Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation

Bogdan Budnik,[1] Ezra Levy,[2] Nikolai Slavov[2,3]
[1]MSPRL, FAS Center for Systems Biology, Harvard University, Cambridge, MA 02138, USA
[2]Department of Biology, Northeastern University, Boston, MA 02115, USA
[3]Department of Bioengineering, Northeastern University, Boston, MA 02115, USA

Cellular heterogeneity is important to biological processes, including cancer[1,2] and development[3]. However, proteome heterogeneity is largely unexplored because of the limitations of existing methods for quantifying protein levels in single cells. To alleviate these limitations, we developed Single Cell ProtEomics by Mass Spectrometry (SCoPE-MS), and validated its ability to identify distinct human cancer cell types based on their proteomes. We used SCoPE-MS to quantify over a thousand proteins in differentiating mouse embryonic stem (ES) cells. The single-cell proteomes enabled us to deconstruct cell populations and infer protein abundance relationships. Comparison between single-cell proteomes and transcriptomes indicated coordinated mRNA and protein covariation. Yet many genes exhibited functionally concerted and distinct regulatory patterns at the mRNA and the protein levels, suggesting that post-transcriptional regulatory mechanisms contribute to proteome remodeling during lineage specification, especially for developmental genes. SCoPE-MS is broadly applicable to measuring proteome configurations of single cells and linking them to functional phenotypes, such as cell type and differentiation potentials.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

Cellular systems, such as tissues, cancers, and cell cultures, consist of a variety of cells with distinct molecular and functional properties. Characterizing such cellular differences is key to understanding normal physiology, combating cancer recurrence[1,2], and enhancing targeted differentiation for regenerative therapies[3]; it demands quantifying the proteomes of single cells.

However, quantifying proteins in single mammalian cells remains confined to fluorescent imaging and antibodies. Fluorescent proteins have proved tremendously useful but are limited to quantifying only a few proteins per cell and sometimes introduce artifacts[4]. Multiple methods for quantifying proteins in single cells have been recently developed, including single-cell Western blots[5], CyTOF[6], and Proseek Multiplex, an immunoassay readout by RT-PCR[7]. These methods enabled quantifying up to a few dozen endogenous proteins but their throughput and accuracy are limited by the availability of highly-specific antibodies that bind their cognate proteins stoichiometrically.

We aimed to overcome these limitations by developing a high-throughput method for Single Cell ProtEomics by Mass Spectrometry (SCoPE-MS) that can quantify thousands of proteins in single mammalian cells. To develop SCoPE-MS, we resolved two major challenges: (i) delivering the proteome of a mammalian cell to a MS instrument with minimal protein losses and (ii) simultaneously identifying and quantifying peptides from single-cell samples. To overcome the first challenge, we manually picked live single cells under a microscope and lysed them mechanically (by Covaris sonication in glass microtubes) in phosphate-buffered saline, Fig. 1a. This method was chosen to obviate chemicals that may undermine peptide separation and ionization or sample cleanup that may incur significant losses. The proteins from each cell lysate were quickly denatured at $90\ ^oC$ and digested with trypsin at $45\ ^oC$ overnight, Fig. 1a; see Methods for full experimental details.

To overcome the second challenge, we made novel use of tandem mass tags (TMT). This technology was developed for multiplexing[8], which affords cost-effective high-throughput. Even more crucial to our application, TMT allows quantifying the level of each TMT-labeled peptide in each sample while identifying its sequence from the total peptide amount pooled across all samples[8]. SCoPE-MS capitalizes on this capability by augmenting each single-cell set with a sample comprised of $\sim 100 - 200$ carrier cells that provide enough ions for peptide sequence identification, Fig. 1a. Increasing the number of carrier cells increases peptide identification rates

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

but decreases quantitative precision. The carrier cells also help with the first challenge by reducing losses from single cells, since most of the peptides sticking to tips and tube walls originate from the carrier cells. Thus, the carrier cells help overcome the two major challenges.

Quantification of TMT-labeled peptides relies on reporter ions (RI) whose levels reflect both peptide abundances and noise contributions, such as coisolation interference and background noise[8,9]. To evaluate the contribution of background noise to single-cell RI quantification, we estimated the signal-to-noise ratio (SNR), Extended Data Fig. 1. The estimates indicated that RI intensities are proportional to the amount of labeled single-cell proteomes, and very low for channels left empty. These data suggest that the signal measured in single cells exceeds the background noise by 10-fold or more. As an added SNR control for every TMT set, SCoPE-MS leaves the 130N channel empty, thus simultaneously avoiding isotopic cross-contamination from the carrier cells in channel 131 and having a channel that reflects the background noise.

