Taming cell-to-cell heterogeneity in Acute Myeloid Leukaemia with machine learning

Yara E. Sánchez-Corrales[1*], Ruben V.C. Pohle[2], Sergi Castellano[1,3], Alice Giustacchini[2,*]

[1]Genetics and Genomic Medicine Department, Great Ormond Street Institute of Child Health, University College London, London, UK.
[2]Molecular and Cellular Immunology Section, Great Ormond Street Institute of Child Health, University College London, London, UK.
[3]UCL Genomics, Great Ormond Street Institute of Child Health, University College London, London, UK.
*** Correspondence:**
Corresponding Authors
y.sanchez-corrales@ucl.ac.uk
a.giustacchini@ucl.ac.uk

Keywords: AML, Machine learning, classification, clustering, leukaemia

Abstract

Acute Myeloid Leukaemia (AML) is a phenotypically and genetically heterogenous blood cancer characterised by very poor prognosis, with disease relapse being the primary cause of treatment failure.

AML heterogeneity arise from different genetic and non-genetic sources, including its proposed hierarchical structure, with leukemic stem cells (LSCs) and progenitors giving origin to a variety of more mature leukemic subsets. Recent advances in single-cell molecular and phenotypic profiling have highlighted the intra and inter-patient heterogeneous nature of AML, which has so far limited the success of cell-based immunotherapy approaches against single targets.

Machine Learning (ML) can be uniquely used to find non-trivial patterns from high-dimensional datasets and identify rare sub-populations. Here we review some recent ML tools that applied to single-cell data could help disentangle cell heterogeneity in AML by identifying distinct core molecular signatures of leukemic cell subsets. We discuss the advantages and limitations of unsupervised and supervised ML approaches to cluster and classify cell populations in AML, for the identification of biomarkers and the design of personalised therapies.

1 Introduction

AML is an aggressive and fast-progressing leukaemia characterised by the accumulation of myeloid progenitors (Tenen, 2003). Although most patients achieve remission after first line chemotherapy and

haematopoietic stem cell transplantation, about 40% later relapse (Tsirigotis et al., 2016). Long-term survival following relapse is below 20% with a median survival of 4-6 months, an outcome that has not improved over the last two decades with conventional approaches (Tsirigotis et al., 2016; Medeiros, 2018; Lonetti et al., 2019) and novel therapies are therefore urgently needed (Lonetti et al., 2019).

AML is a molecularly heterogeneous group of diseases with a complex mutational landscape, characterized by intra- and inter-patient variation (Figure 1A). Advances in next-generation sequencing and single-cell technologies have revealed that AML cells display genetic and epigenetic heterogeneity in different patients and even within the same patient multiple sub-clones co-exist, each carrying its own hierarchical structure and possessing distinct immunophenotypes (Miles et al., 2020).

A non-genetic source of heterogeneity in AML is its proposed hierarchical structure, mimicking the cellular hierarchy in normal hematopoietic development (Figure 1B). In healthy individuals, this involves a stepwise differentiation process, with hematopoietic stem cells (HSCs) giving rise to progressively more mature blood cells (Velten et al., 2017; Karamitros et al., 2018; Liggett & Sankaran, 2020). LSCs lie at the top of AML cellular hierarchies, and carry an unlimited ability to self-renew as well as giving origin to a variety of more mature leukemic subsets (Tenen, 2003), each expressing characteristic patterns of cell surface markers. LSCs can persist in a dormant state, making them selectively unresponsive to conventional chemotherapies and allowing them to eventually fuel disease relapse. For these reasons, the effective targeting of LSCs underpins any successful treatment for AML.

A promising approach is to target LSCs using immunotherapy with autologous T cells genetically redirected to express Chimeric Antigen Receptors (CARs). In fact, CAR-T cells can effectively target tumour cells irrespectively of their quiescent status. However, the lack of surface markers preferentially expressed on LSCs as opposed to healthy HSCs has hindered the development of cell-based immunotherapy strategies for AML, given the high risk of on-target off-tumour toxicity (Perna et al., 2017; Lamble & Tasian, 2019). In addition, some of the targets tested so far (e.g. CD33 or CD123) have heterogenous expression in the LSC compartment, with the risk of relapse due to their incomplete targeting (Mardiana & Gill, 2020). Upon relapse, genetic and immunophenotypic heterogeneity in AML LSCs further increases, complicating the discovery of ‘one fits all’ drug target (Ho et al., 2016).

