Main Manuscript for

Correlated protein modules reflecting functional coordination of interacting proteins are uncovered by single-cell proteomics

Mo Hu [a,b,1], Yutong Zhang [a,b,c,1], Yuan Yuan [a,b,c], Wenping Ma [a,b,d], Yinghui Zheng [a,b], Qingqing Gu [e], X. Sunney Xie [a,b]

[a] Beijing Advanced Innovation Center for Genomics, Peking University,100871 Beijing, China; [b] Biomedical Pioneering Innovation Center, Peking University, 100871 Beijing, China; [c] Collage of Chemistry and Molecular Engineering, Peking University, 100871 Beijing, China; [d] Peking-Tsinghua Center for Life Sciences (CLS), Academy for Advanced Interdisciplinary Studies, Peking University, 100871, Beijing, China and [e] Changping Laboratory, Beijing, China.

[1] Mo Hu and Yutong Zhang contributed equally to this work.

* X. Sunney Xie

Email: sunneyxie@biopic.pku.edu.cn

Keywords: single-cell proteomics; correlated protein modules; mass spectrometry; protein-protein interaction; gene ontology

This PDF file includes:

Main Text
Figures 1 to 4

Abstract

Single-cell proteomics has attracted a lot of attention in recent years because it offers more functional relevance than single-cell transcriptomics. However, most work to date focused on cell typing, which has been widely accomplished by single-cell transcriptomics. Here we report the use of single-cell proteomics to measure the correlations between the translational levels of any pair of proteins in a single mammalian cell. To do so, we chose label-free instead of isobaric labeling quantification for single-cell mass spectrometry in order to minimize measurement noise. In measuring pairwise correlations among ~1,000 proteins in steady-state K562 cells, we found multiple correlated protein modules (CPMs), each containing a group of highly positively correlated proteins that are functional interacting and collectively involved in certain biological functions, such as protein synthesis and oxidative phosphorylation. Some CPMs are shared across different cell types while others are cell-type specific. Although widely studied in omics analyses, pairwise correlations are often measured by introducing perturbations to bulk samples, which would mask the coordination of gene expression in steady-state condition. The single-cell correlations probed in our experiment reflect intrinsic steady-state fluctuations in the absence of perturbation. We note that observed correlations between proteins are experimentally more distinct and functionally more relevant than those between corresponding mRNAs measured in single-cell transcriptomics. By virtue of single-cell proteomics, functional coordination of proteins is manifested through CPMs.

42 **Significance Statement**

43 In a mammalian cell, thousands of proteins are expressed, barely continuously, but at fluctuating levels
44 over time. Yet many proteins need to work together in order to perform a specific biological function. By
45 conducting accurate mass spectrometry-based proteomics of single cells, we discovered multiple protein
46 groups, each exhibiting positive pairwise correlations under steady-state condition. Through such
47 observations, it can be inferred that certain proteins are synchronized to collectively carry out specific
48 biological functions, such as protein synthesis. Our results demonstrate that single-cell proteomics can be
49 used to probe protein-protein interactions and study how proteins work together in a system-wide
50 approach.

51

52

53 **Main Text**

54

55 **Introduction**

56 In a living cell, abundances of proteins and mRNAs fluctuate stochastically over time or even oscillate
57 periodically (1, 2), hence cell-to-cell variations. This is because in a single cell there are only one or two
58 copies of DNA (2–4), whose transcription and translation cannot be synchronized among different cells.
59 On the other hand, a particular cellular function, such as protein synthesis, often requires multiple
60 proteins, which, if each fluctuating, need to be coordinated in order for them to work together. While bulk
61 measurement fails to study the ubiquitous fluctuation, the variation of mRNA abundances and protein
62 abundances can be revealed at single-cell level, thanks to the single-cell omics technologies. Although
63 the genome-wide fluctuation cannot be studied through single-cell omics in real time, the intrinsic
64 dynamics may be deduced through correlation analysis on single cells in steady state (5) (**Fig. 1a**).