To evaluate the ability of SCoPE-MS to distinguish different cell types, we prepared two label-swapped and interlaced TMT sets with alternating single Jurkat and U-937 cells, two blood cancer cell lines with average cell diameter of only 11 μm (Fig. 1b). The levels of all 583 proteins quantified in single cells were projected onto their principle components (PC). The three-dimensional projections of single-cell proteomes clustered by cell type (Fig. 1c), suggesting that SCoPE-MS can identify cell types based on their proteomes. Next, we identified proteins whose levels vary less within a cell type than between cell types. Among the 117 proteins showing such trends at FDR $< 2\%$, we plotted the distributions for seven in Fig. 1d. Some of these proteins are expected to be cell type specific, such as the higher abundance of Complement C3 in the U-937 cells, which are myeloid lineage precursors for macrophages. The consistency of protein fold-changes between Jurkat and U-937 cells is also reflected in the positive correlations among fold-changes estimated from different cells and TMT channels, Extended Data Fig. 2.

Next, we quantified single-cell proteome heterogeneity and dynamics during ES cell differentiation. To initiate differentiation, we withdrew leukemia inhibitor factor (LIF) from ES cell cultures and transitioned to suspension culture; LIF withdrawal results in complex and highly heterogeneous differentiation of epiblast lineages in embryoid bodies (EB). We used SCoPE-MS to quantify over a thousand proteins at FDR $= 1\ \%$ (Extended Data Fig. 3a) and their pair-wise cor-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

relations (averaging across single cells) in days 3, 5, and 8 after LIF withdrawal (Fig. 2a). Cells from different days were processed together to minimize batch biases[10].

We first explored protein covariation as reflected in the overrepresentation of functionally related proteins within highly coherent clusters of protein-protein correlations, Fig. 2a. The large clusters on all days are enriched for proteins with biosynthetic functions. This covariation is consistent with the possibility of heterogeneous and asynchronous slowing of cell growth as cells differentiate. The smaller clusters correspond to lineage-specific proteins and more specialized functions.

Next, we projected the proteomes of single cells from all days (190 cells) onto their PCs, Fig. 2b. The projections cluster by date; indeed, PC 1 loading correlate to the days post LIF withdrawal, Extended Data Fig. 3b. The small clusters of lineage-specific genes (Fig. 2a) suggest that we have quantified proteomes of distinct cell states; thus we attempted to identify cell clusters by projecting the EB proteomes onto their PCs and identifying sets of proteins that are concertedly regulated in each cluster, Fig. 2c,d. The projection resulted in clusters of cells, whose identity is suggested by the dominant proteins in the singular vectors. We identified biological functions overrepresented[11] within the distribution of PC loadings and colorcoded each cell based on the average levels of proteins annotated to these functions. The PCs do not correlate to missing data, indicating that our experimental design has overcome challenges common to high-throughput single-cell data[10]; see Methods. These results suggest that SCoPE-MS data can meaningfully classify cell identity for cells from complex and highly heterogeneous populations.

Klein *et al.*[12] recently quantified mRNA heterogeneity during ES differentiation, and we used their inDrop data to simultaneously analyze mRNA and protein covariation and to directly test whether genes coexpressed at the mRNA level are also coexpressed at the protein level. To this end, we computed all pairwise correlations between RNAs (Fig. 3a) and proteins (Fig. 3b) for all genes quantified at both levels in cells undergoing differentiation for 7 and 8 days. Clustering hierarchically the correlation matrices results in 3 clusters of genes. To compare these clusters, we computed the pairwise Jaccard coefficients, defined as the number of genes present in both classes divided by the number of genes present in either class, i.e., intersection/union). The results (Fig. 3c) indicate that the largest (green) cluster is 55 % identical and the medium (blue) cluster is

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

33 % identical. This cluster stability is also reflected in a positive correlation between corresponding mRNA and protein correlations, Fig. 3d. The magnitude of this correlation is comparable to protein-mRNA correlations from bulk datasets[11,13] and testifies strongly for the quantitative accuracy of both inDrop and SCoPE-MS.

Having established a good overall concordance between mRNA and protein covariation, we next explored whether and how much this concordance varies between genes with different biological functions. The covariation concordance of a gene was estimated as the similarity of its mRNA and protein correlations, i.e., the correlation between the corresponding correlation vectors[14]. The median concordance of ribosomal proteins (RP) of both the 60S (RPL) and 40S (RPS) is significantly higher than for all genes, Fig. 3e. This result indicates that RPL and RPS genes have significantly ($p < 10^{-20}$) more similar gene-gene correlations at the mRNA and the protein levels than the other quantified genes. Some RPs correlate less well to the remaining RPs (Extended Data Fig. 4), which may reflect lineage specific ribosome remodeling, but this possibility needs to be evaluated more directly with isolated ribosomes[15]. In contrast to RPs, genes functioning in tissue morphogenesis, proteolysis, and development have significantly ($p < 10^{-3}$) lower concordance at the mRNA and protein level than all genes, Fig. 3e.

The power of MS proteomics had been circumscribed to bulk samples. Indeed, the TMT manufacturer recommends $100\ \mu g$ of protein per channel, almost 10^6 more than the protein content of a typical mammalian cell. SCoPE-MS bridged this gap by efficient sample preparation and the use of carrier cells. These innovations open the gates to further improvements (e.g., increased multiplexing) that will make single-cell MS proteomics increasingly powerful.

