Title:

A Quantitative Single-Cell Proteomics Approach to Characterize an Acute Myeloid Leukemia Hierarchy

Running title:

Single Cell Proteomics in Leukemia

Author list
Erwin M. Schoof[1,2,3,4,5,6,#,*], Nicolas Rapin[1,2,4,*,@], Simonas Savickas[3], Coline Gentil[1,2], Eric Lechman[5,6], James Seymour Haile[1,2,4], Ulrich auf dem Keller[3], John E. Dick[5,6], Bo T. Porse[1,2,4,#]

Affiliations
[1] The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, University of Copenhagen, Denmark
[2] Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark
[3] Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby, Denmark
[4] Danish Stem Cell Centre (DanStem), University of Copenhagen, Denmark
[5] Princess Margaret Cancer Centre, University Health Network, Toronto, Canada
[6] Department of Molecular Genetics, University of Toronto, Toronto, Canada

[*] These authors contributed equally to this work

[#] Corresponding author:
Erwin M. Schoof, email: erwin.schoof@finsenlab.dk / erws@dtu.dk
Bo T. Porse, email: bo.porse@finsenlab.dk

[@] Current address: Deloitte Consulting, Weidekampsgade 6, 2300 København S, Denmark

Key words: acute myeloid leukemia / computational biology / proteomics / single-cell approaches / tandem mass tags (TMT)

Final character count: 47,782

Abstract

In recent years, cellular life science research has experienced a significant shift, moving away from conducting bulk cell interrogation towards single-cell analysis. It is only through single cell analysis that a complete understanding of cellular heterogeneity, and the interplay between various cell types that are fundamental to specific biological phenotypes, can be achieved. Single-cell assays at the protein level have been predominantly limited to targeted, antibody-based methods. However, here we present an experimental and computational pipeline, which establishes a comprehensive single-cell mass spectrometry-based proteomics workflow.

By exploiting a leukemia culture system, containing functionally-defined leukemic stem cells, progenitors and terminally differentiated blasts, we demonstrate that our workflow is able to explore the cellular heterogeneity within this aberrant developmental hierarchy. We show our approach is capable to quantifying hundreds of proteins across hundreds of single cells using limited instrument time. Furthermore, we developed a computational pipeline (SCeptre), that effectively clusters the data and permits the extraction of cell-specific proteins and functional pathways. This proof-of-concept work lays the foundation for future global single-cell proteomics studies.

Introduction

In recent years, single-cell molecular approaches such as RNAseq (sc-RNAseq) have revolutionized our understanding of molecular cell biology (Treutlein *et al*, 2014; Islam *et al*, 2014; Poulin *et al*, 2016; Trapnell, 2015; Paul *et al*, 2015). Gaining single-cell resolution of what occurs at the molecular level within single cells is of the utmost importance, particularly within in cancer biology, where it has long been known that tumors consist of a multitude of cell types, all acting in concert. (Levitin *et al*, 2018; Jerby-Arnon *et al*, 2018; Rodriguez-Meira *et al*, 2019; Nam *et al*, 2019; van Galen *et al*, 2019) Similarly, in complex biological organs such as the hematopoietic system, it is the complex interplay of various cell types and differentiation stages which defines a healthy or malignant state (Bonnet & Dick, 1997; Mercier & Scadden, 2015; Kreso & Dick, 2014; Bahr *et al*, 2018; Notta *et al*, 2016; Shlush *et al*, 2017; Lauridsen *et al*, 2018; Lapidot *et al*, 1994). Thus, technologies which are unable to resolve the molecular landscapes at single-cell resolution, like the more traditional Mass Spectrometry (MS) based proteomics approaches that typically require an input of hundreds of thousands of cells, are not sufficient to gain an understanding of cellular heterogeneity and of underlying signaling networks within individual cell types. Due to technical and practical limitations in terms of instrument sensitivity and experimental workflows thus far, single-cell MS (scMS) has been elusive, leading to an initial focus on sc-RNAseq based approaches. While RNA-based methods have been informative about the RNA landscapes in a plethora of biological systems and have demonstrated high clinical relevance(Eppert *et al*, 2011; Duployez *et al*, 2019; Ng *et al*, 2016), their accuracy has proven limitation when used as a proxy for protein levels (Vogel & Marcotte, 2012; Khan *et al*, 2013). Therefore, to gain a thorough understanding of what occurs in a cell at the protein level, on a global scale, MS-based approaches are the sole way to accomplish this. Being the cellular workhorses, there is much knowledge to be gained from mechanisms occurring at the protein level, either through enzyme activity, post-translational modifications or protein degradation/proteolysis; hence the great need for protein level approaches at the single-cell level.

A few years ago, a novel type of flow cytometry was established; by combining traditional flow cytometry workflows with mass spectrometry, a new analysis method termed Mass Cytometry was developed, more commonly referred to as CyTOF (Newell *et al*, 2012; Bodenmiller *et al*, 2012). This allows the simultaneous readout of tens of markers simultaneously, allowing single-cell analysis of pre-defined sets of proteins or post-translational modifications (PTMs). This method, however, relies heavily on previously

validated reagents and antibody panels, and is thereby inherently limited in terms of proteome coverage and which cellular signaling networks can be interrogated. Thus, while CyTOF represents a dramatic leap forward in terms of quantitative, protein-level analysis of single cells, there are a few inherent limitations that have prevented the technology from gaining universal recognition as a protein-level alternative to sc-RNAseq.

In this work, we have established a novel experimental workflow that allows the global characterization of single cell proteomes without the need for antibodies, and we conducted a proof-of-concept study in a primary Acute Myeloid Leukemia (AML) hierarchy, consisting of leukemic stem cells (LSC), progenitors and blasts (Lechman *et al*, 2016). This primary culture system, termed '8227', was derived from an AML sample where the patient had relapsed after treatment. By culturing under serum-free, growth-factor supplemented conditions, the hierarchical nature of AML is maintained, with LSC at the apex, differentiating to progenitors and terminally differentiated blasts; most traditional AML cell lines lack such a hierarchical structure. Blasts (characterized as CD34-) are the dominant population in the culture, but *in vivo* and *in vitro* long-term maintenance relies on the LSC (CD34+CD38-) and progenitor (CD34+CD38+) cells, thereby closely mimicking the in vivo hierarchical nature of primary AML (Fig. 1A). In order to successfully eradicate an AML in patients, we must not only target the blasts, but also the LSC in order to prevent relapse. However, due to their low abundance, studying LSC from a molecular perspective is challenging and has been limited to bulk-sorted approaches thus far (Raffel *et al*, 2017). The 8227 culture system provides us with an ideal proof-of-concept system, as the functional heterogeneity within this culture system has previously been evaluated and is readily isolated through FACS sorting (based on classical CD34/CD38 stem cell markers) (Lechman *et al*, 2016; Kaufmann *et al*, 2019). Modeling these functional differences using our molecular data would provide proof-of-principle that our workflow is able to distinguish differentiation stages in a complex cellular hierarchy.

Our approach builds on a series of recent developments in the low-input proteomics field (Kulak *et al*, 2014; Lechman *et al*, 2016; Schoof *et al*, 2016; Wojtowicz *et al*, 2016; Klimmeck *et al*, 2012; Cabezas-Wallscheid *et al*, 2014), and focuses on minimizing sample loss throughout the experimental protocol. Subsequently, by utilizing a 'booster' channel to provide additional peptide copies (and thus, ions for MS identification), combined with Tandem Mass Tag (TMT) labeling, we are able to derive quantitative information about

protein levels in 10 single cells per MS injection. A similar approach has recently been published (Budnik *et al*, 2018), also utilizing the TMT technology, but which does not employ FACS sorting to isolate the cells of interest, has not demonstrated the ability to differentiate between various cell types or differentiation stages originating from the same starting population, and has not applied state-of-the-art computational single-cell analysis. In order for scMS to be a viable alternative to sc-RNAseq, it needs to 1) be able to match the throughput capacity, 2) cover the same order of magnitude in terms of number of unique proteins detected and 3) be easily implementable in a wide range of cellular assays. Therefore, we opted to use a 96-well plate format, into which cells can be sorted by standard FACS sorting, thereby providing medium-high throughput, omitting the requirement for expensive consumables, and being amenable to automated liquid-handling systems. This type of experimental workflow puts our method in line with sc-RNAseq workflows in terms of throughput and ease of implementation. Maximizing proteome coverage was addressed by optimized experimental workflows and utilizing the latest generation of MS instruments. In order to be able to draw conclusions about single cell molecular phenotypes, it is imperative to be able to quantify hundreds of unique proteins, which was a vital criterium of our proposed workflow. Finally, in order to utilize the scMS data to its full potential, we adopted the latest algorithms from the sc-RNAseq field, and implemented them on our protein-level data. Thus, this work represents a proof-of-concept demonstration of the method and a data resource of a leukemia hierarchy, while additionally providing the community with a universally deployable software package that allows researchers from any field to analyze their own single-cell proteomics or other expression data.

Results

Challenges to overcome

One of the main technical challenges in conducting single-cell proteomics using MS, is the inherent sensitivity issue of MS instruments requiring enough ion (i.e. peptide) copies to successfully sequence, and thereby identify, the peptide. Depending on instrument type, this threshold rests anywhere between 10,000 – 100,000 ions and defines the lower limit of detection. To overcome this limitation, we utilized the main strength of TMT technology, namely the possibility to multiplex samples, while still being able to resolve sample-specific quantitative protein levels. Moreover, to boost the number of ions available for MS identification even further, we deploy a ‘booster’ sample, consisting of 500 cells, and dedicate a single TMT channel to that. The current iteration of the TMT technology has

eleven channels available, thus meaning that a single TMT sample can contain up to ten single cells plus the booster channel(s) (Fig. 1B). By multiplexing samples, the MS instrument has the ion equivalent of peptide copies from 510 cells available in total, leading to the successful identification and quantitation of a representative subset of the cellular proteome. For the experiments described below, we created single cell TMT pools for bulk cells (live cells covering the entire culture), blasts, progenitors and LSC, as we are able to FACS sort these populations according to the sorting scheme in Fig. 1B.

An additional challenge is the sample loss associated with the upstream experimental workflow, where proteins and peptides bind non-specifically to the plastic surfaces of the tubes they are contained in. When starting with extremely limited material such as single cells, minimizing these non-specific sample losses is of utmost importance. From our pilot experiments, we found that using Eppendorf LoBind technology was very effective at minimizing these losses and hence all our experiments are done in LoBind PCR plates and microtubes. To assist with minimizing sample loss, we follow and further adapted the iST approach (Kulak *et al*, 2014), by processing samples all in a single reaction chamber. Cells are FACS-sorted directly into a LoBind 96-well PCR plate, the lysis is done directly after sorting, and the solubilized proteins are digested and TMT-labelled in the same well, thereby minimizing transfer-associated sample loss. After acidification, the single cell peptide samples are pooled with their respective booster channel, and desalted using StageTip technology (Rappsilber *et al*, 2007).

