Grouped and Ungrouped Single-Cell Electropherograms Enable Precision DNA Interpretation: Relevancy and Legitimacy of Single-Cell Forensics

LFTDI

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

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

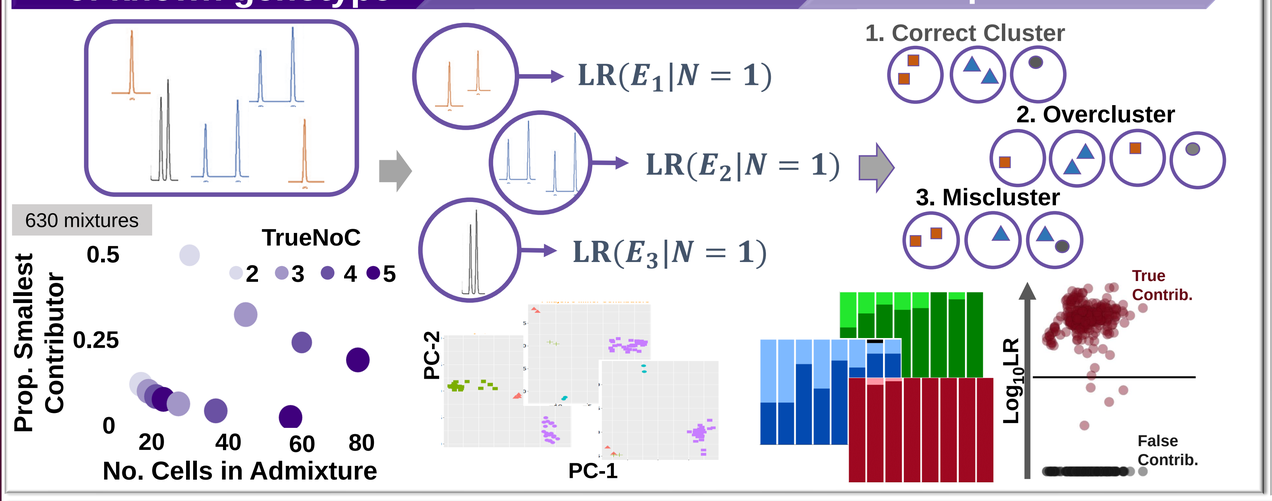
Rutgers University | Camden

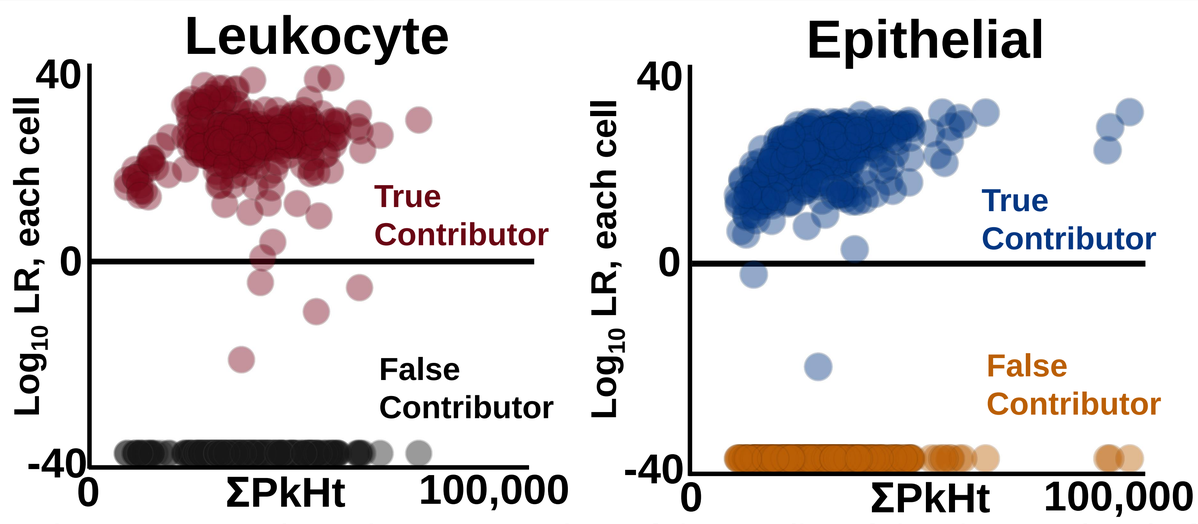
Forensic DNA pipelines are complex systems, where each step influences outcomes of subsequent steps

- The ***E*** in the likelihood ratio, $LR = \frac{\Pr(E|H_p,N,I)}{\Pr(E|H_d,N,I)}$, is influenced by processing, threshold and high-pass artifact filtering decisions
- The ***N*** indicates number of contributors (NOC), and is assigned
- The ***I*** is context, and is assigned

In Single-cell analysis *n* = 1 and *i*= I_{none}, simplifying interpretation

Simulate admixtures of known genotype → **Group scEPGs** → **Assess cluster and scLR performance**