SCoPE-MS enabled us to classify cells and explore the relationship between mRNA and protein levels in single mammalian cells. This first foray into single mammalian proteomes demonstrates that mRNA covariaion is predictive of protein covariaion even in single cells. It further establishes the promise of SCoPE-MS to quantitatively characterize single-cell gene regulation and classify cell types.

Acknowledgments: We thank S. Semrau, M. Jovanovic, R. Zubarev, and members of the Slavov laboratory for discussions and constructive comments, as well as the Harvard University FAS Sci-

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

ence Operations for supporting this research project. This work was funded by startup funds from Northeastern University and a New Innovator Award from the NIGMS from the National Institutes of Health to N.S. under Award Number DP2GM123497.

Competing Interests: The authors declare that they have no competing financial interests.

Contributions: B.B., and N.S. conceived the research. B.B., E.L. and N.S. performed experiments and collected data; N.S. analyzed the data and wrote the manuscript.

The raw MS data have been deposited in MassIVE (ID: MSV000080489) and in the ProteomeX change (ID: 0000398). Supplemental website can be found at:
http://www.northeastern.edu/slavovlab/data_webs.htm

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

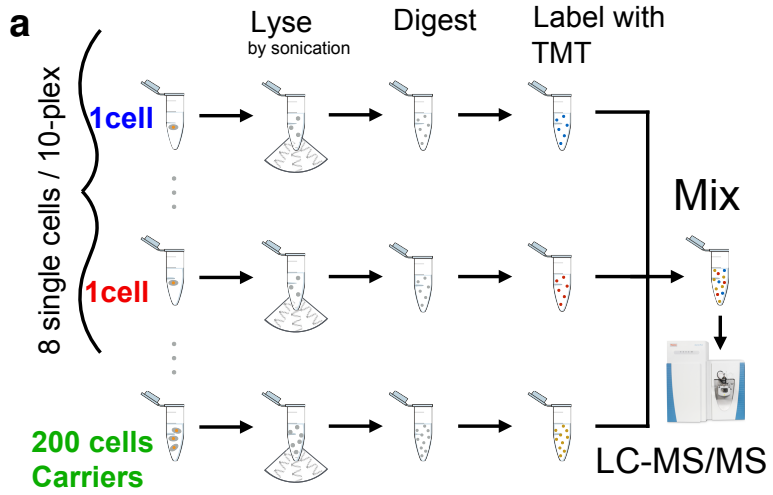
Figure Captions

Figure 1 | Validating SCoPE-MS by classifying single cancer cells based on their proteomes. (**a**) Conceptual diagram and work flow of SCoPE-MS. Individually picked live cells are lysed by sonication, the proteins in the lysates are digested with trypsin, the resulting peptides labeled with TMT labels, combined and analyzed by LC-MS/MS (Orbitrap Elite). (**b**) Design of control experiments used to test the ability of SCoPE-MS to distinguish U-937 cells from Jurkat cells. (**c**) Unsupervised principal component (PC) analysis using data for all quantified proteins from the experiments described in panel (b) stratifies the proteomes of single cancer cells by cell type. (**d**) Distributions of protein levels across single U-937 and Jurkat cells indicate cell-type-specific protein abundances.

Figure 2 | Identifying protein covariation and cell clusters across differentiating ES cells. (**a**) Clustergrams of pairwise protein-protein correlations in cells differentiating for 3, 5, and 8 days after LIF withdrawal. The correlation vectors were hierarchically clustered based on the cosine of the angles between them. (**b**) The proteomes of all single EB cells were projected onto their PCs, and the marker of each cell color-coded by day. The single-cell proteomes cluster by day, a trend also reflected in the distributions of PC 1 loadings by day, Extended Data Fig. 2. (**c, d**) The proteomes of cells differentiating for 8 days were projected onto their PCs, and the marker of each cell color-coded based on the normalized levels of all proteins from the indicated gene-ontology groups.

Figure 3 | Coordinated mRNA and protein covariation in differentiating ES cells. (**a**) Clustergram of pairwise correlations between mRNAs with 2.5 or more reads per cell as quantified by inDrop in single EB cells[12]. (**b**) Clustergram of pairwise correlations between proteins quantified by SCoPE-MS in 12 or more single EB cells. (**c**) The overlap between corresponding RNA from (a) and protein clusters from (b) indicates similar clustering patterns. (**d**) Protein-protein correlations correlate to their corresponding mRNA-mRNA correlations. Only genes with significant mRNA-mRNA correlations were used for this analysis. (**e**) The concordance between corresponding mRNA and protein correlations (computed as the correlation between between corresponding correlations[14]) is high for ribosomal proteins (RPL and RPS) and lower for developmental genes; distribution medians are marked with red pluses. Only the subset of genes quantified at both RNA and protein levels were used for all panels.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017



b

Experimental Design Table

Label	Set 1	Set 2
126	U-937	Jurkat
127N	Jurkat	U-937
127C	U-937	Jurkat
128N	Jurkat	U-937
128C	U-937	Jurkat
129N	Jurkat	U-937
129C	U-937	Jurkat
130N	*empty*	*empty*
130C	Jurkat	U-937
131 Carriers	100 Jurkat 100 U-937	100 Jurkat 100 U-937