The final and biggest technical challenge restricting single-cell proteomics is closely related with impediments in the sc-RNAseq field, namely the computational analysis and resolving the individual cell-types based on the available molecular data. To overcome this challenge, we adapted several of the latest state-of-the-art algorithms from the sc-RNAseq field, and tailored them to be amenable for our MS data (Fig. 1C). The resulting computational pipeline was termed 'SCeptre' (Single Cell proteomics readout of expression). Visualization of the data is facilitated through the use of SCANPY (Wolf *et al*, 2018), which allows one to adapt the embedding of choice (tSNE, UMAP, etc.) and explore the data visually. In order to account for experimental batch effects between sample injections and possible loss of protein groups between the same population, we rely on a proven technique from sc-RNAseq, where proteins most commonly expressed across all cells are used to compute an embedding of cells in order to see their relative positions in expression space. We use batch correction

(Haghverdi *et al*, 2018) on a reduced set of proteins commonly present in all populations (Step 1). The key parameter for defining this is "sigma", which controls the percentage of proteins that are allowed to be missing in a given population. This can be adjusted for each dataset independently, in order to explore the most appropriate threshold levels for individual experimental setups. Once a set of common proteins is defined (Step 2), a feature engineering step is applied in the form of augmenting the data of protein expression with biological pathways (Step 3). In this proof-of-concept work, it was decided to use wiki pathways (Slenter *et al*, 2018), but other sources of pathway information or gene signatures can be used as well, as we have implemented support for the generic Gene Matrix Transposed (.gmt) pathway format widely used in the field(Subramanian *et al*, 2005). Next, we compute a correlation network for all features (protein and pathway expression, step 4). Features are set as nodes in the network and two nodes are linked together when they achieve a correlation coefficient across all cells in the dataset above a specified threshold. The complements larger than four nodes are then used to compute an "Eigenprotein" value from all proteins and pathways within that component for all cells (Step 5). Finally, the Eigenproteins (or so-called 'cliques') are used to compute a UMAP (Uniform Manifold Approximation and Projection) embedding of all cells in the dataset (Step 6), which is the core of the visualizations shown in Figure 4. Combined, this workflow allows the determination of which proteins and pathways might play a role in defining the various cellular phenotypes. SCeptre is available as a docker image that contains all the processing steps and libraries readily available through an ipython notebook.

Designing an optimal experimental workflow

To determine the optimal experimental workflow, we explored various parameters to assess two key parameters for successful single-cell proteome analysis: 1) the choice of booster channel cells, and 2) instrument parameters. For the booster channel, in order to ensure the most accurate protein representation of the single cell proteomes within the booster channel, and exploit the nature of single-cell analyses to the fullest, it is imperative to choose the correct cell type, given the cell-specific protein expression levels that help distinguish cellular phenotype. To investigate the (dis)advantages of using various types of booster channel cells (bulk cells, differentiation stage specific and both), we repeated the experiments using the different types of booster channels and subsequently interpret the results. For the instrument parameters, we wished to investigate the pros and cons of conducting TMT quantitation at MS2 level or MS3 using the SPS TMT MS3 methodology (Ting *et al*, 2011; Hogrebe *et al*,

2018; McAlister *et al*, 2014); while the former method generally results in greater numbers of proteins identified, the quantitative accuracy tends to be affected by co-isolating peptides, which is resolved by the latter method, at the cost of sequencing speed, and consequently, number of protein identifications. Combined, these two types of investigations should provide us with a comprehensive catalogue on how to extract and utilize single cell proteomics data most effectively.

Effect of the choice of Booster Channel cells

To investigate the importance of the choice of booster cells, we set out to compare the results using 1) bulk, 2) differentiation stage specific, and 3) a combination of both bulk and differentiation stage specific booster channels (i.e. using two TMT booster channels). A brief overview of the result statistics can be found in Table 1, where it is clear that when using two types of booster channels simultaneously, the gain in extra ions for identification significantly increases the number of proteins identifiable in the single cells (2,138 proteins, compared to 1,259 proteins in the bulk cell booster channel only). Likely, this increase in number of proteins identifiable is also due to the fact that differentiation stage specific proteins (such as CD34, which has elevated expression levels in progenitors and LSC) can be captured with the cell-specific booster channels, which would be lost in the population average provided by the bulk cell boosters. Moreover, comparison of PCA clustering of the three types of booster channels reveals that distinguishing various differentiation stages and accurately clustering them together becomes hampered when not using cell-specific boosters (Fig. 2A). Indeed, only when a cell-specific booster channel is used, do the blasts separate clearly from the progenitors and LSC, which is a key indicator of the ability to determine cellular phenotype from the molecular protein-level data. However, standard PCA analysis often falls short for interpreting single cell molecular analyses comprehensively; hence, we processed the data using SCeptre, in order to determine whether the single cell proteomics data would enable us to differentiate between the three differentiation stages, and whether specific cell clusters could be identified within the respective cell populations. As portrayed in Figure 2B, it is evident that, in fact, SCeptre is able to distinguish the various differentiation stages, irrespective of booster channel used, and clusters together single cells of the same differentiation stage in all three datasets. Especially the UMAP embedding is successful at highlighting the correct cell clusters, although tSNE is also able to generally cluster according to differentiation stage. As expected, bulk cells are generally placed between the blasts and progenitors/LSC, given that they consist of all three differentiation stages. The fact

that they are generally most closely located to the blasts can be explained by the fact that ~90% of the 8227 cell culture system is made up of blast cells, which are therefore over-represented in the bulk cell population, thereby rendering them molecularly most similar to the blasts. When taken together, these results strongly suggest that our experimental workflow is able to capture enough proteome depth to decipher cellular phenotype at the single cell level, and that SCeptre is able to resolve these cell populations from the protein expression data alone.

Effect of MS Instrument Type

To exploit the latest capabilities in resolving TMT co-isolation effects, we utilized the ThermoFisher Orbitrap Fusion instrument to analyse a subset of samples (40 single cells for each of the three differentiation stages and bulk cells, i.e. 160 single cells total, combined with a differentiation stage specific booster channel of 500 cells) using the TMT SPS MS3 workflow(McAlister *et al*, 2014; Ting *et al*, 2011). We were interested in exploring whether the increased quantitative accuracy of such an instrument workflow would be beneficial for deciphering cellular phenotypes. Given the inherent lower proteome coverage with such an approach however, we also wanted to investigate whether the resulting proteome depth would still be sufficient for clustering the different differentiation stages correctly. When plotting this data using a standard PCA analysis (Fig. 2C), very little degree of separation between the different differentiation stages can be observed, and not as extensive as from our dataset using MS2-based quantitation. However, when analysing the same data using SCeptre (Fig. 2D), it becomes clear that the protein-level data was in fact sufficient to cluster the differentiation stages correctly, with the single cells generally clustering together by differentiation stage, especially in the UMAP-embedded visualization of the data.

Booster Channel control experiments

Encouraged by our initial results of being able to cluster differentiation stages correctly, irrespective of which booster channel we used, we next set out to explore whether mixing single cells of different differentiation stages, combined with a booster channel of one differentiation stage only, would still allow us to resolve differentiation stages correctly. We hypothesize that if we are truly reading out single cell proteomes, irrespective of type of booster channel used, the booster channel of one differentiation stage should not influence the final clustering of the single cells from other differentiation stages that were analysed in unison. To this end, we sorted two 96-well plates according to Figure 3A, where for each

differentiation stage, we prepared two TMT pools with that specific differentiation stage as booster channel, while mixing in single cells of all three differentiation stages and bulk within the same TMT pool. Thus, in total, 16 single cells of each differentiation stage were analysed, in combination with a booster channel originating from each of the differentiation stages. Resulting samples were then analysed using both MS2 and MS3-level quantitation, and we tested whether the single cells clustered according to differentiation stage. Total proteome coverage was around 1,000 proteins across 72 cells (Fig. 3B). When subjected to standard PCA analysis, there was no clustering according to differentiation stage (Fig. 3C); in fact, the single cells clustered almost exclusively according to the TMT pool they originated from. No significant difference was observed between MS2 and MS3-level quantitation, and even with the more accurate MS3 approach, no correct differentiation stage clustering was apparent. However, when analysed using SCeptre, single cells strikingly clustered perfectly according to differentiation stage, especially in the MS3 dataset (Fig. 3D). These results are highly critical, as they indicate we truly are measuring single cell proteomes as opposed to being significantly influenced by the booster channel contents; while differentiation stage specific booster samples are important for detecting cell-specific proteins, they are not imperative for being able to correctly cluster the differentiation stages using the protein-level information only. Moreover, these experiments also show that even using small cell numbers, with limited per-cell proteome coverage of several hundred proteins (Fig. 3D), we are still able to detect different differentiation stages and extract those proteins that may be key to their functional phenotypes.

Eigenprotein Analysis of scMS data

Having established that the molecular protein expression data from our experimental pipeline is of enough proteome depth and quantitative accuracy to correctly cluster differentiation stages, we next set out to improve the clustering ability even further, and try to decipher those proteins and functional pathways which may be fundamental to defining cellular phenotype. Being able to find cellular sub-clusters within a cell type of interest is of great benefit when trying to deduce cellular heterogeneity, as this may allow the determination of even purer cell populations within a previously deemed homogeneous cell type. To this end, we employed a concept from the RNAseq field, namely that of Eigenproteins. By combining protein expression patterns with protein interaction (i.e. pathway) information of those proteins, one can potentially derive functionally unique clusters within a cell type of interest, thereby spanning the bridge between protein expression and functional implications of those proteins.

This has been applied in the genomics field before (Han *et al*, 2017; Cheng *et al*, 2017; Agrahari *et al*, 2018), but to the best of our knowledge, has not been explored in the proteomics field thus far. By combining protein expression with protein interaction information, our pipeline generates so-called 'cliques', which are subsequently used to re-cluster the single cells on an embedding of choice (UMAP or tSNE).