As a result of AML’s heterogenous nature, CAR-T cell approaches against a single target are unlikely to be effective, thus the design of combinations of CAR-T cells against multiple targets requires a systematic characterization of the expression levels of surface antigens in AML cell populations at single-cell resolution (Figure 1C) (Perna et al., 2017).

The unprecedented resolution achieved with single-cell technologies has enabled the dissection of cell populations, including tumour and rare cell types that could not be identified using conventional bulk sequencing (Giustacchini et al., 2017; Aldridge & Teichmann, 2020). In AML, the quantitative phenotyping of leukemic cell profiles has allowed the identification of leukemic subsets without prior knowledge of phenotypic markers for their prospective isolation, opening up new analytical challenges

for their clinical interpretation (Van Galen et al., 2019; Petti et al., 2019; Miles et al., 2020; Wu et al., 2020; Velten et al., 2021; Triana et al., 2021).

Despite Machine Learning (ML) techniques having shown prognostic utility in classifying patients at high risk of relapse and having been applied to risk-adapted treatments (review by Eckardt et al. (2020)), they have only been recently applied to resolve heterogeneity in single-cell datasets from AML patients (Van Galen et al., 2019; Triana et al., 2021). Fortunately, there has been an explosion of new algorithms based on ML for the characterization of cell populations in single-cell datasets (Table 1) that could be applied to identify molecular markers specific to AML subpopulations.

Here, we review some recent state-of-the-art ML methods with the potential to shed light into cell heterogeneity in AML and identify biomarkers for specific cell populations in single-cell datasets. Benchmarking of some recent methods has been done by Abdelaal et al. (2019) and Zhao et al. (2020). Rather than an extensive discussion of algorithms, we provide a general overview of tools available to identify cell populations in single-cell studies, highlighting ones that have the potential to reveal new and rare cell types in AML and aid the design of personalised treatments.

2 Machine learning for cell type identification in single-cell datasets and biomarker discovery for personalized immunotherapy

Single-cell high-throughput techniques, such as scRNA-seq, quantitatively characterise **cell types** within a tissue (Trapnell, 2015). Typical workflows in single-cell transcriptional profiling include dimensionality reduction and clustering of cells based on their gene expression patterns followed by manual annotation of cell clusters from known cell type **markers** (Kolodziejczk et al., 2015). In the context of AML and other cancers, transcriptionally similar malignant cells are expected to group together, and can be unambiguously identified by the expression of certain feature genes that can be used as biomarkers for designing personalised treatments.

The identification of cell types using typical workflows has several drawbacks: first, rare cell types are easily missed and grouped together with some more prevalent ones; second, cell identity is often not discrete but lies in a continuum (for instance, cells with mixed identities or in transition); and third, the clustering can reflect other sources of variability unrelated to cell types (Kiselev et al., 2019). To address these issues, ML tools have recently been developed allowing quantitative identification and probabilistic assignment of cell types, thus aiding the identification of rare and heterogeneous cell populations.

In general, ML approaches are either **unsupervised** or **supervised** (Figure 1D). The main difference being the use of prior knowledge. Supervised methods are **trained** on an **annotated reference** with known **classes** of cell types, whereas unsupervised models identify patterns in the data without prior knowledge. A summary of recent methods is shown in Table 1.

2.1 Recent ML unsupervised methods

A common task for unsupervised methods is to use the intrinsic structure of the data to find clusters of cells. The advantage of these approaches is that cells can be grouped in an automatic and unbiased manner and thus, have the potential to discover unknown cell populations.

The popular single-cell processing packages Seurat (Butler et al., 2018) and Scanpy (Wolf et al., 2018) use a graph-based clustering approach combined with modularity optimization to group transcriptionally-similar cells together. Markers differentially expressed in each cluster can be found using different methods, including logistic regression. The cell identity of each cluster is assigned manually according to previous knowledge of cell-type specific markers. The main disadvantage of this approach is that the number of clusters depends on a resolution parameter assigned by the user (higher values will lead to a greater number of clusters) and thus, they may not faithfully reflect cell types.

The recently developed Single-Cell Clustering Assessment Framework (SCCAF) (Miao et al, 2020) generates an optimal number of clusters automatically. After the data has been clustered, SCCAF builds an ML classifier (logistic regression) using part of the data (training). By applying this model to the rest of the dataset (test), it iteratively merges clusters that appear indistinguishable to the ML classifier to produce the final optimum clustering. The output of the model is a weighted list of feature genes characteristic of every cluster that often include known markers for a given cell type and could potentially be used to detect common biomarkers of leukemic cell subsets from AML patients.