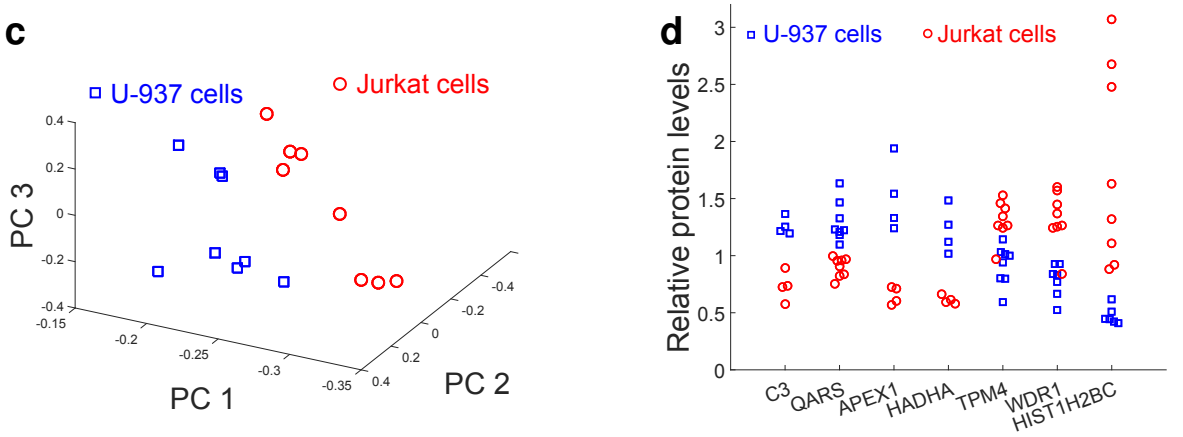


Figure 1 | Validating SCoPE-MS by classifying single cancer cells based on their proteomes. (**a**) Conceptual diagram and work flow of SCoPE-MS. Individually picked live cells are lysed by sonication, the proteins in the lysates are digested with trypsin, the resulting peptides labeled with TMT labels, combined and analyzed by LC-MS/MS (Orbitrap Elite). (**b**) Design of control experiments used to test the ability of SCoPE-MS to distinguish U-937 cells from Jurkat cells. (**c**) Unsupervised principal component (PC) analysis using data for all quantified proteins from the experiments described in panel (b) stratifies the proteomes of single cancer cells by cell type. (**d**) Distributions of protein levels across single U-937 and Jurkat cells indicate cell-type-specific protein abundances.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