To demonstrate the utility of the Eigenprotein approach, we focused on dataset number three (using differentiation stage specific booster channels), as it contains the largest number of cells and total number of proteins identified, and it should contain the differentiation stage specific proteins that would be lost when using bulk cells as booster channels. When supplementing the protein expression levels with protein interaction information, a different clustering patterns emerges (Fig. 4A), where cells of one differentiation stage are clustered in a more cloud-like fashion. In this analysis, a total of 54 cliques were identified (Supplementary Figure 1), which are sufficient and able to cluster the single cells according to differentiation stage. When looking at the UMAP representations of these Eigenprotein cliques (Fig. 4B), one can clearly distinguish differentiation stage specific clusters, suggesting that those cells are utilizing functional signalling pathways in a similar manner. These clusters can then be related back to the Eigenprotein differentiation stage clustering (Fig. 4A) to determine which specific differentiation stages are having those pathways up- or down-regulated. The pathways associated with the Eigenprotein cliques shown here (Eigenproteins 16, 26 and 22) correspond to "Wnt Signaling/Pluripotency", "VEGF/Fas ligand/p38_MAPK signaling", and "mir-124 predicted interactions with cell cycle and differentiation"; relating back to the differentiation stage clustering, it appears that these pathways are important for Progenitors/LSC, LSC and blasts respectively. For a complete overview of the Eigenprotein cliques identified in this dataset, they are all listed in Supplementary Table 1. This analysis workflow presents a meaningful way to explore the data, both visually and functionally, and to get insights into what functional pathways are active within the various differentiation stages, thereby providing input for downstream functional validation.

Towards finding cell-specific proteins

To further enhance the differentiation stage specific analysis, at the single-cell level, we wanted to be able to find those proteins whose expression levels most extensively define cellular phenotype. To this end, we deployed the "rank_genes" function from SCANPY (Wolf *et al*, 2018), which is integrated in SCeptre. This allows us to extract those proteins

that are ranked highest for defining a differentiation stage of interest, and plot them (Fig. 4C). For this illustration, the Wilcoxon-Rank-Sum statistical test was used. We then plotted the statistically significant proteins in a heatmap, illustrating the expression level of those proteins across the various cell populations, within each individual cell separately (Fig. 4D). In order to link these results back to our Eigenprotein embedding, the computational pipeline also enables the plotting of individual protein expression levels on any embedding of choice. To illustrate this more clearly, we plotted the top three proteins in the LSC population and the blast population in Figure 4E. From these results, we can find exactly those cells that show an increased or decreased expression level for the proteins of interest, and thereby interpret the statistical results in more detail. When looking at the targets highlighted by the Wilcoxon analysis, in the LSC population specifically, it is interesting to note that SWAP70 has previously been linked to AML development in a murine setting (Erkeland *et al*, 2004). Moreover, it is predicted to interact with NPM1 (Stelzer *et al*, 2016), which is frequently mutated in AML (McKerrell *et al*, 2015; Krönke *et al*, 2013). Combined with our single cell analysis potentially highlighting it as an LSC protein, seems to warrant future functional follow-up to investigate its exact role in disease development. The next target on the list, DDX46, has previously been shown to be required in hematopoietic stem cell activity in zebrafish (Hirabayashi *et al*, 2013). While it has not been shown in a leukemic context thus far, its role in hematopoietic stem cell differentiation, a process which has gone awry in AML, renders it a potentially highly interesting target. Together, these results indicate that the data is depicting several potentially relevant proteins for leukemia disease morphology, and that our experimental and computational pipeline is able to derive them effectively, from single cell proteomics data, something which is completely unprecedented.

Discussion

This work represents a proof-of-concept study, investigating whether current technology is able to conduct single-cell proteomics analysis on a bio-therapeutically relevant model system. By spending considerable effort not only on the sample preparation and data generation, but also on the subsequent data analysis efforts, we managed to establish a scMS workflow package which is able to 1) quantify thousands of unique proteins, 2) analyse hundreds of cells per day of instrument time, 3) visualize the data using the latest state-of-the-art single cell computational algorithms, and 4) derive differentiation stage specific proteins which may pose as potential therapeutic targets or other functionally relevant candidates.

As single-cell approaches put significant strain on throughput requirements, we focused especially on having an experimental workflow that would allow easy preparation of large cell numbers. As is the case for any single cell analysis, the more cells that can be analysed, the more knowledge can be extracted about the model system under investigation. By using standard FACS sorting methodology, even for very low frequency cell populations, we are able to sort several 96-well plates per hour, including the eight booster channel samples per plate, consisting of 500 cells each. Simultaneously, this allows for including index sorting in the experimental setup, thereby allowing a link to be drawn between fluorescent surface markers and expression levels of all detected proteins within the cells. Future experiments will focus on porting the workflow to a 384-well plate format, to further increase the throughput capacity, which will be very well matched with the newly released TMT 16-plex reagents. By simultaneously decreasing sample volumes, reaction kinetics should be improved, thereby potentially boosting proteome coverage as well. Similar approaches using nanofluidics have been shown very effective when analysing small sample amounts (Zhu *et al*, 2018), and it is likely that this platform would be beneficial to scMS as well; however, the expensive consumables associated with such an approach make large-scale investigations very costly, hence why 384-well plates may be an effective compromise.

We demonstrated the utility of the booster channel, and that when paired with appropriate computational workflows, we are able to correctly cluster cells irrespective of which type of booster cells were used. Nevertheless, if possible, we would recommend the use of a cell-specific booster channel, in order to detect the cell-specific proteins; in our case, we used CD34 as a trial candidate, as we know this should only be found at high abundance in the LSC/Progenitor cells. We were able to confirm this (Supplementary Figure 2), but more importantly, this protein was only detected when a cell-specific booster channel was used, thereby underlining the importance of using such a booster type. However, the fact that cells clustered correctly in all cases suggests that, e.g. in cases of very low abundant cell types, other cells could be used. This would be of great advantage in studies focused on very rare cell populations. With improvements at the MS instrument level, lower booster cell numbers (tens of cells rather than hundreds) would still be able to produce useful proteome depth, so in future studies, cell abundance should not be a limiting factor for scMS. Alternatively, it could be opted to deploy targeted peptide libraries as booster channel; by TMT-labeling those reference peptides, the MS instrument can be steered towards analysing a set of proteins of

interest, without needing additional cell numbers to provide the peptide copies. This will require further optimization, but in principle can be a powerful complementary booster method, especially in cases where cell-specific booster channels are difficult to generate, e.g. in the case of smaller multi-cellular organisms. Simultaneously, this can help ensure that proteins of interest will be quantified, compared to the partially stochastic nature of data-dependent analysis workflows. While it remains to be tested, peptide libraries could theoretically even open up the possibility of studying PTMs such as phosphorylation events, and could thus be employed to study common cell processes such as cell cycle, kinase activity etc.

Regarding the instrument parameters related to TMT quantitation, we did not observe a significant improvement when using MS3 level quantitation compared to MS2. This may be due to the fact that we analysed fewer cells, and due to slower cycle time, were able to quantify fewer proteins. It is likely that a newer generation of the ThermoFisher Tribrid instruments (such as the Orbitrap Eclipse) would offset this difference in number of proteins identified compared to MS2 quantitation. However, this lack of improvement, combined with the efficient clustering capabilities of MS2 level data, also suggests that the TMT co-isolation effect is not strong enough to negatively affect our readouts; nevertheless, alternative tagging technologies such as EASI-tags could be explored in future work to determine their compatibility with scMS (Winter *et al*, 2018). The suitability of TMT for this workflow is further supported by the fact that we can use the raw intensities as provided by Proteome Discoverer for our computational analyses, and are therefore not subject to normalization and correction effects that are sometimes opted to include when using TMT. This could potentially have implications for merging several experiments retrospectively, since using the raw intensities means that different datasets should be more compatible due to the lack of post-acquisition processing requirements.

By porting several of the latest state-of-the-art algorithms developed in the more established sc-RNAseq field, we are able to utilize the knowledge that has been gained over the past years to our advantage. One of the main strengths of our computational pipeline is to significantly enhance the ability to cluster cell types according to different data types (e.g expression levels or Eigenprotein cliques), the subsequent visualization thereof and extraction of highly relevant proteins. As the main embeddings (tSNE and UMAP) commonly used in single-cell approaches each have their own (dis)advantages (Zhu *et al*, 2018), we opted to

give the user the opportunity to plot both and choose according to their experimental setting. When comparing the clustering based on expression levels only with the clustering based on the Eigenprotein integration, it becomes very clear why single-cell analysis is so important when trying to understand cellular phenotype; while specific protein expression patterns may not always be consistent amongst all the single cells of a differentiation stage, the pathway activity for a particular pathway often does seem to span across a large subset of that differentiation stage (Supplementary Figure 1), indicating that functionally, they are displaying a similar phenotype that would have gone unnoticed when using expression data only. It simultaneously translates protein expression levels to a more functional interpretation thereof, which can be powerful when trying to functionally validate certain observations. Furthermore, it allows detection of functionally similar cells across different cell types, which can be informative of which cellular pathways are shared between, and which ones are unique to a certain differentiation stage. This may not have been picked up by protein expression levels alone.

In conclusion, this work presents the first time that a true, pure LSC proteome is published, while simultaneously being the first single-cell analysis of a leukemia hierarchy at the global proteome level. While it focuses on a single AML sample, it should nevertheless be a good foundation as a resource for follow-up studies, and paves the way for studying primary leukemias. Furthermore, by providing the community with the experimental protocols, combined with a powerful computational analysis pipeline, we strongly believe our scMS approach is now a real alternative to conducting sc-RNAseq analyses, especially in those cases where protein-level information is desirable. This opens up a plethora of research avenues, spanning across many biological fields, and proteome coverage will only improve as instrument sensitivity and experimental workflows develop even further, closing the gap between RNA-based and protein-based approaches.

Materials and Methods

Cell Culture and FACS Sorting

8227 cells were grown in StemSpan SFEM II media, supplemented with growth factors (Miltenyi Biotec, IL-3, IL-6 and G-CSF (10ng/mL), h-SCF and FLt3-L (50ng/mL), and TPO (25ng/mL) to support the hierarchical nature of the leukemia hierarchy captured within the cell culture system. On day 6, cells were harvested (8e6 cells total), washed, counted and resuspended in fresh StemSpan SFEM II media on ice at a cell density of 5e6 cells / ml. Staining was done for 30mins on ice, using a CD34 antibody (CD34-APC-Cy7, Biolegend, clone 581) at 1:100 (vol/vol) and CD38 antibody (CD38-PE, BD, clone HB7) at 1:50 (vol/vol). Cells were washed with extra StemSpan SFEM II media, and subsequently underwent three washes with ice cold PBS to remove any remaining growth factors or other contaminants from the growth media. Cells were resuspended for FACS sorting in fresh, ice cold PBS at 2e6 cells / ml. Cell sorting was done on a FACSAria I or III instrument, controlled by the DIVA software package and operating with a 100um nozzle. Cells were sorted at single-cell resolution, into a 96-well Eppendorf LoBind PCR plate (Eppendorf AG) containing 40ul of 50mM HEPES pH 8.5. In each row, wells 1-10 were filled with single cells, and well 11 was filled with 500 cells for the booster channel. Directly after sorting, plates were briefly spun and then boiled at 95C for 5mins, followed by sonication in a waterbath sonicator (VWR) for 2 mins to complete the lysis. Plates were then stored at -80C until further sample preparation.