Another unsupervised method, single-cell consensus clustering (SC3) uses the first 4-7% * N (number of cells) **eigenvectors** to build multiple **k-means clustering** solutions (Kiselev et al., 2017). After hierarchical grouping, the final clustering is driven by the combination of multiple clustering solutions. The output is a list of marker genes that define each consensus cluster. While SC3 may not be the most sensitive method to find rare populations (such as LSCs), SC3 was successful in identifying clusters of prevalent genetic subclones with different mutations in myeloproliferative neoplasms (Kiselev et al., 2017). A disadvantage of this method is that it does not scale well for datasets with more than 5,000 cells (Andrews et al., 2021).

A recent unsupervised method, weighted-nearest neighbour (WNN), was used to cluster cells using multiple data modalities (e.g. surface proteins and transcriptomes) measured in the same cell (Hao et al., 2020). This method uses **k-nearest neighbours** (kNN) to learn cell-specific modality “weights”. When applied to a multiomics dataset generated from human bone marrow samples (Stuart et al., 2019), it showed that the combination of surface proteins and gene expression was superior for identifying cell populations than using one data modality alone. Multiomic single-cell technologies quantifying both surface proteins and transcriptomes of individual cells (e.g. CITE-seq), could be ideally applied to the identification of surface targets for the design of cell based immunotherapies (Stoeckius et al., 2017).

Other unsupervised methods rely on Non-negative matrix factorization (NMF) methods (Kotliar et al., 2019; Stein-O'Brien et al., 2019). These methods allow for the identification of cell types and, simultaneously, **cell states.** Given the great transcriptional heterogeneity seen in AML even within clonal populations carrying the same mutational patterns (Petti et al., 2019), it may be helpful to consider cell identities and activities separately when clustering leukemic populations. Moreover, NMF is potentially useful to identify LSC populations in AML, where the classical surface proteins defining primitive cell types are present in highly similar patterns to healthy HSCs, but a 'malignant stem-like' profile can still be identified (Levine et al., 2015).

2.2 Recent ML supervised methods

Supervised methods to classify cell types exploit previously identified cell types and use either known marker genes or annotated reference datasets as an input to probabilistically assign new cells to a given category.

Some methods take a list of markers for each cell type as input (Lee & Hemberg, 2019). For example, CellAssign (Zhang et al., 2019) uses predefined cell types input as a marker gene list to build a hierarchical model that produces a statistical classification of cells. This approach was used to delineate the composition of the tumour microenvironment in serial samples (treatment and relapse) from follicular lymphoma. Garnett (Pliner et al., 2019) also takes as input a list of markers. The format of the input list permits accounting for cellular hierarchy (i.e, cell subtypes) and can include positive and negative markers to define cell types (Pliner et al., 2019).

Other supervised methods use an annotated reference dataset to classify cell types but differ in the features and the ML methods used to train models (see Table 1). For instance, SingleCellNet (Tan et al., 2019) uses the most discriminative **gene pairs** (top pair transformation) to build a **random forest** classifier while methods such as scPred (Alquicira-Hernandez et al., 2019) and Moana (Wagner & Yanai, 2018) use principal components as features to fit a **support vector machine** (SVM). Some methods rely on one or several similarity metrics (such as SingleR Aran et al. (2019)) and **k-nearest neighbours** (kNN) to map query datasets into a known reference (e.g. scmap (Kiselev et al., 2018) and scClassify (Lin et al., 2020)). Other methods use the training dataset to build an **Artificial Neural Network** (ANN) model such as SuperCT (Xie et al., 2019) and ACTINN (Ma & Pellegrini, 2019) with an input layer containing as many nodes as the number of genes in the training set and an output layer with nodes equal to the number of cell types. Interestingly, both ANN methods provide pre-trained models that could be used to classify new AML datasets.

An advantage of supervised ML approaches is that cell types are assigned probabilistically and some approaches allow for the possibility of an "unassigned" category (Kiselev et al., 2018, Zhang et al., 2019, Tan et al., 2019, Pliner et al., 2019, Ma & Pellegrini, 2019). The unassigned label for cells that are absent or are very different in the reference dataset is key to limit misclassification and to allow the discovery of new cell types.

Algorithms such as CHETAH (de Kanter et al., 2019) and scClassify (Lin et al., 2020) allow for intermediate categories that can highlight populations with a mixture of identities as previously reported in AML (Smith et al., 1983). These methods are based on hierarchal correlation trees to classify test datasets (de Kanter et al., 2019, Lin et al., 2020).