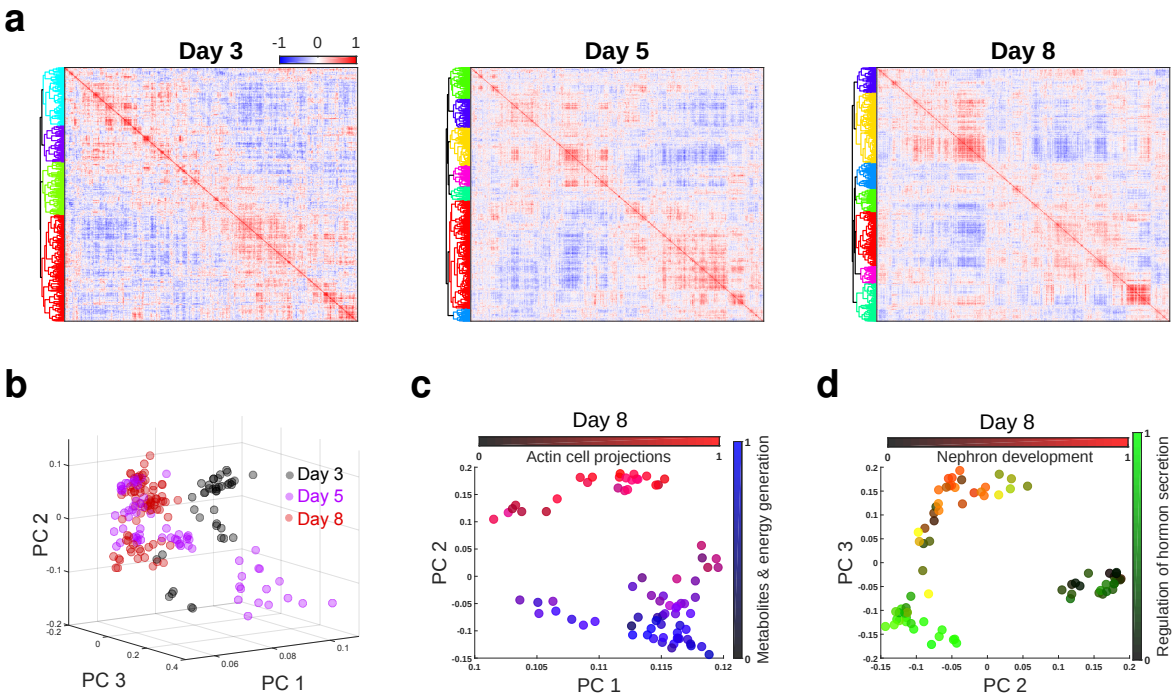


Figure 2 | Identifying protein covariation and cell clusters across differentiating ES cells. (**a**) Clustergrams of pairwise protein-protein correlations in cells differentiating for 3, 5, and 8 days after LIF withdrawal. The correlation vectors were hierarchically clustered based on the cosine of the angles between them. (**b**) The proteomes of all single EB cells were projected onto their PCs, and the marker of each cell color-coded by