Mass Spectrometry Sample Preparation

After thawing, the lysates were treated with Benzonase (Sigma cat. nr. E1014), diluted to 1:500 (vol/vol) for 1hr at 37C to digest any DNA that would interfere with downstream processing. Subsequently, 25ng of Trypsin (Sigma cat. nr. T6567) was added to the single cell samples, 50ng of Trypsin was added to the 500-cell booster channel samples, and the

plates were vortexed and kept at 37C overnight to complete the protein digestion. The next morning, peptides were labelled with TMT (tandem mass tag) reagents according to manufacturer's instructions. Briefly, 85mM of each label was added to the single-cell samples, while the 500-cell booster channel samples were labelled with 170mM of reagent. The labelling reaction was quenched with 2.5% Hydroxylamine for 15mins, after which peptides were acidified to a final concentration of 1% TFA, and TMT pools were mixed from 10 single cells + 1 booster channel to make up one 11-plex TMT sample each. The acidified TMT pools were subsequently desalted using in-house packed StageTips (Rappsilber *et al*, 2007). For each sample, 2 discs of C18 material (3M Empore) were packed in a 200ul tip, and the C18 material activated with 40ul of 100% Methanol (HPLC grade, Sigma), then 40ul of 80% Acetonitrile, 0.1% formic acid. The tips were subsequently equilibrated 2x with 40ul of 1%TFA, 3% Acetonitrile, after which the samples were loaded using centrifugation at 4,000x rpm. After washing the tips twice with 100ul of 0.1% formic acid, the peptides were eluted into clean 500ul Eppendorf tubes using 40% Acetonitrile, 0.1% formic acid. The eluted peptides were concentrated in an Eppendorf Speedvac, and re-constituted in 1% TFA, 2% Acetonitrile, containing iRT peptides (Biognosys AG, Switzerland) for Mass Spectrometry (MS) analysis.

Mass Spectrometry Data Collection

Peptides were loaded onto a 2cm C18 trap column (ThermoFisher 164705), connected in-line to a 50cm C18 reverse-phase analytical column (Thermo EasySpray ES803) using 100% Buffer A (0.1% Formic acid in water) at 750bar, using either the Thermo EasyLC 1200, or the ThermoFisher Ultimate 3000 UHPLC system, and the column oven operating at 45°C. Peptides were eluted over a 100 or 140 minute gradient, ranging from 6 to 60% of 80% acetonitrile, 0.1% formic acid at 250 nl/min.

For samples analysed using the MS2-level quantitation feature of TMT, the Q-Exactive HF-X instrument (ThermoFisher Scientific) was operated in DD-MS2 mode. The instrument was run with a top 16 method, collecting MS2 spectra at 45,000 resolution and a 1e5 AGC target. Ions were collected for 120ms, and isolated with an isolation width of 1.2 or 0.7 m/z. Precursors with a charge of 2-7 were included, and those that have been sequenced once were put on an exclusion list for up to 60 seconds.

For samples analysed using MS3-level SPS TMT quantitation (McAlister *et al*, 2014), the Orbitrap Fusion instrument (ThermoFisher Scientific) was operated in DD-MS3 mode. MS1 scans were collected at 120,000 resolution, scanning from 375-1500 m/z, collecting ions for 50ms or until the AGC target of 4e5 was reached. Precursors with a charge state of 2-7 were included for MS2 analysis, which were isolated with an isolation window of 0.7 m/z. Ions were collected for up to 50ms or until an AGC target value of 1e4 was reached, and fragmented using CID at 35% energy; these were then read out on the linear ion trap in rapid mode. Subsequently, up to 10 notches were selected for MS3 analysis, isolated with an m/z window of 2 m/z, and fragmented with HCD at 65% energy. Resulting fragments were read out in the Orbitrap at 50,000 resolution, with a maximum injection time of 105ms or until the AGC target value of 1e5 was reached.

Mass Spectrometry Raw Data Analysis

To translate .raw files into protein identifications and TMT reporter ion intensities, Proteome Discoverer 2.2 (ThermoFisher Scientific) was used with the built-in TMT Reporter ion quantification workflows. Default settings were applied, with Trypsin as enzyme specificity. Spectra were matched against the 9606 human database obtained from Uniprot. Dynamic modifications were set as Oxidation (M), and Acetyl on protein N-termini. Cysteine carbamidomethyl was set as a static modification, together with the TMT tag on both peptide N-termini and K residues. All results were filtered to a 1% FDR. The mass spectrometry data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository (Perez-Riverol *et al*, 2019) with the dataset identifier PXD015112. Reviewer account details: Username: reviewer87620@ebi.ac.uk, Password: 6BnVxQ9F.

Computational Analysis of Single Cell Data

The Proteome Discoverer output file is filtered for single cell data only (through removal of the booster channel samples), and the raw TMT reporter ion intensities are taken through a computational pipeline that aims at denoising the data and find meaningful cell clusters. The first step in the SCeptre pipeline is to find proteins that are in common with all the cell populations. We set a modular threshold s that controls the percentage of a given protein being present on average in each population. When the most common proteins are found, a non-parametric Bayesian batch correction method (Johnson *et al*, 2007) is used to account for the repeated injections over several TMT pools, for each of the cell populations. The final dataset of cells and batch corrected protein expression is then reported and taken forward for

further analysis. The next step consists of the computation of gene signature scores, for each individual cell. To this end, an average value of protein expression for each protein expressed in a given cell is reported. For the scope of this article, it was decided to use Wikipathways gene sets (Pico *et al*, 2008). The resulting augmented dataset with gene signature expression and protein expression is then subjected to a correlation analysis where the correlation coefficient of each cell against all other cells in the dataset is reported. This correlation matrix is then used to build a network of cells. Cells close to each other, e.g. with a correlation coefficient above a threshold of 0.6 are linked together. We then find sub-components (cliques) in the network (Cazals & Karande, 2008), which groups proteins and gene signatures together, resulting in the definition of an Eigenprotein. To compute the Eigenprotein score, a principal component analysis is run using the proteins and gene signatures in the Eigenprotein on all cells of the dataset. The Eigenprotein score is the first principal component (PC1). Using the Eigenprotein score, an embedding of the cells in the dataset is created and used to report cell proximity in Eigenprotein space. The pipeline and all source code in Python is available as an iPython notebook run from a docker container (kuikuisven/sceptre).

Acknowledgements

Erwin M. Schoof is a Lundbeck Fellow (fellowship # 2017-389) and a former EMBO Fellow (ALTF 1595–2014) and is co-funded by the European Commission (LTFCOFUND2013, GA-2013-609409) and Marie Curie Actions. This work was supported by the Kirsten & Freddy Johansen Foundation, the Independent Research Fund Denmark and through a centre grant from the Novo Nordisk Foundation (Novo Nordisk Foundation Centre for Stem Cell Biology, DanStem; Grant Number NNF17CC0027852). U. auf dem Keller acknowledges support by a Novo Nordisk Foundation Young Investigator Award (Grant Number NNF16OC0020670). Work in the Dick lab was supported by funds from the: Kirsten & Freddy Johansen International Prize, Princess Margaret Cancer Centre Foundation, Ontario Institute for Cancer Research with funding from the Province of Ontario, Canadian Institutes for Health Research, Canadian Cancer Society Research Institute, Terry Fox Foundation, Genome Canada through the Ontario Genomics Institute, and a Canada Research Chair.

Author contributions

E.M. Schoof designed and carried out the experiments, wrote the manuscript and conducted data analysis. N. Rapin designed and implemented the computational workflow, conducted

data analysis and wrote the manuscript. E. Lechman, C. Gentile and J.S. Haile contributed with cellular assays. S. Savickas and U. auf dem Keller assisted with the MS analysis, helped develop the instrument methods and contributed to the overall functioning of the method. All authors proof-read and contributed to the manuscript. U. auf dem Keller, J.E. Dick and B.T. Porse oversaw the study, wrote the manuscript and helped design the experiments.

Conflict of interest

The authors declare no conflict of interest

References

Agrahari R, Foroushani A, Docking TR, Chang L, Duns G, Hudoba M, Karsan A & Zare H (2018) Applications of Bayesian network models in predicting types of hematological malignancies. *Sci. Rep.* **8:** 6951 Available at: http://www.nature.com/articles/s41598-018-24758-5

Bahr C, von Paleske L, Uslu V V., Remeseiro S, Takayama N, Ng SW, Murison A, Langenfeld K, Petretich M, Scognamiglio R, Zeisberger P, Benk AS, Amit I, Zandstra PW, Lupien M, Dick JE, Trumpp A & Spitz F (2018) A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* **553:** 515–520 Available at: http://www.nature.com/articles/nature25193

Bodenmiller B, Zunder ER, Finck R, Chen TJ, Savig ES, Bruggner R V, Simonds EF, Bendall SC, Sachs K, Krutzik PO & Nolan GP (2012) Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nat. Biotechnol.* **30:** 858–867 Available at: http://www.ncbi.nlm.nih.gov/pubmed/22902532

Bonnet D & Dick JE (1997) Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* **3:** 730–7 Available at: http://www.ncbi.nlm.nih.gov/pubmed/9212098

Budnik B, Levy E, Harmange G & Slavov N (2018) SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* **19:** 161 Available at: http://www.ncbi.nlm.nih.gov/pubmed/30343672

Cabezas-Wallscheid N, Klimmeck D, Hansson J, Lipka DB, Reyes A, Wang Q, Weichenhan D, Lier A, von Paleske L, Renders S, Wünsche P, Zeisberger P, Brocks D, Gu L,

Herrmann C, Haas S, Essers MAG, Brors B, Eils R, Huber W, et al (2014) Identification of Regulatory Networks in HSCs and Their Immediate Progeny via Integrated Proteome, Transcriptome, and DNA Methylome Analysis. *Cell Stem Cell* **15:** 507–522 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25158935

Cazals F & Karande C (2008) A note on the problem of reporting maximal cliques. *Theor. Comput. Sci.* **407:** 564–568 Available at: https://www.sciencedirect.com/science/article/pii/S0304397508003903

Cheng J, Zhang J, Han Y, Wang X, Ye X, Meng Y, Parwani A, Han Z, Feng Q & Huang K (2017) Integrative Analysis of Histopathological Images and Genomic Data Predicts Clear Cell Renal Cell Carcinoma Prognosis. *Cancer Res.* **77:** e91–e100 Available at: http://www.ncbi.nlm.nih.gov/pubmed/29092949