65 Currently, single-cell transcriptomics study has been widely used to distinguish different types of cells.
66 Apart from cell typing, high-precision single-cell transcriptomics can also measure correlations in steady-
67 state gene expression (6, 7). However, proteins, rather than mRNAs, are direct executers of most
68 biological processes. Since protein abundance cannot be accurately predicted by mRNA (8–10), the
69 dynamics of protein abundances and their correlations can only be assessed at the protein level.
70 Moreover, small copy numbers of mRNA in single cells combined with limited detectability of current
71 technologies (11, 12) would lead to high measurement noise (**Fig. 1b**), hindering accurate correlation
72 measurements (**Fig. 1c**). Despite a much lower level of detectability and limited number of cells,

73 correlations measured at the protein level still proved to be more faithful than those measured at RNA
74 level (**Fig. 1c, d**).

75 To study the functional correlations among proteins, various methods have been developed, including
76 protein correlation profiling (PCP) (13–16) ,which is based on the hypothesis that interacting proteins (or
77 proteins in the same pathway) should be positively correlated. Also, correlation-centric approach, with
78 accurate quantification by mass spectrometry, proved to be powerful in capturing indirect interactions and
79 proteins that interact but do not co-localize (17). However, these methods all rely on perturbations and
80 bulk measurement where protein abundances are averaged among different cells, and cannot probe
81 intrinsic dynamics of steady-state cells. Therefore, system-wide single-cell proteomics is needed.

82 Current mass spectrometry-based single-cell proteomics studies have achieved identification and
83 quantification of hundreds or even thousands of proteins in a single cell using isobaric labeling (18, 19) or
84 label-free (20, 21) quantification. Most existing studies focus on distinguishing different types of cells (18–
85 20), which has been widely accomplished by single-cell transcriptomics analysis. To date, single-cell
86 proteomics has not been used to study proteome-wide pairwise protein correlation.

87 Here we report a robust mass spectrometry-based single-cell proteomic pipeline, which allowed us to
88 measure correlations among ~1,000 proteins in K562 cells. By measuring pairwise correlations based on
89 protein abundance fluctuations in homogeneous cells, we identified correlated protein modules (CPMs),
90 which are groups of proteins with highly positive pairwise correlations. CPMs were characterized by
91 enrichment of certain cellular functions and known protein complexes. Furthermore, we demonstrated

92 that CPMs identified from single-cell proteomic measurement could provide clearer and more solid
93 evidence on functional relationships among proteins than single cell RNA-seq based approach.

94 **Results**

95 **Development of a single-cell proteomics pipeline for accurate protein correlation measurement**

96 To measure pairwise correlations between proteins, protein abundances need to be measured accurately
97 in the first place. To this end, we evaluated two established methods for MS-based single-cell proteomics
98 analysis, *i.e.*, the MS^1-based label-free quantification and the carrier proteome-based isobaric labeling
99 with TMT quantification (**Fig. 2a**). We optimized these two pipelines by considering the throughput,
100 coverage of proteome and robustness of the quantification performance. For label-free pipeline,
101 commonly used gradient for single-cell proteomics ranges from 40 minutes to 2 hours. Longer gradient
102 generally improves the protein identification, with a penalty on the number of cells to be analyzed each
103 day. Here we ran single-shot LC-MS experiments at a moderate gradient (16 cells per day) to balance the
104 throughput and quantification accuracy. For isobaric labeling pipeline, to improve the accuracy of
105 measurements, real-time search (RTS) and MS^3 were used to reduce the interference from co-isolation of
106 other precursors. The throughput of the isobaric labeling pipeline is higher than that of the label-free
107 pipeline, reaching more than 100 cells each day.