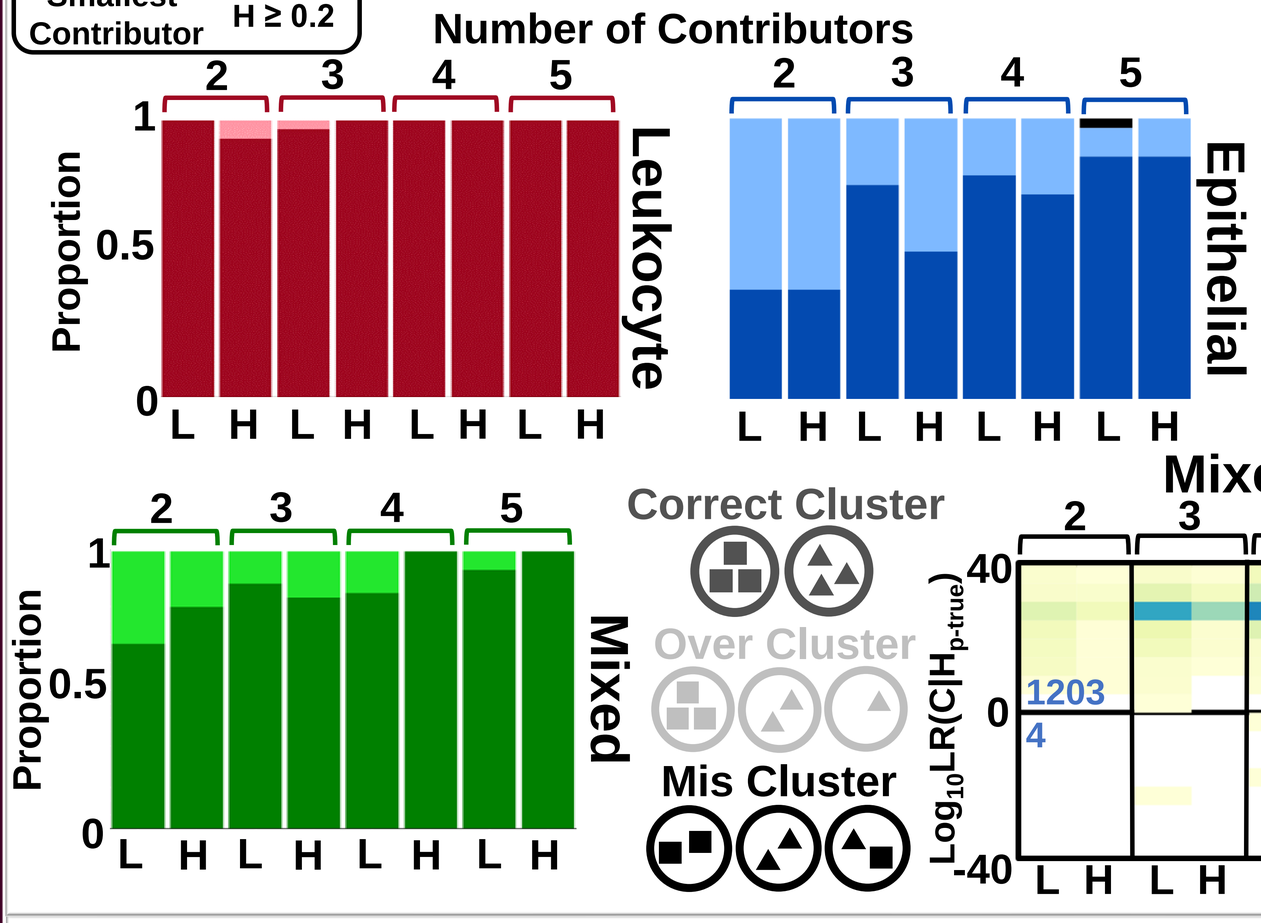


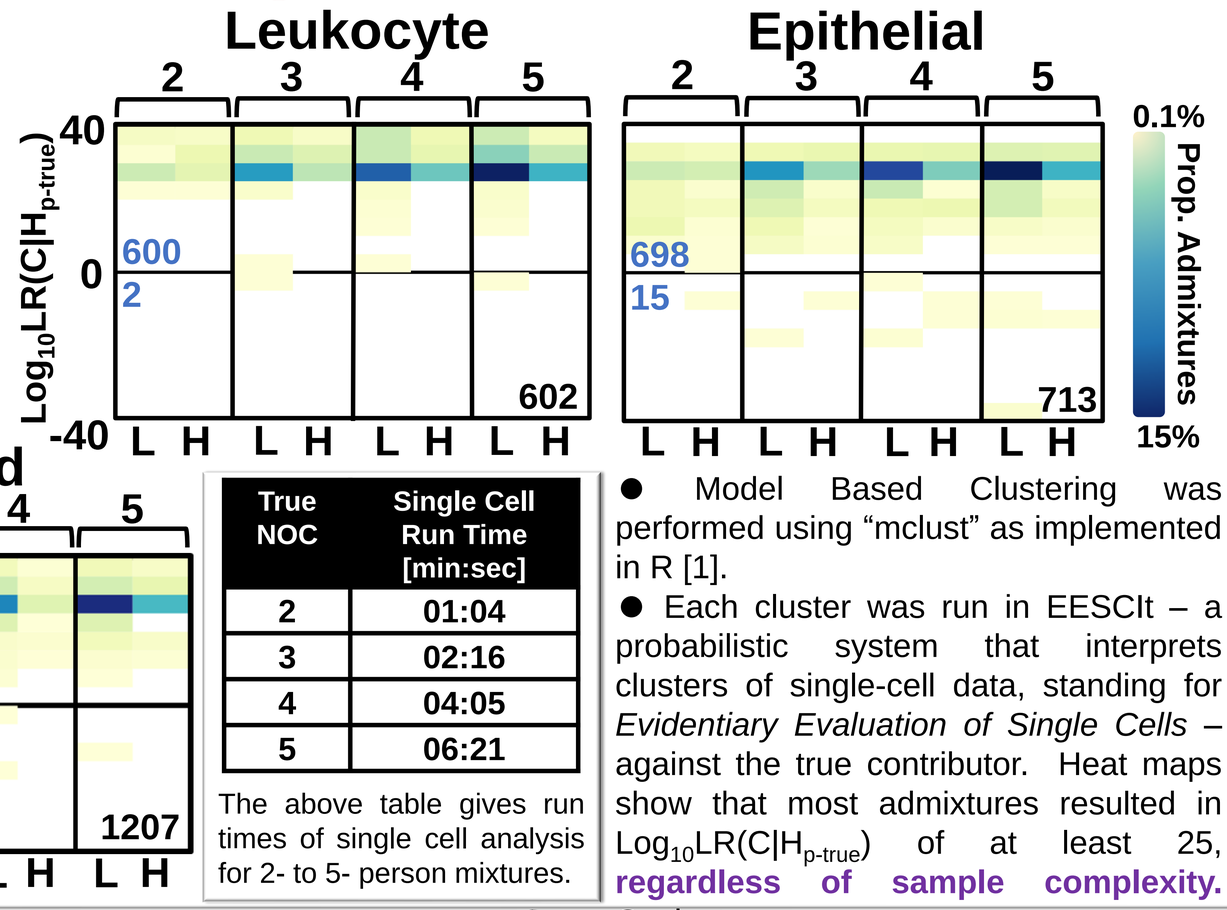


- These scatter plots give an overview of the quality of the data, or level of genetic information, inherent in each of the 643 single cell electropherograms used to construct 630 single-cell admixtures ranging from 2- to 5- contributors and containing up to 75 cells, where the proportion of the smallest contributor was as low as 3.5%.
- The scatter plots show $Log_{10}LR$ for each cell against the sum of the peak height (ΣPkHt) across the single cell electropherogram. When each cell was run against a true contributor, only 6 of 643 resulted in $Log_{10}LR < 0$. These were generally the result of overexpressed stutter. When each sample was run against a false contributor the $Log_{10}LR$ was always ≤ -40.
- Overall, scEPGs carry sufficient data for single-cell analysis.

Model Based Clustering

Proportion Smallest Contributor: L < 0.2, H ≥ 0.2





True NOC	Single Cell Run Time [min:sec]
2	01:04
3	02:16
4	04:05
5	06:21

The above table gives run times of single cell analysis for 2- to 5- person mixtures.

- Model Based Clustering was performed using "mclust" as implemented in R [1].
- Each cluster was run in EESCIt – a probabilistic system that interprets clusters of single-cell data, standing for *Evidentiary Evaluation of Single Cells* – against the true contributor. Heat maps show that most admixtures resulted in $Log_{10}LR(C|H_{p\text{-}true})$ of at least 25, **regardless of sample complexity.**

C=cluster

- The histograms show that most of the time there is correct clustering, and when overclustering occurs, it usually only renders one additional cluster.
- The heat maps show $Log_{10}LR(C|H_{p\text{-}true})$ for each cluster, C, plotted against the number of cells in a cluster, and separated by clustering outcome. **$Log_{10}LR(C|H_{p\text{-}true})$ remain constant regardless of whether there were more clusters than TrueNOC, or the number of scEPGs in a cluster.**

Conclusion

1. Model Based Clustering is a **credible** means by which to cluster scEPGs.
2. Single cell pipelines have forensic **relevance** since $Log_{10}(C|H_{p\text{-}true}) \geq 25$ for most clusters, and stays at that value regardless of admixture complexity.
3. $Log_{10}LR(C|H_{p\text{-}true})$ are robust to overclustering, showing the **legitimacy** of single cell genetics to the forensic domain.

References

[1] Scrucca, Luca et al. "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R journal* vol. 8,1 (2016): 289-317.

Collaborations

Maynooth University National University of Ireland Maynooth

Rutgers

Funding

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, U.S. 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.