Duployez N, Marceau-Renaut A, Villenet C, Petit A, Rousseau A, Ng SWK, Paquet A, Gonzales F, Barthélémy A, Leprêtre F, Pottier N, Nelken B, Michel G, Baruchel A, Bertrand Y, Leverger G, Lapillonne H, Figeac M, Dick JE, Wang JCY, et al (2019) The stem cell-associated gene expression signature allows risk stratification in pediatric acute myeloid leukemia. *Leukemia* **33:** 348–357 Available at: http://www.nature.com/articles/s41375-018-0227-5

Eppert K, Takenaka K, Lechman ER, Waldron L, Nilsson B, van Galen P, Metzeler KH, Poeppl A, Ling V, Beyene J, Canty AJ, Danska JS, Bohlander SK, Buske C, Minden MD, Golub TR, Jurisica I, Ebert BL & Dick JE (2011) Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat. Med.* **17:** 1086–1093 Available at: http://www.nature.com/articles/nm.2415

Erkeland SJ, Valkhof M, Heijmans-Antonissen C, van Hoven-Beijen A, Delwel R, Hermans MHA & Touw IP (2004) Large-scale identification of disease genes involved in acute myeloid leukemia. *J. Virol.* **78:** 1971–80 Available at: http://www.ncbi.nlm.nih.gov/pubmed/14747562

van Galen P, Hovestadt V, Wadsworth Ii MH, Hughes TK, Griffin GK, Battaglia S, Verga JA, Stephansky J, Pastika TJ, Lombardi Story J, Pinkus GS, Pozdnyakova O, Galinsky I, Stone RM, Graubert TA, Shalek AK, Aster JC, Lane AA & Bernstein BE (2019) Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176:** 1265-1281.e24 Available at: https://linkinghub.elsevier.com/retrieve/pii/S0092867419300947

Haghverdi L, Lun ATL, Morgan MD & Marioni JC (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*

36: 421–427 Available at: http://www.nature.com/articles/nbt.4091

Han Z, Johnson T, Zhang J, Zhang X & Huang K (2017) Functional Virtual Flow Cytometry: A Visual Analytic Approach for Characterizing Single-Cell Gene Expression Patterns. *Biomed Res. Int.* **2017:** 1–9 Available at: https://www.hindawi.com/journals/bmri/2017/3035481/

Hirabayashi R, Hozumi S, Higashijima S-I & Kikuchi Y (2013) Ddx46 is required for multi-lineage differentiation of hematopoietic stem cells in zebrafish. *Stem Cells Dev.* **22:** 2532–42 Available at: http://www.ncbi.nlm.nih.gov/pubmed/23635340

Hogrebe A, von Stechow L, Bekker-Jensen DB, Weinert BT, Kelstrup CD & Olsen J V. (2018) Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nat. Commun.* **9:** 1045 Available at: http://www.nature.com/articles/s41467-018-03309-6

Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P & Linnarsson S (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11:** 163–166 Available at: http://www.nature.com/articles/nmeth.2772

Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su M-J, Melms JC, Leeson R, Kanodia A, Mei S, Lin J-R, Wang S, Rabasha B, Liu D, Zhang G, Margolais C, Ashenberg O, Ott PA, Buchbinder EI, Haq R, Hodi FS, et al (2018) A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell* **175:** 984-997.e24 Available at: https://www.sciencedirect.com/science/article/pii/S0092867418311784?via%3Dihub

Johnson WE, Li C & Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8:** 118–127 Available at: http://www.ncbi.nlm.nih.gov/pubmed/16632515

Kaufmann KB, Garcia-Prat L, Liu Q, Ng SWK, Takayanagi S-I, Mitchell A, Wienholds E, van Galen P, Cumbaa CA, Tsay MJ, Pastrello C, Wagenblast E, Krivdova G, Minden MD, Lechman ER, Zandi S, Jurisica I, Wang JCY, Xie SZ & Dick JE (2019) A stemness screen reveals *C3orf54/INKA1* as a promoter of human leukemia stem cell latency. *Blood* **133:** 2198–2211 Available at: http://www.ncbi.nlm.nih.gov/pubmed/30796022

Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK & Gilad Y (2013) Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342:** 1100–4 Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1242379

Klimmeck D, Hansson J, Raffel S, Vakhrushev SY, Trumpp A & Krijgsveld J (2012) Proteomic Cornerstones of Hematopoietic Stem Cell Differentiation: Distinct Signatures of Multipotent Progenitors and Myeloid Committed Cells. *Mol. Cell. Proteomics* **11:** 286–302 Available at: http://www.ncbi.nlm.nih.gov/pubmed/22454540

Kreso A & Dick JE (2014) Evolution of the Cancer Stem Cell Model. *Cell Stem Cell* **14:** 275–291 Available at: https://www.sciencedirect.com/science/article/pii/S1934590914000575?via%3Dihub

Krönke J, Bullinger L, Teleanu V, Tschürtz F, Gaidzik VI, Kühn MWM, Rücker FG, Holzmann K, Paschka P, Kapp-Schwörer S, Späth D, Kindler T, Schittenhelm M, Krauter J, Ganser A, Göhring G, Schlegelberger B, Schlenk RF, Döhner H & Döhner K (2013) Clonal evolution in relapsed NPM1-mutated acute myeloid leukemia. *Blood* **122:** 100–8 Available at: http://www.ncbi.nlm.nih.gov/pubmed/23704090

Kulak NA, Pichler G, Paron I, Nagaraj N & Mann M (2014) Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11:** 319–324 Available at: http://www.nature.com/articles/nmeth.2834

Lapidot T, Sirard C, Vormoor J, Murdoch B, Hoang T, Caceres-Cortes J, Minden M, Paterson B, Caligiuri MA & Dick JE (1994) A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367:** 645–648 Available at: http://www.ncbi.nlm.nih.gov/pubmed/7509044

Lauridsen FKB, Jensen TL, Rapin N, Aslan D, Wilhelmson AS, Pundhir S, Rehn M, Paul F, Giladi A, Hasemann MS, Serup P, Amit I & Porse BT (2018) Differences in Cell Cycle Status Underlie Transcriptional Heterogeneity in the HSC Compartment. *Cell Rep.* **24:** 766–780 Available at: http://www.ncbi.nlm.nih.gov/pubmed/30021172

Lechman ER, Gentner B, Ng SWK, Schoof EM, van Galen P, Kennedy JA, Nucera S, Ciceri F, Kaufmann KB, Takayama N, Dobson SM, Trotman-Grant A, Krivdova G, Elzinga J, Mitchell A, Nilsson B, Hermans KG, Eppert K, Marke R, Isserlin R, et al (2016) miR-126 Regulates Distinct Self-Renewal Outcomes in Normal and Malignant Hematopoietic Stem Cells. *Cancer Cell* **29:** 214–28 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26832662

Levitin HM, Yuan J & Sims PA (2018) Single-Cell Transcriptomic Analysis of Tumor Heterogeneity. *Trends in Cancer* **4:** 264–268 Available at: https://www.sciencedirect.com/science/article/abs/pii/S2405803318300384?via%3Dihub

McAlister GC, Nusinow DP, Jedrychowski MP, Wühr M, Huttlin EL, Erickson BK, Rad R,

Haas W & Gygi SP (2014) MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes. *Anal. Chem.* **86:** 7150–7158 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24927332

McKerrell T, Park N, Moreno T, Grove CS, Ponstingl H, Stephens J, Crawley C, Craig J, Scott MA, Hodkinson C, Baxter J, Rad R, Forsyth DR, Quail MA, Zeggini E, Ouwehand W, Varela I & Vassiliou GS (2015) Leukemia-Associated Somatic Mutations Drive Distinct Patterns of Age-Related Clonal Hemopoiesis. *Cell Rep.* **10:** 1239–1245 Available at: https://www.sciencedirect.com/science/article/pii/S2211124715001138?via%3Dihub

Mercier FE & Scadden DT (2015) Not All Created Equal: Lineage Hard-Wiring in the Production of Blood. *Cell* **163:** 1568–1570 Available at: https://www.sciencedirect.com/science/article/pii/S0092867415016372?via%3Dihub

Nam AS, Kim K-T, Chaligne R, Izzo F, Ang C, Taylor J, Myers RM, Abu-Zeinah G, Brand R, Omans ND, Alonso A, Sheridan C, Mariani M, Dai X, Harrington E, Pastore A, Cubillos-Ruiz JR, Tam W, Hoffman R, Rabadan R, et al (2019) Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature* **571:** 355–360 Available at: http://www.nature.com/articles/s41586-019-1367-0

Newell EW, Sigal N, Bendall SC, Nolan GP & Davis MM (2012) Cytometry by Time-of-Flight Shows Combinatorial Cytokine Expression and Virus-Specific Cell Niches within a Continuum of CD8+ T Cell Phenotypes. *Immunity* **36:** 142–152 Available at: http://www.ncbi.nlm.nih.gov/pubmed/22265676

Ng SWK, Mitchell A, Kennedy JA, Chen WC, McLeod J, Ibrahimova N, Arruda A, Popescu A, Gupta V, Schimmer AD, Schuh AC, Yee KW, Bullinger L, Herold T, Görlich D, Büchner T, Hiddemann W, Berdel WE, Wörmann B, Cheok M, et al (2016) A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature* **540:** 433–437 Available at: http://www.nature.com/articles/nature20598

Notta F, Zandi S, Takayama N, Dobson S, Gan OI, Wilson G, Kaufmann KB, McLeod J, Laurenti E, Dunant CF, McPherson JD, Stein LD, Dror Y & Dick JE (2016) Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351:** aab2116 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26541609

Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, Winter D, Lara-Astiaso D, Gury M, Weiner A, David E, Cohen N, Lauridsen FKB, Haas S, Schlitzer A, Mildner A, Ginhoux F, Jung S, Trumpp A, Porse BT, et al (2015) Transcriptional Heterogeneity

and Lineage Commitment in Myeloid Progenitors. *Cell* **163:** 1663–77 Available at: https://linkinghub.elsevier.com/retrieve/pii/S0092867415014932

Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, Pérez E, Uszkoreit J, Pfeuffer J, Sachsenberg T, Yilmaz S, Tiwary S, Cox J, Audain E, Walzer M, Jarnuczak AF, et al (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47:** D442–D450 Available at: https://academic.oup.com/nar/article/47/D1/D442/5160986

Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR & Evelo C (2008) WikiPathways: Pathway Editing for the People. *PLoS Biol.* **6:** e184 Available at: http://www.ncbi.nlm.nih.gov/pubmed/18651794