108 Human leukemia cell line K562 was used as single-cell samples to test the performance of these two
109 pipelines. We first compared the number of identified proteins in each run between two methods. The
110 label-free pipeline identified more proteins than did the isobaric labeling pipeline (**Fig. 2b**). Compared with
111 the isobaric labeling pipeline, the number of proteins identified using the label-free pipeline vary more

112 largely across different runs, ranging from 600 to 1000 proteins each run. In MS-based proteomics, a
113 protein identified is not necessarily quantified. Nevertheless, with the help of match-between-runs (MBR)
114 method, the label-free pipeline was able to quantify more proteins compared to the isobaric labeling
115 pipeline. Due to the unavailability of related software, no MBR or other peptide identity propagation
116 methods was applied to the isobaric labeling-based quantification.

117 We then analyzed the coefficient of variance (CV) of protein abundances among cells, which was
118 substantially lower for label-free pipeline compared with isobaric labeling. Notably, there were
119 substantially more proteins with CV lower than 0.3 in the label-free pipeline (**Fig. 2c**). As for the
120 distribution of raw MS signals, the signals for quantification in single-cell channels in the isobaric labeling
121 pipeline (MS^3 signal-to-noise ratio) are much lower than in the label-free pipeline (MS^1 intensity) (**Fig. 2d**).
122 We further evaluated the performance of two pipelines in measuring pairwise correlations of protein
123 abundances (SI Appendix, **Fig. S1a**). The Pearson correlation coefficient of protein abundances
124 measured by two methods both centered around zero and symmetrically distributed from negative region
125 to positive region (SI Appendix, **Fig. S1b**). Ribosomal proteins were chosen to assess the quality of
126 measured correlation coefficient. The label-free pipeline was able to quantify more ribosomal proteins and
127 discover more highly positive pairwise correlations than the isobaric labeling pipeline, while the
128 correlations between randomly picked proteins measured by the two pipelines distributed similarly (**Fig.
129 2e, f**). Taken together, the label-free pipeline was selected to be used in this study.

130 **Correlated protein modules in K562 cell line**

131 Pairwise correlations of protein abundances were measured based on single-cell measurements of the
132 K562 cell line. In total, 1409 proteins were quantified in 74 cells. After removing contaminates, low-quality
133 proteins and outlier cells, the final dataset was established comprising 1249 proteins from 69 cells.
134 Pearson correlation coefficients of the abundances of these proteins centered around zero (SI Appendix,
135 **Fig. S1b**). After unsupervised hierarchy clustering based on Pearson correlation coefficients of protein
136 abundances, we found groups of highly positive correlated proteins, named as correlated protein modules
137 (CPMs). 66 CPMs were found in K562 cells, containing 546 proteins in total (**Fig. 3a**). Each CPM
138 contained 5 to 38 proteins. The correlation coefficients between protein pairs within a CPM, with average
139 values ranging from 0.29 to 0.53, was much higher than those outside CPMs (SI Appendix, **Fig. S2a**).
140 To examine potential functional relationships among proteins in each CPM, we compared our data with
141 CORUM (22) to see if CPMs represent certain protein complexes. Several subunits from the same
142 complex appeared in the same CPM. For example, CPM #7 contained 5 spliceosome subunits and CPM
143 #34 contained 4 EIF subunits. CPM #19 was almost entirely composed of T-complex proteins (5/7) and
144 CPM #28 mostly consisted of ribosomal proteins (37/38).
145 Proteins in the same CPMs tend to have the same subcellular localization, as shown by enrichment
146 analysis of the Gene Ontology cellular component (GOCC). For example, CPM #15 was enriched with the
147 ribonucleoprotein complex and telomeric region while CPM #7 was enriched with the CMG complex. Over
148 half of the CPMs had significant enrichment of GOCC terms.
149 Functional relationships indicated by the Gene Ontology biological process (GOBP), KEGG pathway (23)
150 and Reactome (24) pathway enrichment were highly prevalent among CPMs (**Fig. 3c,** SI Appendix, **Fig.**