day. The single-cell proteomes cluster by day, a trend also reflected in the distributions of PC 1 loadings by day, Extended Data Fig. 2. (**c, d**) The proteomes of cells differentiating for 8 days were projected onto their PCs, and the marker of each cell color-coded based on the normalized levels of all proteins from the indicated gene-ontology groups.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

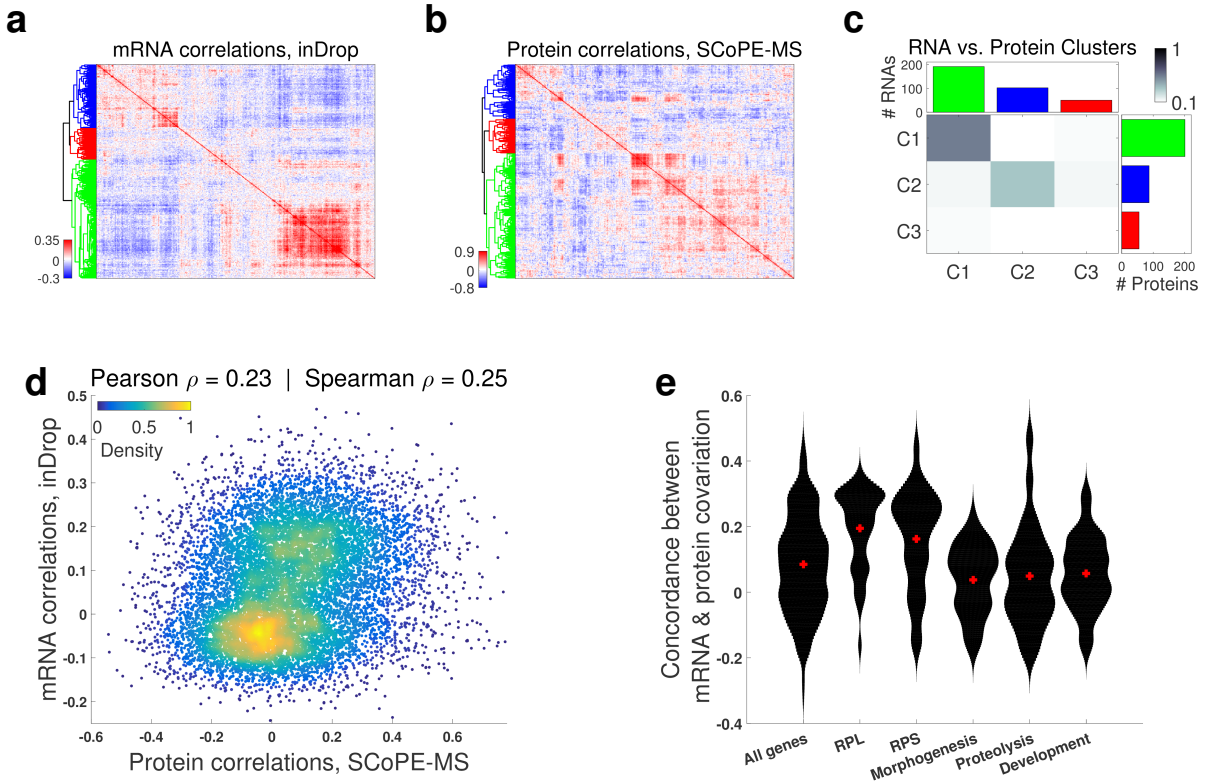
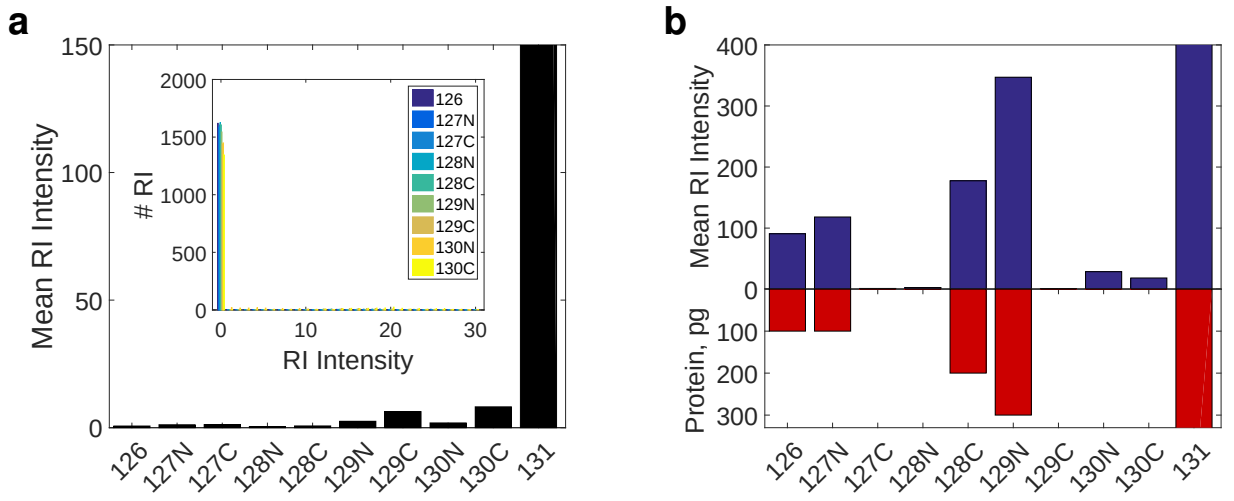