As more annotated single-cell datasets become available, the primary advantage of supervised methods is leveraging previous knowledge. Reference datasets of human bone marrow cells from healthy individuals are available from resources such as the Human Cell Atlas (Regev et al., 2017). Distinct cell populations or patient-specific tumour clones could be identified as unknown (because they are very different or absent in the reference data sets). As AML single-cell datasets become more abundant, they can be integrated with healthy single or multimodal references using ML methods (Hao et al., 2020).

A disadvantage of supervised methods is that they rely on known markers or accurate cell type annotations to build classification models. Often, markers for rare cell populations, such as LSCs, are unknown, not robust (Pollyea & Jordan, 2017) or can be expressed by more than one cell type (Van Galen et al., 2019). Further, in many cases, annotation of single-cell datasets requires additional standardisation (de Kanter et al., 2019).

3 Discussion

ML techniques are able to find non-trivial patterns in high-dimensional data (Geron, 2019). In fact, ML has already proven useful in identifying markers in bulk studies in prospectively isolated leukemic sub-populations (Ng et al., 2016; Li et al., 2020). However, ML has not reached its full potential for the characterisation of AML cell populations at single-cell resolution, partly due to the recent development of large datasets (Van Galen et al., 2019; Petti et al., 2019; Miles et al., 2020; Triana et al., 2021; Velten et al., 2021).

Here we have reviewed tools to aid biomarker discovery using ML at single-cell level resolution. Many ML models explicitly quantify the contribution of individual features (genes) for a given classification. Importantly, genes identified in microarray data as important for classifying samples into “AML” or “no-AML” were not always differentially expressed (Warnat-Herresthal et al., 2020). This means that traditional differential expression analysis could fail to identify biomarkers that are good predictors for assigning a given group of cells (Alquicira-Hernandez et al., 2019). Thus, ML algorithms can find biomarkers that otherwise will be missed, expediting the design of suitable target combinations for immunotherapy.

Recently, it was shown that single-cell transcriptomics is capable of dissecting genetic subclones in AML, such as GATA2^{R361C}, which cluster separately from normal hematopoietic cell types (Petti et al., 2019). This observation suggests that subclonal diversity in AML could be associated with distinct gene expression profiles which ML techniques can leverage to identify mutated populations. Some AML mutations create subtle differences in expression profiles (Van Galen et al., 2019; Petti et al.,

2019; Velten et al., 2021) and isolating these populations represents an analytical challenge contemporary ML methods could address.

Moreover, recent experimental innovations allowing for the simultaneous quantitative assessment of cellular and molecular information at single-cell resolution promise to better dissect cell heterogeneity in AML. Particularly important is the ability to detect mutations in single cells combined with their transcriptional profiling, offering an unprecedent opportunity to identify specific leukemic cell populations (Giustacchini et al., 2017; Rodriguez-Meira et al., 2019; Van Galen et al., 2019; Petti et al., 2019; Ludwig et al., 2019; Velten et al., 2021). For instance, the combination of single-cell transcriptomics and mutational profiles allowed the distinction of pre-leukemic clones, LSC and healthy HSC (Velten et al., 2021). ML such as SVM could be used next to identify molecules that maximise this classification as done before for bulk RNA-seq and microarray data (Li et al., 2020).

In addition, the identification of mutant and non-mutant cells allows for applying ML methods to both all and only mutated cells to further characterise subpopulations (Petti et al., 2019), and can be used to fine-tune ML classification algorithms. For instance, a two-step ML classification strategy was applied to bone marrow samples of AML patients (Van Galen et al., 2019). First, a fraction of mutant cells was identified by genotyping and these were classified into one of six normal haematopoietic cell types (monocyte-like, progenitor-like, etc.). Subsequently, these malignant cell types were incorporated as additional classes in a second classifier that successfully identified mutant and normal cells from their transcriptome profiles.

The simultaneous characterization of surface proteins at single-cell resolution (Stoeckius et al., 2017) is especially important for isolation of heterogeneous cell populations. There are some analytical challenges with the integration of multiple data modalities (Efremova & Teichmann, 2020), but combining different data types from the same cell has already shown to improve cell population identification in AML datasets (Petti et al., 2019; Triana et al., 2021) and healthy bone marrow samples (Hao et al., 2020), thus we anticipate that multimodal datasets will improve the performance of ML models in isolating specific cell populations and may facilitate the identification of relevant surface targets for precision immunotherapy.

All the methods reviewed here will incur a certain degree of **underfitting** and **overfitting**. Thus, it is wise to compare algorithms in the initial cell composition assessment. Some, such as hierarchical methods, are potentially more suitable for AML samples, where there is an intrinsic hierarchy shared with normal hematopoietic development (Figure 1B). Also, methods that enable the recognition of intermediate cell types, mixed identities or different cell states would be more suitable for the identification of abnormally differentiated leukemic cells, known to be characteristic of AML (Smith et al., 1983).