Poulin J-F, Tasic B, Hjerling-Leffler J, Trimarchi JM & Awatramani R (2016) Disentangling neural cell diversity using single-cell transcriptomics. *Nat. Neurosci.* **19:** 1131–1141 Available at: http://www.nature.com/articles/nn.4366

Raffel S, Falcone M, Kneisel N, Hansson J, Wang W, Lutz C, Bullinger L, Poschet G, Nonnenmacher Y, Barnert A, Bahr C, Zeisberger P, Przybylla A, Sohn M, Tönjes M, Erez A, Adler L, Jensen P, Scholl C, Fröhling S, et al (2017) BCAT1 restricts αKG levels in AML stem cells leading to IDHmut-like DNA hypermethylation. *Nature* **551:** 384–388 Available at: http://www.nature.com/articles/nature24294

Rappsilber J, Mann M & Ishihama Y (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2:** 1896–1906 Available at: http://www.nature.com/articles/nprot.2007.261

Rodriguez-Meira A, Buck G, Clark S-A, Povinelli BJ, Alcolea V, Louka E, McGowan S, Hamblin A, Sousos N, Barkas N, Giustacchini A, Psaila B, Jacobsen SEW, Thongjuea S & Mead AJ (2019) Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Mol. Cell* **73:** 1292-1305.e8 Available at: https://linkinghub.elsevier.com/retrieve/pii/S1097276519300097

Schoof EM, Lechman ER & Dick JE (2016) Global proteomics dataset of miR-126 overexpression in acute myeloid leukemia. *Data Br.* **9:** 57–61 Available at: http://www.ncbi.nlm.nih.gov/pubmed/27656662

Shlush LI, Mitchell A, Heisler L, Abelson S, Ng SWK, Trotman-Grant A, Medeiros JJF, Rao-Bhatia A, Jaciw-Zurakowsky I, Marke R, McLeod JL, Doedens M, Bader G, Voisin V, Xu C, McPherson JD, Hudson TJ, Wang JCY, Minden MD & Dick JE (2017) Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature* **547:**

104–108 Available at: http://www.nature.com/articles/nature22993

Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, et al (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **46:** D661–D667 Available at: http://academic.oup.com/nar/article/46/D1/D661/4612963

Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary D, Warshawsky D, Guan-Golan Y, Kohn A, Rappaport N, Safran M & Lancet D (2016) The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. In *Current Protocols in Bioinformatics* pp 1.30.1-1.30.33. Hoboken, NJ, USA: John Wiley & Sons, Inc. Available at: http://doi.wiley.com/10.1002/cpbi.5

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES & Mesirov JP (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102:** 15545–15550 Available at: https://www.pnas.org/content/102/43/15545.long

Ting L, Rad R, Gygi SP & Haas W (2011) MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8:** 937–40 Available at: http://www.ncbi.nlm.nih.gov/pubmed/21963607

Trapnell C (2015) Defining cell types and states with single-cell genomics. *Genome Res.* **25:** 1491–8 Available at: http://www.ncbi.nlm.nih.gov/pubmed/26430159

Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA & Quake SR (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509:** 371–375 Available at: http://www.nature.com/articles/nature13173

Vogel C & Marcotte EM (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13:** 227–232 Available at: http://www.ncbi.nlm.nih.gov/pubmed/22411467

Winter SV, Meier F, Wichmann C, Cox J, Mann M & Meissner F (2018) EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. *Nat. Methods* **15:** 527–530 Available at: http://dx.doi.org/10.1038/s41592-018-0037-8

Wojtowicz EE, Lechman ER, Hermans KG, Schoof EM, Wienholds E, Isserlin R, van Veelen

PA, Broekhuis MJC, Janssen GMC, Trotman-Grant A, Dobson SM, Krivdova G, Elzinga J, Kennedy J, Gan OI, Sinha A, Ignatchenko V, Kislinger T, Dethmers-Ausema B, Weersing E, et al (2016) Ectopic miR-125a Expression Induces Long-Term Repopulating Stem Cell Capacity in Mouse and Human Hematopoietic Progenitors. *Cell Stem Cell*

Wolf FA, Angerer P & Theis FJ (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19:** 15 Available at: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1382-0

Zhu Y, Piehowski PD, Zhao R, Chen J, Shen Y, Moore RJ, Shukla AK, Petyuk VA, Campbell-Thompson M, Mathews CE, Smith RD, Qian W-J & Kelly RT (2018) Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. *Nat. Commun.* **9:** 882 Available at: http://www.nature.com/articles/s41467-018-03367-w

Figure legends

Figure 1 - A) Overview of a typical AML hierarchy, with LSC at the apex. B) Sorting scheme for sorting the bulk, blast, progenitor and LSC populations from the 8227 culture system, followed by the TMT labeling setup for single-cell proteomics, consisting of 9-10 single cells and 1-2 booster channels. C) Overview of the SCeptre computational single cell MS pipeline.

Figure 2 – A) Standard PCA clustering of the single cell proteomics data, when using 1) a bulk cell booster, 2) a bulk cell + cell-specific booster, or 3) a cell-specific only booster channel. B) Overview of the tSNE and UMAP clusterings of the bulk, blast, LSC and progenitor cells in the different datasets, combined with the number of proteins found in the specific cells as overlaid on the same computational embedding. C) Standard PCA clustering of the TMT SPS MS3 data. D) tSNE and UMAP clustering of the TMT SPS MS3 data, with the number of proteins found in the specific cells overlaid on the same computational embedding.

Figure 3 - A) Sorting overview for the control TMT pools, where single cells of each differentiation stage were combined with a booster channel of one differentiation stage, until all possible combinations were met. B) Total number of protein identifications for the control samples, using either MS2 and MS3 level TMT quantitation. C) Standard PCA

analysis of single cells, coloured either by cell differentiation stage or TMT pool. The TMT colour scheme highlights a significant batch effect where cells cluster mainly according to TMT pool rather than differentiation stage. D) SCeptre analysis of control samples, highlighting a clear clustering pattern according to cellular differentiation stage. Number of proteins identified in each cell is indicated in the accompanying plots, overlaid on the same tSNE/UMAP embedding as the original clustering.

Figure 4 - A) UMAP clustering of the different differentiation stages when using protein expression values only, or when using the Eigenprotein information derived from pathway integration. B) The Eigenprotein cliques overlaid on the Eigenprotein-based UMAP embedding, highlighting a Progenitor/LSC, an LSC and a blast-specific Eigenprotein clique respectively. C) Wilcoxon-ranked-sum ranking of top proteins defining the three differentiation stages and bulk. D) Heatmap visualization of the top three proteins from the Wilcoxon-ranked-sum testing. E) Protein expression levels of top three LSC and Blast proteins, overlaid on the Eigenprotein UMAP embedding.

Supplementary Figure 1 - Overview of all 54 Eigenprotein cliques identified by SCeptre, overlaid on the Eigenprotein UMAP embedding.

Supplementary Figure 2 - UMAP embedding plot, highlighting the CD34 expression levels within the single cells (lower panels), compared to the UMAP cell differentiation clusters (upper panels). Values are plotted both on protein expression-based embedding, and Eigenprotein embedding.

Tables and their legends

Dataset	Booster Cells	MS Instrument	# of Cells	# of Protein IDs	Average # of Protein IDs p. cell (after batch correction)
1	Bulk	QExactive HFX (MS2)	320	1,259	389 (min: 24, max: 522)
2	Bulk + Cell-specific	QExactive HFX (MS2)	320	2,138	994 (min: 318, max: 1157)
3	Cell-specific	QExactive HFX (MS2)	400	2,452	389 (min: 143, max: 532)
4	Cell-specific	Orbitrap Fusion (MS3)	160	1,216	259 (min: 63, max: 375)

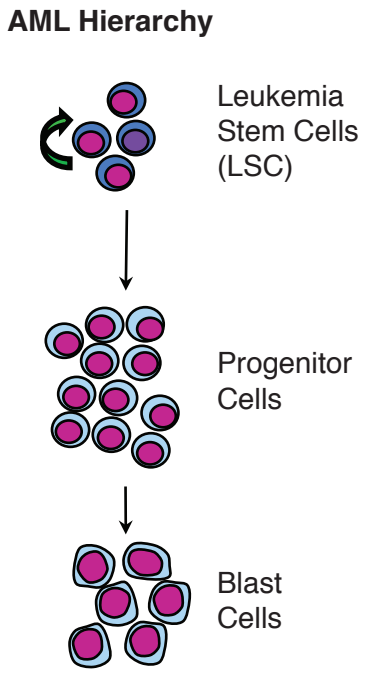
Table 1 - Overview of Protein identification numbers across the various datasets

Expanded View Figure legends

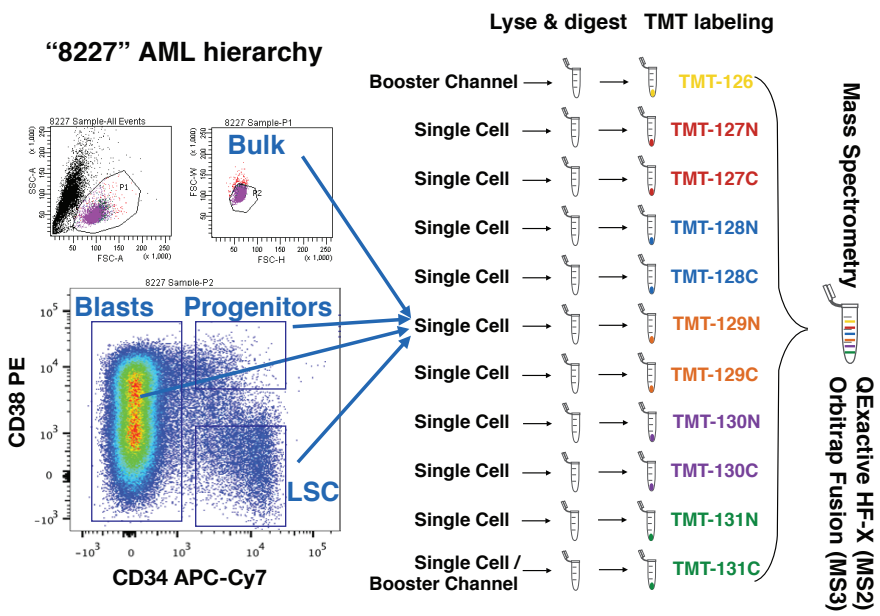
Supplementary Table 1 – Overview of Eigenprotein cliques and the pathways and genes contained therein.