151 **S3)**. For example, CPM #56 was enriched with “tRNA aminoacylation (GO:0006418)” while CPM #41 was
152 enriched with "antigen processing and presentation (KEGG: hsa04612)". That most CPMs were
153 significantly enriched with GOBP terms suggests that involving in the same biological process is a major
154 contributor to protein correlations in steady state.
155 Some CPMs were enriched with multiple functions. For example, in CPM #6, TRIM28 and USP7 were
156 involved in “DNA demethylation (GO:0080111)” while DCD and S100A9 were related to “antimicrobial
157 humoral response (GO:0019730)”. This implies that DNA demethylation and antimicrobial humoral
158 response are closely related, consistent with a previous study (25).
159 Finally, we studied the relationship between each protein pair within a CPM in the STRING database (26)
160 to see how well protein correlations reflect actual protein-protein interactions. It was found that all CPMs
161 except four had corresponding PPIs. Over 90% of protein pairs in CPM #56 and CPM #61 had known
162 PPIs in the STRING database. To further evaluate the relationship of pairwise protein correlations, CPMs
163 and protein-protein interactions, we calculated the fraction of known protein-protein interaction at different
164 correlation level between pairs of proteins in steady-state single-cells. For a pair of proteins, a larger
165 correlation coefficient means higher possibility that the two proteins have known interactions (**Fig. 3d**).
166 We also found the protein pairs within CPMs have much higher possibility to have known PPI in the
167 STRING database compared to a pair of proteins not in the same CPM (**Fig. 3d**), which is consistent with
168 the fact that CPMs reflect functional coordination of interacting proteins.
169 As CPM carries out specific cellular functions, correlations among CPMs may imply relationships among
170 different cellular activities. We defined the abundance of a CPM as the sum of the abundances of all

171 proteins included in that CPM. We further measured the relationship between CPMs based on the
172 pairwise correlations of CPM abundances. Most of the CPMs were poorly correlated, with correlation
173 coefficients from 0 to 0.2. However, some CPM pairs showed negative correlations lower than -0.3, which
174 means the abundance of one CPM goes up in a certain cell when the other one goes down. For example,
175 CPM #11 and CPM #57 had a negative correlation coefficient (SI Appendix, **Fig. S2b**). The 11 proteins in
176 CPM #11 were related to "rRNA processing (GO:0006364)" and "DNA post-replication repair
177 (GO:0006301)", while the 9 proteins in CPM #57 were involved in "ATP generation (GO:0006757)" and
178 "NADH metabolism (GO:0006734)". For cells that had high CPM #11 abundance, the abundance of CPM
179 #57 was rather low, consistent with a previous study reporting that ribosome biosynthesis is controlled by
180 fluctuation of NAD(+)/NADH ratio (27). CPM #57 was also negatively correlated with CPM #37 (SI
181 Appendix, **Fig. S2c)** which contains proteins TOP2A and CDK1. These two proteins were found to be
182 involved in regulating the circadian rhythm (28). Thus, this negative correlation suggested that CPM #57
183 was a circadian rhythm-dependent fluctuating CPM.

184 **CPM is a universal feature of functional proteome**

185 To confirm that CPM is a universal feature of functional proteome, we further examined protein
186 correlations in the human 293T cell line and mouse MII oocytes. The correlation map of 1150 proteins in
187 63 293T cells revealed 50 CPMs (SI Appendix, **Fig. S4a**). Despite the differences between K562 and
188 293T cell line in protein levels and in sizes of CPMs, the functions of CPMs in the two cell lines appeared
189 similar. For example, T-complex CPM was also found in 293T cells (SI Appendix, **Fig. S4d**) as in K562
190 cells. Proteins of tRNA aminoacylation module in K562 were also highly correlated in 293T, despite of

191 marked differences between the two cell lines in the abundance of these proteins (SI Appendix, **Fig.**
192 **S4c**). This shows that although proteins with shared function could be tuned at different levels, the
193 coordination of synchronized fluctuation is consistent among different types of cells.
194 Experiments on 137 mouse MII oocytes quantified 3422 proteins, which is about three-fold the number of
195 quantified proteins in regular-size mammalian cells. We discovered 86 CPMs from these proteins (SI
196 Appendix, **Fig. S4b**). Compared with K562 and 293T, CPMs of mouse oocytes are generally bigger in
197 size and tend to infer more complex biological functions. CPMs of certain protein complexes identified in
198 K562 and 293T cells, such as T-complex (SI Appendix, **Fig. S4d, e**), were also found in mouse oocytes,
199 indicating that these CPMs are conserved across different species and cell types. CPMs associated with
200 certain functions, such as proteasome complex, could only be identified in oocytes (SI Appendix, **Fig.**
201 **S4f**). Therefore, the modules identified from single-cells can infer cellular characteristics and are affected
202 by the total number of identified proteins.