Finally, we anticipate that single-cell resolution phenotyping will be important for the design of cell-based immunotherapy combinatorial strategies accounting for clonality and differentiation states of

AML populations, with ML likely playing a pivotal role in the selection of optimal therapeutic targets for the design of personalised workflows tailored to each patient.

4 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

5 Author Contributions

All authors contributed to the conception and editing of this review and approved the final manuscript. YSC conducted literature review and wrote the manuscript in consultation with RP. SG and AG critically revised the work.

6 Funding

This work was supported by the NIHR GOSH BRC.

7 Acknowledgments

This work was supported by the NIHR GOSH BRC, the views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. We thank George Hall for helpful feedback to the manuscript.

8 References

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., et al. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology, 20*(1). doi:10.1186/s13059-019-1795-z

Aldridge, S.,and Teichmann, S. A. (2020). Single cell transcriptomics comes of age. *Nature Communications, 11*(1). doi:10.1038/s41467-020-18158-5

Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q.,andPowell, J. E. (2019). scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biology, 20*(1). doi:10.1186/s13059-019-1862-5

Andrews, T. S., Kiselev, V. Y., McCarthy, D.,andHemberg, M. (2021). Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nature Protocols, 16*(1), 1-9. doi:10.1038/s41596-020-00409-w

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology, 20*(2), 163-172. doi:10.1038/s41590-018-0276-y

Butler, A., Hoffman, P., Smibert, P., Papalexi, E.,andSatija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology, 36*(5), 411-420. doi:10.1038/nbt.4096

de Kanter, J., Lijnzaad, P., Candelli, T., Margaritis, T.,andHolstege, F. C. P. (2019). CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Research, 47*(16), e95-e95. doi:10.1093/nar/gkz543

Eckardt, J.-N., Bornhäuser, M., Wendt, K.,andMiddeke, J. M. (2020). Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects. *Blood Advances, 4*(23), 6077-6085. doi:10.1182/bloodadvances.2020002997

Efremova, M.,and Teichmann, S. A. (2020). Computational methods for single-cell omics across modalities. *Nature Methods, 17*(1), 14-17. doi:10.1038/s41592-019-0692-4

Geron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (Second edition ed.): O'Reilly.

Giustacchini, A., Thongjuea, S., Barkas, N., Woll, P. S., Povinelli, B. J., Booth, C. A. G., et al. (2017). Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nature Medicine, 23*(6), 692-702. doi:10.1038/nm.4336

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., et al. (2020). *Integrated analysis of multimodal single-cell data*. Cold Spring Harbor Laboratory. Retrieved from https://dx.doi.org/10.1101/2020.10.12.335331

Ho, T.-C., Lamere, M., Stevens, B. M., Ashton, J. M., Myers, J. R., O'Dwyer, K. M., et al. (2016). Evolution of acute myelogenous leukemia stem cell properties after treatment and progression. *Blood, 128*(13), 1671-1678. doi:10.1182/blood-2016-02-695312

Karamitros, D., Stoilova, B., Aboukhalil, Z., Hamey, F., Reinisch, A., Samitsch, M., et al. (2018). Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. *Nature Immunology, 19*(1), 85-97. doi:10.1038/s41590-017-0001-2

Kiselev, V. Y., Andrews, T. S.,andHemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics, 20*(5), 273-282. doi:10.1038/s41576-018-0088-9

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods, 14*(5), 483-486. doi:10.1038/nmeth.4236

Kiselev, V. Y., Yiu, A.,andHemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. *Nature Methods, 15*(5), 359-362. doi:10.1038/nmeth.4644

Kolodziejczk, A. A., Kim, J. K., Svensson, V., Marioni, J. C.,andTeichmann, S. A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell, 58*(4), 610-620. doi:10.1016/j.molcel.2015.04.005

Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A., et al. (2019). Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife, 8*. doi:10.7554/elife.43803

Lamble, A. J.,and Tasian, S. K. (2019). Opportunities for immunotherapy in childhood acute myeloid leukemia. *Blood Advances, 3*(22), 3750-3758. doi:10.1182/bloodadvances.2019000357

Lee, J. T. H.,and Hemberg, M. (2019). Supervised clustering for single-cell analysis. *Nature Methods, 16*(10), 965-966. doi:10.1038/s41592-019-0534-4

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., et al. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell, 162*(1), 184-197. doi:10.1016/j.cell.2015.05.047