Figure 1

A)



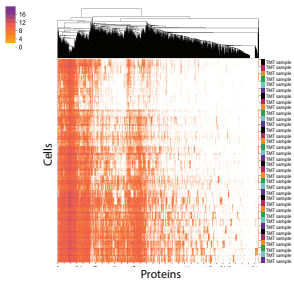
B)



C)

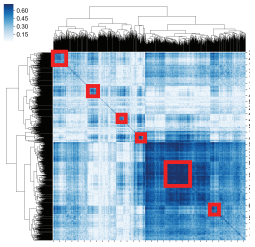
SCeptre Workflow

Step 1



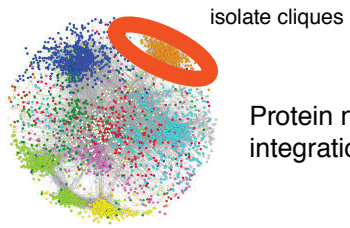
Batch correction by cell type and TMT pool at protein expression level

Step 2



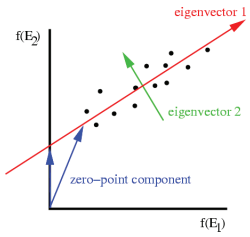
Protein correlation

Step 3



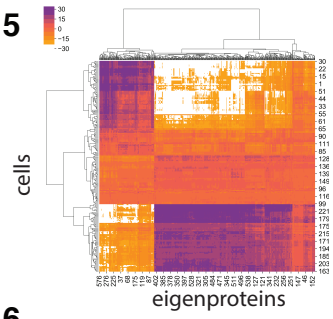
Protein network integration

Step 4



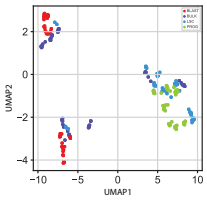
Computing EigenProteins

Step 5



Data cleaning and clustering

Step 6

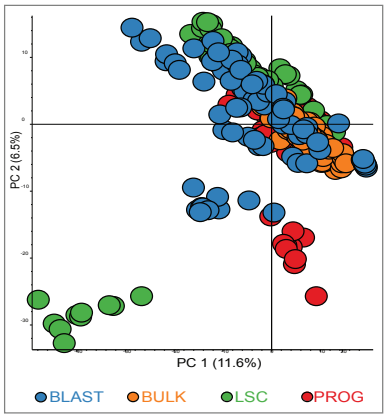


UMAP visualization

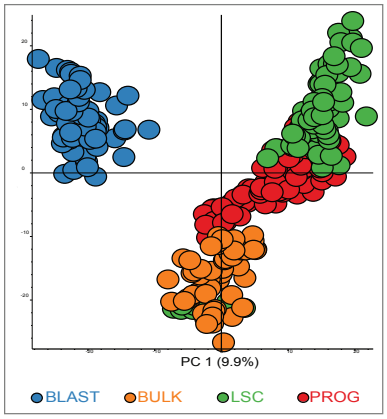
Figure 2

A)

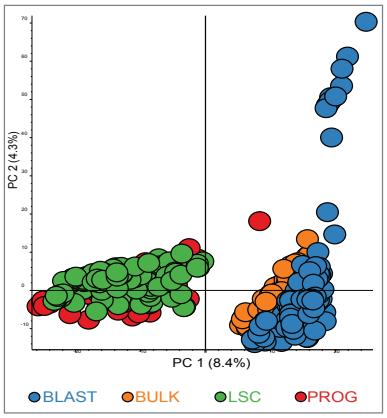
Bulk Booster



Bulk + Cell type-specific Booster



Cell type-specific Booster

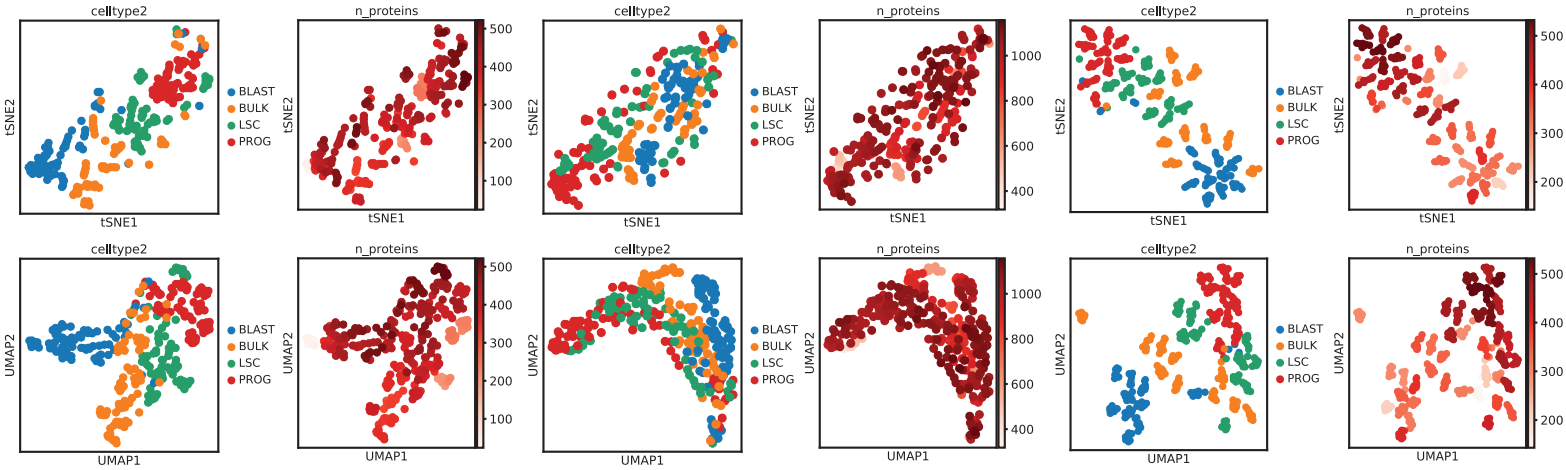


B)

Bulk Booster

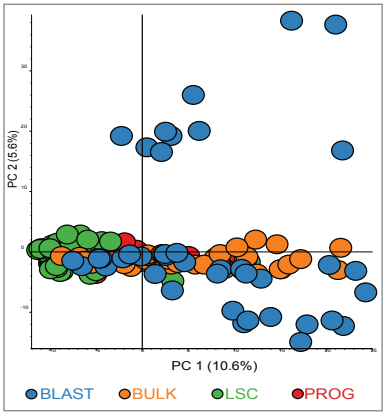
Bulk + Cell type-specific Booster

Cell type-specific Booster



C)

TMT SPS MS3



D)

TMT SPS MS3

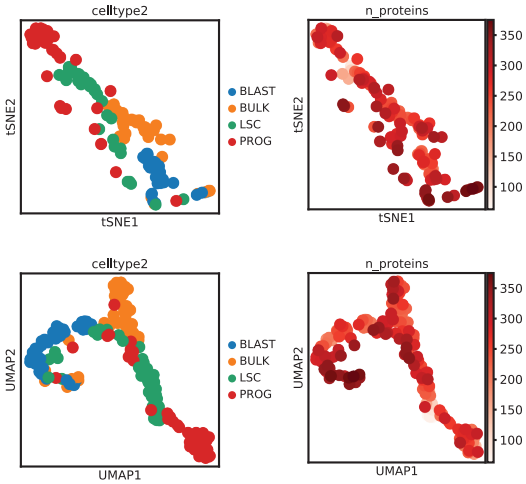


Figure 3

A)

Cells Sorted

MS Type	Booster Cells	96 well plate 1 (126)	2 (127N)	3 (127C)	4 (128N)	5 (128C)	6 (129N)	7 (129C)	8 (130N)	9 (130C)	10 (131N)	11 (131C)	12
MS2	Bulk	500 cells	PROG	BULK	LSC	LSC	-	BULK	BLASTS	BLASTS	PROG	PROG	-
MS2	Bulk	500 cells	PROG	BULK	LSC	LSC	-	BULK	BLASTS	BLASTS	PROG	PROG	-
MS3	Bulk	500 cells	PROG	BULK	LSC	LSC	-	BULK	BLASTS	BLASTS	PROG	LSC	-
MS3	Bulk	500 cells	PROG	BULK	LSC	LSC	-	BULK	BLASTS	BLASTS	PROG	LSC	-
MS2	Blasts	500 cells	BLASTS	BLASTS	BULK	BULK	PROG	PROG	LSC	LSC	-	PROG	-
MS2	Blasts	500 cells	BLASTS	BLASTS	BULK	BULK	PROG	PROG	LSC	LSC	-	PROG	-
MS3	Blasts	500 cells	BLASTS	BLASTS	BULK	BULK	PROG	PROG	LSC	LSC	-	LSC	-
MS3	Blasts	500 cells	BLASTS	BLASTS	BULK	BULK	PROG	PROG	LSC	LSC	-	LSC	-