Figure 3 | Coordinated mRNA and protein covariation in differentiating ES cells.
(**a**) Clustergram of pairwise correlations between mRNAs with 2.5 or more reads per cell as quantified by inDrop in single EB cells[12]. (**b**) Clustergram of pairwise correlations between proteins quantified by SCoPE-MS in 12 or more single EB cells. (**c**) The overlap between corresponding RNA from (a) and protein clusters from (b) indicates similar clustering patterns. (**d**) Protein-protein correlations correlate to their corresponding mRNA-mRNA correlations. Only genes with significant mRNA-mRNA correlations were used for this analysis. (**e**) The concordance between corresponding mRNA and protein correlations (computed as the correlation between between corresponding correlations[14]) is high for ribosomal proteins (RPL and RPS) and lower for developmental genes; distribution medians are marked with red pluses. Only the subset of genes quantified at both RNA and protein levels were used for all panels.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

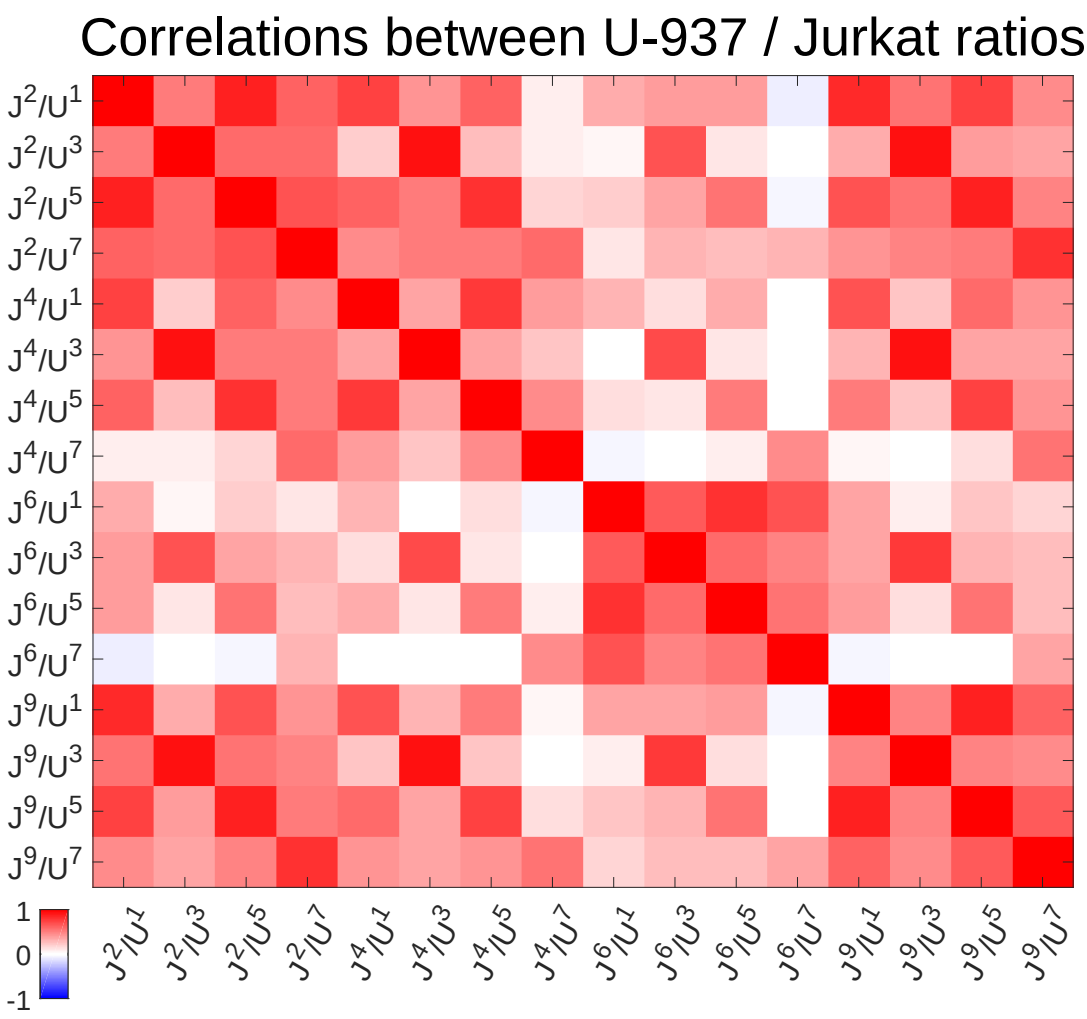
Extended Data Figures



Extended Data Figure 1 | Contribution of background noise to quantification of peptides in single cells. (**a**) Reporter ion (RI) intensities in a SCoPE set in which the single cells were omitted while all other steps were carried out, i.e., trypsin digestion, TMT labeling and addition of carrier cells in channel 131. Thus, RI intensities in channels $126 - 130C$ correspond to background noise. The distribution of RI intensities in the inset shows that the RI for most peptides in channels $126 - 130C$ are zero, i.e., below the MaxQuant noise threshold. The y-axis is limited to 150 to make the mean RI intensities visible. The mean RI intensity for single-cell channels is about 500. (**b**) Mean RI intensities for a TMT set in which only 6 channels contained labeled proteome digests and the other 4 were left empty. Channels 126, 127N, 128C, and 129N correspond to peptides diluted to levels corresponding to 100, 100, 200 and 300 picograms of cellular proteome, channel 131 corresponds to the carrier cells (bars truncated by axes), and the remaining channels were left empty. The RI for most peptides are not detected in the empty channels, and their mean levels very low. This suggests that background noise is low compared to the signal from peptides corresponding to a single cell.

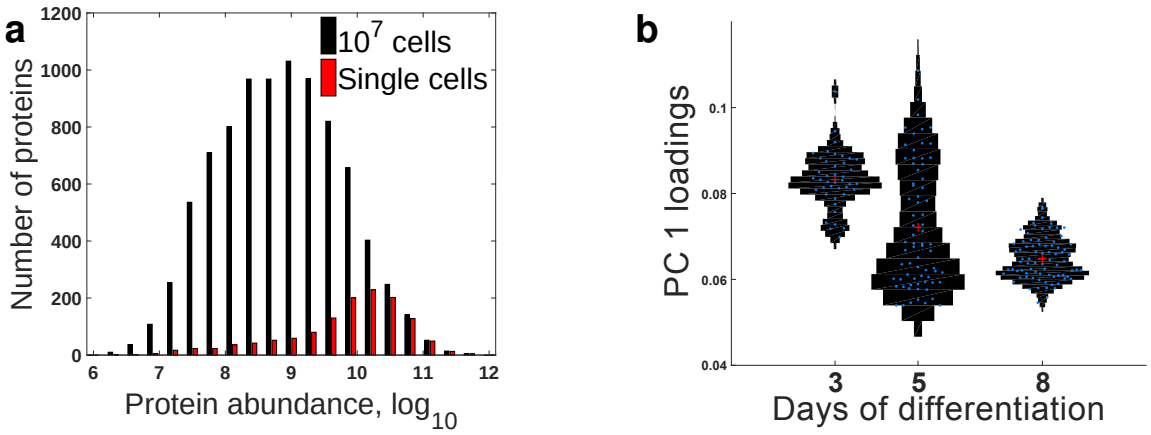
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

a



Extended Data Figure 2 | Consistency of protein ratios between Jurkat and U-937 cells estimated from different combinations of TMT channels. (**a**) A correlation matrix of all pairwise Pearson correlations among the ratios of peptide abundances in U-937 and in Jurkat cells from Set 1 in Fig. 1b. The superscripts corresponds to the TMT labels ordered by mass, with 1 being 126, 2 being 127N and so on.