Li, J., Lu, L., Zhang, Y.-H., Xu, Y., Liu, M., Feng, K., et al. (2020). Identification of leukemia stem cell expression signatures through Monte Carlo feature selection strategy and support vector machine. *Cancer Gene Therapy, 27*(1-2), 56-69. doi:10.1038/s41417-019-0105-y

Liggett, L. A.,and Sankaran, V. G. (2020). Unraveling Hematopoiesis through the Lens of Genomics. *Cell, 182*(6), 1384-1400. doi:10.1016/j.cell.2020.08.030

Lin, Y., Cao, Y., Kim, H. J., Salim, A., Speed, T. P., Lin, D. M., et al. (2020). scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Molecular Systems Biology, 16*(6). doi:10.15252/msb.20199389

Lonetti, A., Pession, A.,andMasetti, R. (2019). Targeted Therapies for Pediatric AML: Gaps and Perspective. *Frontiers in Pediatrics, 7*. doi:10.3389/fped.2019.00463

Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., et al. (2019). Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell, 176*(6), 1325-1339.e1322. doi:10.1016/j.cell.2019.01.022

Ma, F.,and Pellegrini, M. (2019). ACTINN: automated identification of cell types in single cell RNA sequencing. *Bioinformatics*. doi:10.1093/bioinformatics/btz592

Mardiana, S.,and Gill, S. (2020). CAR T Cells for Acute Myeloid Leukemia: State of the Art and Future Directions. *Frontiers in Oncology, 10*. doi:10.3389/fonc.2020.00697

Medeiros, B. C. (2018). Is there a standard of care for relapsed AML? *Best Practice & Research Clinical Haematology, 31*(4), 384-386. doi:10.1016/j.beha.2018.09.006

Miao, Z., Moreno, P., Huang, N., Papatheodorou, I., Brazma, A.,andTeichmann, S. A. (2020). Putative cell type discovery from single-cell gene expression data. *Nature Methods, 17*(6), 621-628. doi:10.1038/s41592-020-0825-9

Miles, L. A., Bowman, R. L., Merlinsky, T. R., Csete, I. S., Ooi, A. T., Durruthy-Durruthy, R., et al. (2020). Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature, 587*(7834), 477-482. doi:10.1038/s41586-020-2864-x

Ng, S. W. K., Mitchell, A., Kennedy, J. A., Chen, W. C., McLeod, J., Ibrahimova, N., et al. (2016). A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature, 540*(7633), 433-437. doi:10.1038/nature20598

Perna, F., Berman, S. H., Soni, R. K., Mansilla-Soto, J., Eyquem, J., Hamieh, M., et al. (2017). Integrating Proteomics and Transcriptomics for Systematic Combinatorial Chimeric Antigen Receptor Therapy of AML. *Cancer Cell, 32*(4), 506-519.e505. doi:10.1016/j.ccell.2017.09.004

Petti, A. A., Williams, S. R., Miller, C. A., Fiddes, I. T., Srivatsan, S. N., Chen, D. Y., et al. (2019). A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nature Communications, 10*(1). doi:10.1038/s41467-019-11591-1

Pliner, H. A., Shendure, J.,andTrapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. *Nature Methods, 16*(10), 983-986. doi:10.1038/s41592-019-0535-3

Pollyea, D. A.,and Jordan, C. T. (2017). Therapeutic targeting of acute myeloid leukemia stem cells. *Blood, 129*(12), 1627-1635. doi:10.1182/blood-2016-10-696039

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., et al. (2017). The Human Cell Atlas. *eLife, 6*. doi:10.7554/elife.27041

Rodriguez-Meira, A., Buck, G., Clark, S.-A., Povinelli, B. J., Alcolea, V., Louka, E., et al. (2019). Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Molecular Cell, 73*(6), 1292-1305.e1298. doi:10.1016/j.molcel.2019.01.009

Smith, L., Curtis, J., Messner, H., Senn, J., Furthmayr, H.,andMcCulloch, E. (1983). Lineage infidelity in acute leukemia. *Blood*, 61 (66): 1138–1145. doi:https://doi.org/10.1182/blood.V61.6.1138.1138

Stein-O'Brien, G. L., Clark, B. S., Sherman, T., Zibetti, C., Hu, Q., Sealfon, R., et al. (2019). Decomposing Cell Identity for Transfer Learning across Cellular Measurements,

Platforms, Tissues, and Species. *Cell Systems, 8*(5), 395-411.e398. doi:10.1016/j.cels.2019.04.004

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., et al. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods, 14*(9), 865-868. doi:10.1038/nmeth.4380