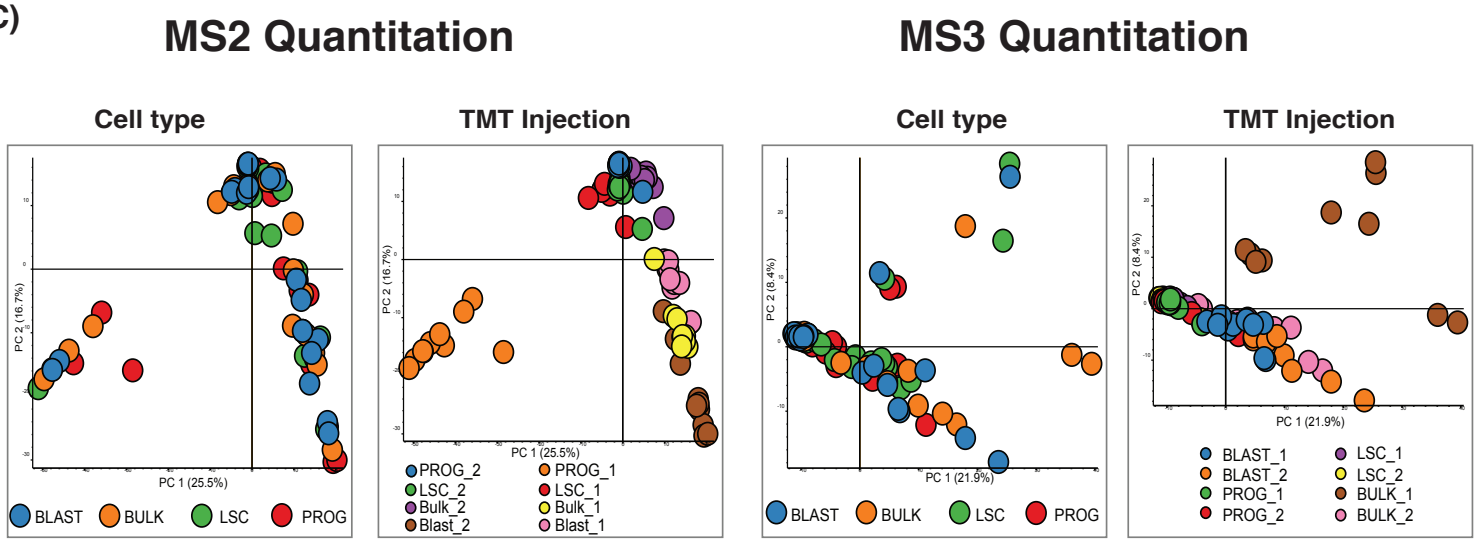
BULK	
BLASTS	
PROG	
LSC	

MS Type	Booster Cells	96 well plate 1 (126)	2 (127N)	3 (127C)	4 (128N)	5 (128C)	6 (129N)	7 (129C)	8 (130N)	9 (130C)	10 (131N)	11 (131C)	12
MS2	Prog	500 cells	-	PROG	BLASTS	BLASTS	BULK	LSC	BULK	PROG	LSC	BULK	-
MS2	Prog	500 cells	-	PROG	BLASTS	BLASTS	BULK	LSC	BULK	PROG	LSC	BULK	-
MS3	Prog	500 cells	-	PROG	BLASTS	BLASTS	BULK	LSC	BULK	PROG	LSC	BLASTS	-
MS3	Prog	500 cells	-	PROG	BLASTS	BLASTS	BULK	LSC	BULK	PROG	LSC	BLASTS	-
MS2	LSC	500 cells	LSC	LSC	PROG	PROG	BLASTS	BLASTS	-	BULK	BULK	BULK	-
MS2	LSC	500 cells	LSC	LSC	PROG	PROG	BLASTS	BLASTS	-	BULK	BULK	BULK	-
MS3	LSC	500 cells	LSC	LSC	PROG	PROG	BLASTS	BLASTS	-	BULK	BULK	BLASTS	-
MS3	LSC	500 cells	LSC	LSC	PROG	PROG	BLASTS	BLASTS	-	BULK	BULK	BLASTS	-

B)

	Total Protein IDs
Ctrls_MS2	1,437
Ctrls_MS3	937

C)



D)

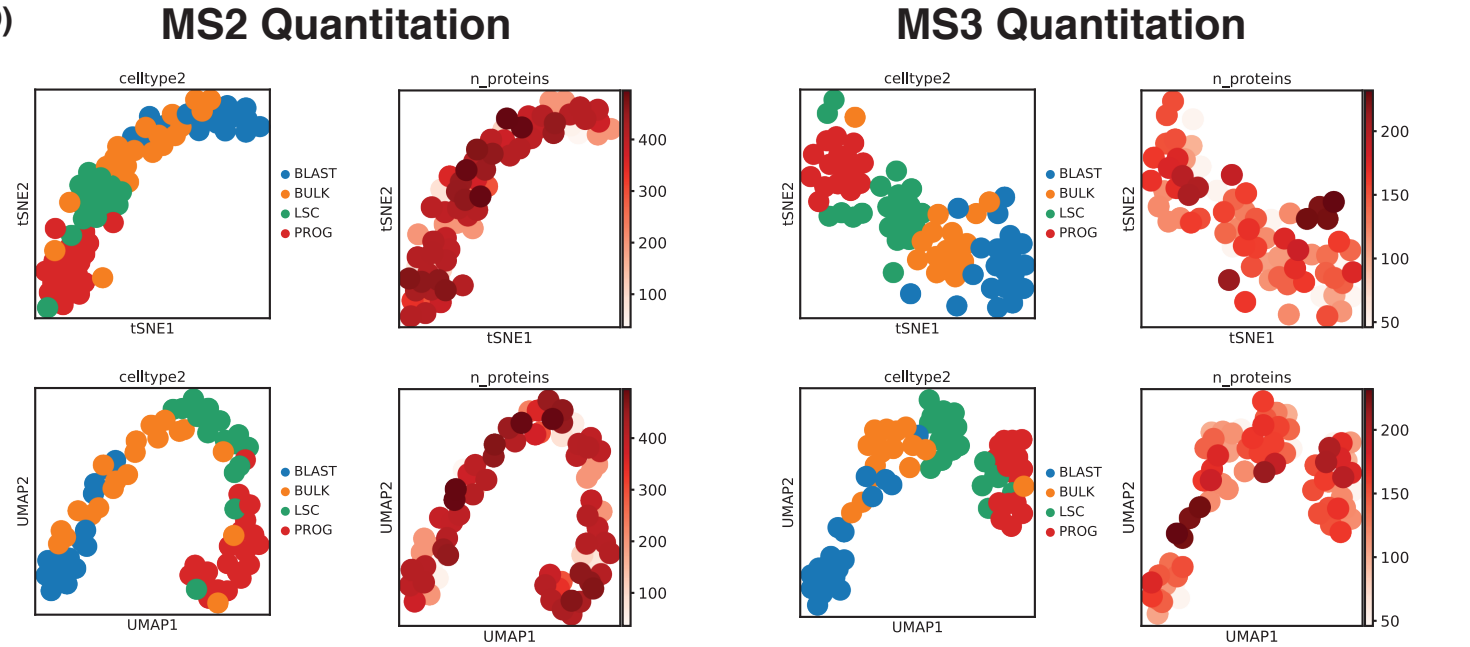
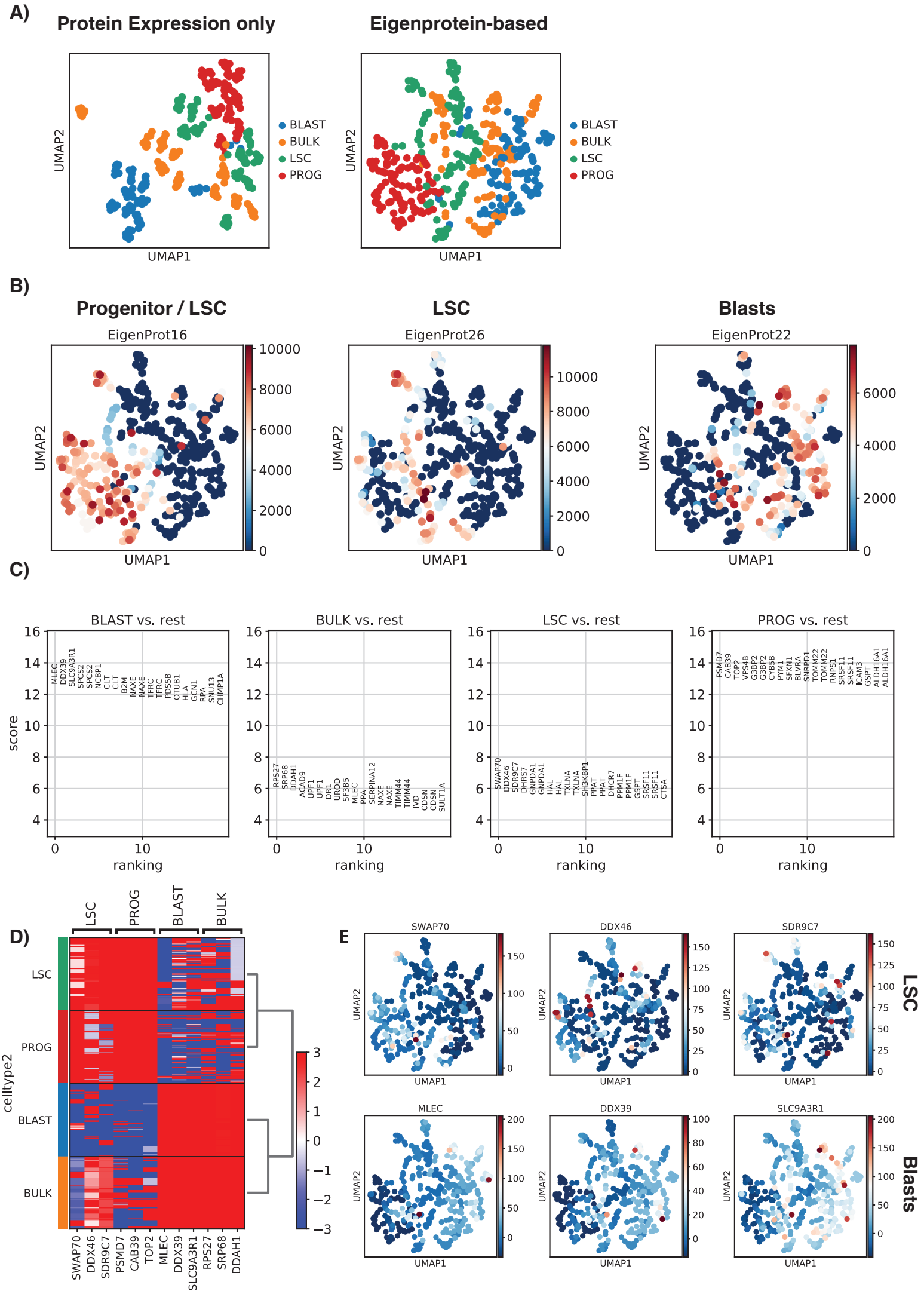


Figure 4


A)
Protein Expression only
Eigenprotein-based
UMAP2
UMAP1
BLAST
BULK
LSC
PROG
B)
Progenitor / LSC
LSC
Blasts
EigenProt16
EigenProt26
EigenProt22
C)
BLAST vs. rest
BULK vs. rest
LSC vs. rest
PROG vs. rest
score
ranking
MLEC DDX39 SLC9A3R1 SPCS2 SPCS2 NCBP1 CLT CLT B2M NAXE NAXE TFRC TFRC PDS5B OTUB1 HLA GCN1 RPA SNU13 CHMP1A
RPS27 SRP68 DDAH1 ACAD9 UPF1 UPF1 DR1 UROD SF3B5 MLEC PPA SERPINA12 NAXE NAXE TIMM44 TIMM44 IVD CDSN CDSN SULT1A
SWAP70 DDX46 SDR9C7 DHRS7 GNPDA1 GNPDA1 HAL HAL TXLNA TXLNA SH3KBP1 PPAT PPAT DHCR7 PPM1F PPM1F GSPT SRSF11 SRSF11 CTSA
PSMD7 CAB39 TOP2 VPS4B G3BP2 G3BP2 CYB5B PYM1 SFXN1 BLVRA SNRPD1 TOMM22 TOMM22 RNPS1 SRSF11 SRSF11 ICAM3 GSPT ALDH16A1 ALDH16A1
D)
LSC
PROG
BLAST
BULK
celltype2
SWAP70 DDX46 SDR9C7 PSMD7 CAB39 TOP2 MLEC DDX39 SLC9A3R1 RPS27 SRP68 DDAH1
E
SWAP70
DDX46
SDR9C7
MLEC
DDX39
SLC9A3R1
LSC
Blasts