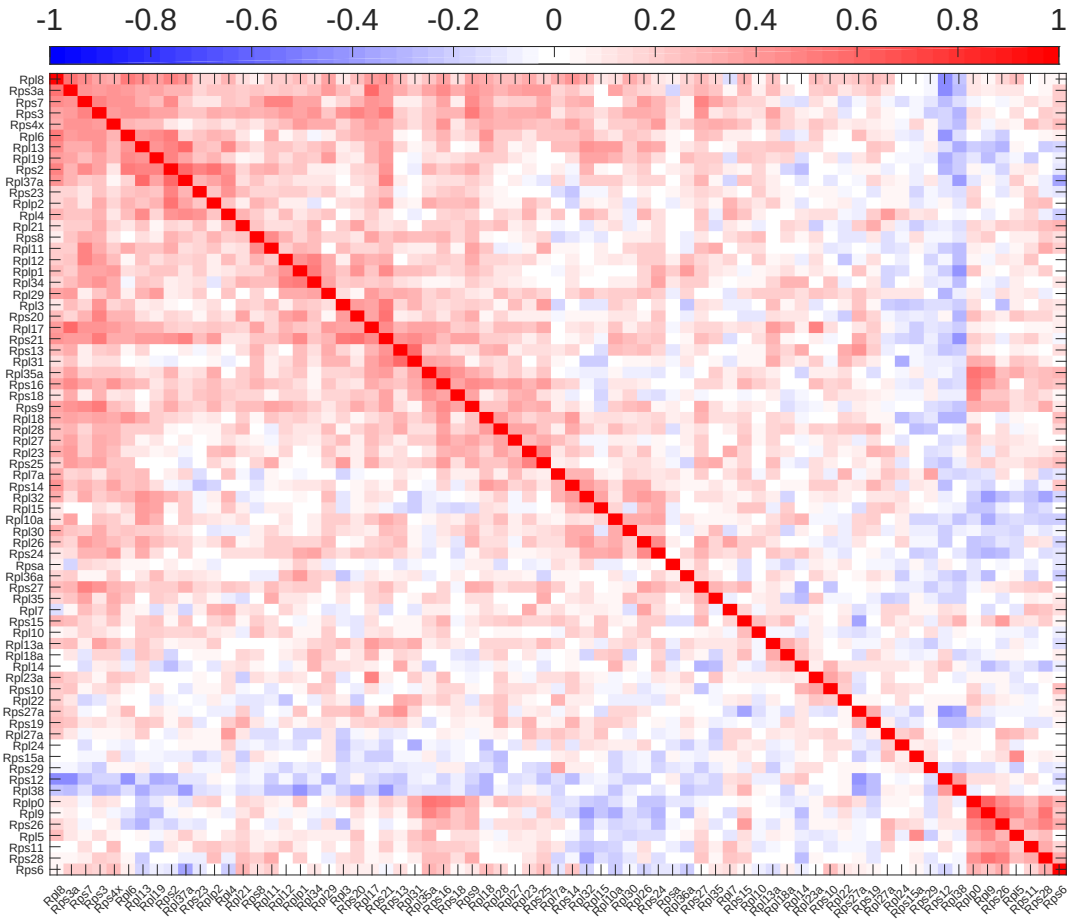
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017



Extended Data Figure 3 | Proteome coverage of differentiating ES cells and distributions of the PC 1 loadings by day of differentiation. (**a**) Distribution of protein abundances for all proteins quantified from 10^7 differentiating ES cells or in at least one single-cell SCoPE-MS set at FDR ≤ 1 %. The probability of quantifying a protein by SCoPE-MS is close to 100 % for the most abundant proteins quantified in bulk samples and decreases with protein abundance, for total of 1526 quantified proteins. (**b**) The proteomes of all differentiating single cells were decomposed into singular vectors and values, and distributions of the loading (elements) of the singular vector with the largest singular value, i.e., PC 1, shown as violin plots. Individual blue circles correspond to single cells, and the red crosses correspond to the medians for each day.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

a



Extended Data Figure 4 | **Correlations between ribosomal proteins** (**a**) All pairwise Pearson correlations between ribosomal proteins on day 8 were computed by averaging across cells. The correlations matrix was clustered, using the cosine between the correlation vectors as a similarly measure.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

References

1. Dean, M., Fojo, T. & Bates, S. Tumour stem cells and drug resistance. *Nature Reviews Cancer* **5,** 275–284 (2005).
2. Cohen, A. A. *et al.* Dynamic proteomics of individual cancer cells in response to a drug. *Science* **322,** 1511–1516 (2008).
3. Semrau, S. & van Oudenaarden, A. Studying lineage decision-making in vitro: emerging concepts and novel tools. *Annual review of cell and developmental biology* **31,** 317–345 (2015).
4. Landgraf, D., Okumus, B., Chien, P., Baker, T. A. & Paulsson, J. Segregation of molecules at cell division reveals native protein localization. *Nature methods* **9,** 480–482 (2012).
5. Hughes, A. J. *et al.* Single-cell western blotting. *Nature methods* **11,** 749–755 (2014).
6. Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332,** 687–696 (2011).
7. Darmanis, S. *et al.* Simultaneous multiplexed measurement of RNA and proteins in single cells. *Cell reports* **14,** 380–389 (2016).
8. Ross, P. L. *et al.* Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics* **3,** 1154–1169 (2004).
9. Savitski, M. M. *et al.* Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *Journal of proteome research* **12,** 3586–3598 (2013).
10. Hicks, S. C., Teng, M. & Irizarry, R. A. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv* **1,** 025528 (2015).
11. Franks, A., Airoldi, E. & Slavov, N. Post-transcriptional regulation across human tissues. *bioRxiv* **1,** DOI: 10.1101/020206 (2016).
12. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161,** 1187–1201 (2015).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017

13. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).

14. Slavov, N. & Dawson, K. A. Correlation signature of the macroscopic states of the gene regulatory network in cancer. *Proceedings of the National Academy of Sciences* **106,** 4079–4084 (2009).

15. Slavov, N., Semrau, S., Airoldi, E., Budnik, B. & van Oudenaarden, A. Differential stoichiometry among core ribosomal proteins. *Cell Reports* **13,** 865–873 (5 2015).

16. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26,** 1367–1372 (2008).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2767v1 | CC BY 4.0 Open Access | rec: 31 Jan 2017, publ: 31 Jan 2017