203 **CPMs reveal more functional information than CGMs and bulk proteomic approach**

204 Similar to CPMs, correlated gene modules (CGMs) have also been reported based on pairwise
205 correlations of RNA expression levels via MALBAC-dt, a single-cell RNA-seq method with high
206 detectability and high accuracy (6). To compare the efficacy of CPM versus CGM in functional analyses,
207 we examined the correlation coefficients of genes that have been identified from both methods. Measured
208 correlation coefficients of RNAs were more concentrated around zero compared to those of proteins (**Fig.**
209 **4a**). We also found the pairwise correlation coefficients of proteins in functional modules are generally
210 higher than those of corresponding RNAs. For example, for ribosome module, the average correlation

211 coefficients among proteins is 0.55, compared with 0.32 for RNAs (**Fig. 4b**). Thus, compared to
212 transcripts of ribosome proteins, the correlation among ribosomal proteins are more distinguishable from
213 the overall pairwise correlations (**Fig. 4c**). Furthermore, some functional modules identified at the protein
214 level cannot be observed at the RNA level, such as CPM #11 (rRNA processing and mismatch repair)
215 and CPM #3 (inner mitochondrial membrane oxidation phosphorylation) (**Fig. 4d**). These differences are
216 likely due to the limited detectability of single-cell RNA-seq, or to the possibility that some genes with a
217 shared function are coordinated at post-transcriptional level but not at transcriptional level. Therefore,
218 CPM with higher correlation coefficients compared to CGM can provide more relevant and pointed
219 information on gene coordination.
220 In traditional proteomics study, correlation-centric analysis of proteins has been done with perturbations
221 and bulk measurement, as in ProteomeHD (17). We compared CPMs from single-cell analysis with those
222 from ProteomeHD. Although some modules, such as ribosome and T-complex, were observed in both
223 datasets, some biological processes like tRNA aminoacylation (CPM #56) were not found in bulk data
224 (**Fig**. **4e**). This shows that our method, which relies on single-cell measurement without perturbations, is
225 more sensitive in probing intrinsic correlations among proteins.
226

227 **Discussion**

228 High-throughput single-cell analysis has proved to be a valuable strategy for understanding the functional
229 mechanisms of cells and tissues (7, 29). However, identifying functional gene expression programs from
230 large-scale single-cell dataset has been challenging. Most methods emphasize on heterogeneity among

231 cells, such as principal component analysis (PCA) (30, 31) or non-negative matrix factorization (NMF)
232 (32). These methods can uncover the variation of gene expression and cellular activity among different
233 cell types, but cannot identify the networks of gene expression in homogeneous cell populations.
234 Nonetheless, investigating correlations on the proteome level is still pertinent. This is because while
235 mRNA fluctuation arises from stochastic expression of a single copy of gene in a live cell (2, 4), protein
236 fluctuation arises from additional variation including translation, post-translational modification and protein
237 degradation (9). Thus, direct measurement of protein correlation is required to uncover the intrinsic
238 dynamics of protein abundance in steady state. Moreover, as the copy number of proteins in single cell is
239 much larger than that of mRNA, measurement of protein can reveal real correlation of genes at much
240 higher accuracy.
241 Here, through global correlation analysis of single-cell proteome measured by mass spectrometry, we
242 revealed the functional coordination of proteins despite stochastic gene expression under steady-state
243 conditions. Our method probed the intrinsic fluctuations of more than a thousand of proteins and found
244 groups of proteins that were highly correlated. Based on functional analysis of CPMs and the
245 relationships between different CPMs, we discovered coordinated fluctuation for proteins involved in the
246 same functions, related pathways or conflicting processes, which is inaccessible by traditional single-cell
247 studies and bulk studies.
248 The functional coordination and interaction of proteins is a universal feature among different types of
249 cells, yet it is determined by the cellular characteristic and affected by the detectability of the protein
250 measuring technology. Currently, the single-cell proteomics still has limited coverage of proteome,