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., et al. (2019). Comprehensive Integration of Single-Cell Data. *Cell, 177*(7), 1888-1902.e1821. doi:10.1016/j.cell.2019.05.031

Tan, Y.,and Cahan, P. (2019). SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Systems, 9*(2), 207-213.e202. doi:10.1016/j.cels.2019.06.004

Tenen, D. G. (2003). Disruption of differentiation in human cancer: AML shows the way. *Nature Reviews Cancer, 3*(2), 89-101. doi:10.1038/nrc989

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Research, 25*(10), 1491-1498. doi:10.1101/gr.190595.115

Triana, S. H., Vonficht, D., Jopp-Saile, L., Raffel, S., Lutz, R., Leonce, D., et al. (2021). *Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states*. Cold Spring Harbor Laboratory. Retrieved from https://dx.doi.org/10.1101/2021.03.18.435922

Tsirigotis, P., Byrne, M., Schmid, C., Baron, F., Ciceri, F., Esteve, J., et al. (2016). Relapse of AML after hematopoietic stem cell transplantation: methods of monitoring and preventive strategies. A review from the ALWP of the EBMT. *Bone Marrow Transplantation, 51*(11), 1431-1438. doi:10.1038/bmt.2016.167

Van Galen, P., Hovestadt, V., Wadsworth Ii, M. H., Hughes, T. K., Griffin, G. K., Battaglia, S., et al. (2019). Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell, 176*(6), 1265-1281.e1224. doi:10.1016/j.cell.2019.01.031

Velten, L., Haas, S. F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B. P., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nature Cell Biology, 19*(4), 271-281. doi:10.1038/ncb3493

Velten, L., Story, B. A., Hernández-Malmierca, P., Raffel, S., Leonce, D. R., Milbank, J., et al. (2021). Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics. *Nature Communications, 12*(1). doi:10.1038/s41467-021-21650-1

Wagner, F.,and Yanai, I. (2018). *Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data*. Cold Spring Harbor Laboratory. Retrieved from https://dx.doi.org/10.1101/456129

Warnat-Herresthal, S., Perrakis, K., Taschler, B., Becker, M., Baßler, K., Beyer, M., et al. (2020). Scalable Prediction of Acute Myeloid Leukemia Using High-Dimensional Machine Learning and Blood Transcriptomics. *iScience, 23*(1), 100780. doi:10.1016/j.isci.2019.100780

Wolf, F. A., Angerer, P.,andTheis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology, 19*(1). doi:10.1186/s13059-017-1382-0

Wu, J., Xiao, Y., Sun, J., Sun, H., Chen, H., Zhu, Y., et al. (2020). A single-cell survey of cellular hierarchy in acute myeloid leukemia. *Journal of Hematology & Oncology, 13*(1). doi:10.1186/s13045-020-00941-y

Xie, P., Gao, M., Wang, C., Zhang, J., Noel, P., Yang, C., et al. (2019). SuperCT: a supervised-learning framework for enhanced characterization of single-cell

transcriptomic profiles. *Nucleic Acids Research, 47*(8), e48-e48. doi:10.1093/nar/gkz116

Zhang, A. W., O'Flanagan, C., Chavez, E. A., Lim, J. L. P., Ceglia, N., McPherson, A., et al. (2019). Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nature Methods, 16*(10), 1007-1015. doi:10.1038/s41592-019-0529-1

Zhao, X., Wu, S., Fang, N., Sun, X.,andFan, J. (2020). Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. *Briefings in Bioinformatics, 21*(5), 1581-1595. doi:10.1093/bib/bbz096

9 Figure legend

Figure 1. The high cell-to-cell heterogeneity in AML tumours can be dissected using machine learning methods. A) The schematic representing clonal diversity in two putative AML patients highlights the complex intra and inter-patient variation of cell diversity (schematics adapted from Petti et al., 2019). Importantly, each clone carries its own hierarchical structure (here shown for one clone as an example). **B)** Leukemic populations share the hierarchical organization of normal hematopoietic development, where hematopoietic stem cells (HSCs) differentiate into multiple cell lineages, giving rise to all mature blood cells (blue lineages). Genetic mutations induce malignant transformation and give rise to leukemic stem cells (LSCs) that share some characteristics of their normal counterparts such as unlimited ability to self-renew and the potential to give origin to a variety of more mature leukemic subsets (red lineages). **C)** Ideal targets for immunotherapy with engineered T cells are those present in both leukemic blast and LSC cells and absent in healthy cell types. Targets that are ubiquitously expressed will fail to target specific leukemic populations and will be toxic for normal cells (on target off, tumour toxicity). Targets that are absent from LSC will render the treatment prone to relapse. Due to the high cell heterogeneity in AML more than one molecule is likely to fulfil these requirements. **D)** Machine learning methods to identify cell populations can be unsupervised and supervised. The former uses the intrinsic structure of the data to cluster cells in an automatic fashion. The second uses a predefined set of groups to classify unknown cells, leveraging previous knowledge.

