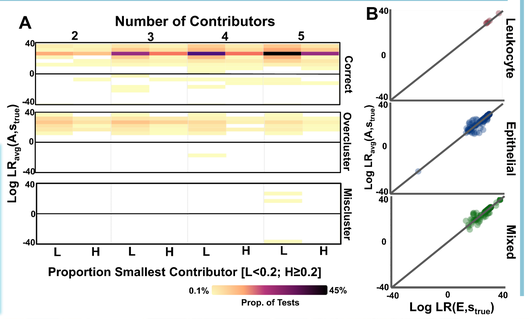
Summary of clustering results, and the likelihood ratio for each cluster, LR(*C*,*s*). (A) Stacked plots of the performance of model-based clustering (MBC) showing the proportion of admixtures resulting in correct, over clustered or misclustered outcomes, separated by mixture type and whether the smallest donor contributed < 20% [L] or ≥ 20% [H] of the cells to the admixture. The 'Mixed' type denotes mixtures of epithelial cells and leukocytes. (B) Heatmaps of $\log \mathrm{LR}(C, s_{\mathrm{true}})$ separated by mixture cell type and the proportion of the smallest contributor. Values above the zero axis are the number of $\log \mathrm{LR}(C, s_{\mathrm{true}}) > 0$, and those below it are the number of $\log \mathrm{LR}(C, s_{\mathrm{true}}) < 0$. We also tested each cluster of scEPGs against all other contributors to the admixtures. The 8,496 $\log \mathrm{LR}(C, s_{\mathrm{false}})$ were all <-28. (C) Histograms of the difference between true number of contributors (NoC) to the admixture and the number of clusters obtained by MBC, separated by the number of donors. On the top of the bar is shown the number of admixtures falling within that value. The value on the top right is the total number of admixtures. (D) Heatmaps of $\log \mathrm{LR}(C, s_{\mathrm{true}})$, separated by clustering outcome and the number of scEPGs in a cluster. The value in the bottom right is the total number of tests.

Evidential/Admixture Weights of Evidence, $LR_{avg}(A,s)$

Single cell weights of evidence across the entire admixture of cells, and the effects of over clustering on the average LR. (A) Heatmap of $\log \mathrm{LR}_{avg}(A, s_{\mathrm{true}})$, separated by clustering result, true NoC, and whether the proportion of the minor contributor constituted < 20% [L] or ≥20% [H] of the admixture. (B) Scatter plots of $\log \mathrm{LR}_{avg}(A, s_{\mathrm{true}})$ against $\log \mathrm{LR}(E, s_{\mathrm{true}})$, which is the weight of evidence based on ground-truth clustering, for admixtures where the number of MBC groups were greater than the true number of donors.

We discuss the informativeness of LR_{avg} by a thought experiment and contrast it with the procedure that only reports the largest LRs. Since $\log \mathrm{LR}_{\mathrm{avg}}(A,s) \cong \log \mathrm{LR}_{\mathrm{max}}(A,s) - \log n$, the difference between $\log \mathrm{LR}_{\mathrm{avg}}(A,s)$ and $\log \mathrm{LR}_{\mathrm{max}}(A,s)$ is small if n is small or $\mathrm{LR}_{\mathrm{max}}$ is large. However, let us consider the limit where n is so large that there is one cluster for every possible genotype in the population and the signal to noise of each scEPG is such that all alleles are detected and confounding signal from noise and stutter are absent. In this limit, there is one cluster corresponding to the PoI genotype, so $\mathrm{LR}_{\mathrm{max}}(A,s)$ is large but $\mathrm{LR}_{\mathrm{avg}}(A,s)$ is on the order of 1, which is the representative result since, in its totality, H_p is about as likely as H_d at this limit.

<u>Suspect-agnostic clustering and averaging LRs across clusters</u> **is an important step in single-cell interpretation**

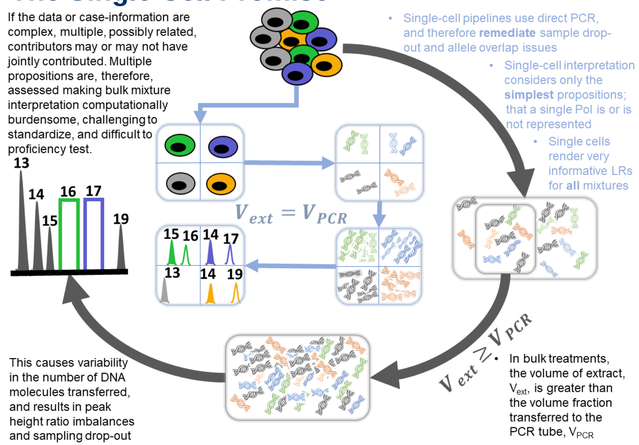


More Information

***Presenting author**: Catherine M. Grgicak
617.913.9728
grgicak@rutgers.edu

The Single Cell Promise

- If the data or case-information are complex, multiple, possibly related, contributors may or may not have jointly contributed. Multiple propositions are, therefore, assessed making bulk mixture interpretation computationally burdensome, challenging to standardize, and difficult to proficiency test.
- Single-cell pipelines use direct PCR, and therefore **remediate** sample drop-out and allele overlap issues
- Single-cell interpretation considers only the **simplest** propositions; that a single PoI is or is not represented
- Single cells render very informative LRs for **all** mixtures



- This causes variability in the number of DNA molecules transferred, and results in peak height ratio imbalances and sampling drop-out
- In bulk treatments, the volume of extract, V_{ext}, is greater than the volume fraction transferred to the PCR tube, V_{PCR}

Acknowledgments & Funding

- This work was performed by students and researchers at **Rutgers University Camden, Departments of Chemistry and Computer Science** and the **National University of Ireland Maynooth, Hamilton Institute**
- This work was supported by NIJ2020-R2-CX-0032 and NIJ2018-DU-BX-K0185 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not reflect those of the Departments of Justice
- US Patent Application for SYSTEMS AND METHODS FOR AUTOMATED ANALYSES OF A BIOLOGICAL SAMPLE (US Patent App. 17/669,790) supported by Rutgers University Innovation Ventures