10 Tables

Table 1. Summary of recent ML-based methods to identify cell types.

Algorithm name	Classification type	Method	Input data	Important contribution	Reference
SC3	Unsupervised	Consensus clustering and hierarchical clustering	Normalised expression matrix	Transcriptome-based identification of genetic subclones in myeloproliferative neoplasms	Kiselev et al. (2017)

cNMF	Unsupervised	Non-negative matrix factorization	Expression matrix and several parameters	Identification of previously misclassified immature skeletal muscle cells in a published dataset from brain organoids	Kotliar et al. (2019)
scCOGAPS	Unsupervised	Non-negative matrix factorization	Normalised and log-scaled expression matrix	Identification of gene expression signatures characteristic of discrete cell types in the developing retina	Stein-O'Brien et al. (2019)
SCCAF	Unsupervised	Logistic Regression and self-projection	Expression matrix and several parameters	Identification of cell states associated with different stages of erythroid maturation in mouse	Miao et al. (2020)
WWN	Unsupervised	K-nearest neighbours and Jaccard distance	Expression matrix and protein matrix (or any other single-cell measurement)	Single-cell multimodal analysis improves resolution of cell states in the immune system and identify previously unreported subpopulations	Hao et al. (2020)
CellAssign	Supervised	Expectation-Maximization hierarchical model	List of cell markers, subset of expression matrix containing the marker	Resolution of malignant and non-malignant cells and their molecular dynamics during disease progression	Zhang et al. (2019)

			genes and some parameters	in follicular lymphoma	
Garnett	Supervised	Multinomial elastic-net regression	Hierarchical list of cell markers (positive and negative) and expression matrix	The model trained on a mouse lung dataset is successfully applied to detect both healthy cell types and tumor cells in a human lung cancer dataset	Pliner et al. (2019)
scmap	Supervised	k-means (scmap-cluster) and k-nearest-neighbour (scmap-cell)	Annotated reference dataset and query expression matrix	Cell types in a test datasets are annotated with high accuracy irrespectively of batch effect	Kiselev et al. (2018)
CHETAH	Supervised	Hierarchical Spearman correlation	Annotated reference dataset and query expression matrix (both normalised and log – scaled)	The cell type identification algorithm correctly identifies cancer cells absent in the reference dataset as "unassigned" or "intermediate"	de Kanter et al. (2019)
scClassify	Supervised	Hierarchical ordered partitioning, ensemble learning and weighted k-nearest-neighbour	Annotated reference dataset and query expression matrix (both log – transformed)	Identification of cell types from the Tabula Muris single cell dataset that were unidentified in the original publication, including very rare populations	Lin et al. (2020)

SingleR	Supervised	Correlation to training set	Annotated reference dataset and query expression matrix (both normalised and log-transformed)	Identification of a subgroup of macrophages whose molecular markers are upregulated in samples from patients with idiopathic pulmonary fibrosis.	Aran et al. (2019)
SingleCellNet	Supervised	Random Forest	Annotated reference dataset and expression matrix (both raw)	Cells from pancreatic tissue that were “unclassified” in the original study are identified as Schwann cells and gamma cells	Tan, and Cahan (2019)
SuperCT	Supervised	Artificial Neural Network	Pre-trained ANN model and a query expression matrix	The model predicts cell types with high accuracy in multiple single cell test datasets including cord blood mononuclear cells and mouse pancreatic cancer.	Xie et al. (2019)
ACTINN	Supervised	Artificial Neural Network	Annotated reference dataset and query expression matrix	Model trained on a T cell subtype reference accurately predicts T cell subtypes from an independent peripheral blood mononuclear cells dataset	Ma, and Pellegrini (2019)

Moana	Supervised	Support Vector Machine	Pre-trained model and raw query expression matrix	Identification of common and cell type-specific gene expression responses to IFN-B treatment in peripheral blood cells	Wagner, and Yanai (2018)
scPred	Supervised	Support Vector Machine	Annotated reference dataset and query expression matrix (both normalised)	Prediction of pathological cell states in gastric and colorectal cancer	Alquicira-Hernandez et al. (2019)

11 Supplementary Material

A glossary is included as a Supplementary Material.