251 especially when compared to single-cell transcriptomics which can identify more than 5,000 genes in a
252 single cell. Therefore, while most key regulators in signaling networks, such as transcription factors, can
253 be readily measured by current single-cell transcriptomics, these low-abundance proteins cannot be
254 detected by single-cell proteomics yet.
255 Finally, we reclaimed the importance of accurate quantification in single-cell proteomics. Suffering from
256 multiple isolation events and longer cycle times, the isobaric labeling quantification pipeline had much
257 lower raw MS signal for quantification, leading to higher CV and less faithful measurement of correlation
258 coefficient. In comparison, the label-free pipeline demonstrated advantages in protein identification and
259 quantification. However, current label-free quantification pipeline has limited proteome coverage and cell
260 throughput (up to 40 cells per day (20, 33)). As the proteome coverage of single-cell proteomics
261 improves, we may be able to gain more functional insights from CPM, such as coregulation mechanism of
262 transcription factors. This possibility is showcased by our mouse oocyte dataset, where the total protein
263 amount per cell is about 100-fold of that in K562. Therefore, substantial effort is needed to make CPM
264 widely applicable to cell cultures or tissues, where much larger number of cells need to be measured and
265 higher proteome coverage is preferred. This is possible as sample preparation (34–36), LC separation
266 (37, 38), MS instrumentation (33), data acquisition (39), and informatics (40, 41) are witnessing rapid
267 advances. We believe that CPM is a good demonstration of single-cell omics methods for functional
268 annotation of the proteome.
269
270 **Materials and Methods**
271 **Preparation of single cell samples**

272 K562 cells were resuspended from culture dish by cold PBS. 293T cells were treated by trypsin for 3 min
273 before washed from dish by cold PBS. Then both cells were washed two more times by cold PBS. Single
274 cell or 10 cells were sorted to 96-well plates (Eppendorf) containing 0.5 μL 1% DDM in water by FACS
275 (FACSAria, BD). Then plates were stored at -80 °C before use. Oocyte-cumulus complexes were
276 collected from C57/6J strain mice superovulated with pregnant mare serum gonadotropin (PMSG) and
277 HCG injections. Hyaluronidase was used to remove cumulus cells. MII oocytes were exposed to PBS
278 solution. Single oocyte cell was then picked by pipetting with an aspirator tube. Then cells were stored at
279 -80 °C before use.

280 **Proteomic sample preparation**

281 We used an acoustic liquid handling workstation (Echo 525, Beckman) to increase throughput and handle
282 small volumes. Sorted single cells in 96-well plate or 200 μL PCR tube were digested directly by trypsin at
283 37 °C for 3 hours. For label-free pipeline, 4.4 μL 0.43% TFA and 1% acetonitrile in water was added to
284 terminate digestion (3.5 μL for oocytes). For isobaric labeling pipeline with TMTpro, channel 126C
285 (equivalent to 10 cells) was used as carrier; channel 127N (equivalent to 3 cells) was used as reference;
286 128C-130C and 131C-133C were single-cell channels; 131N and 134N were negative control channels
287 (no cell). 0.85 μL of 11 mM TMTpro reagent was added to plate. The plate was kept at 22 °C for 1 h.
288 Then 0.2 μL of 0.5% hydroxylamine was added to plate. The plate was held at 22 °C for 30 min. Then
289 channels of isobaric labeling were mixed with 5 μL of carrier/reference mix and 10 μL of 30% acetonitrile.
290 Then peptides were dried in a concentrator (LabCondo) at 45 °C. 5 μL of 0.1% TFA and 1% ACN in water
291 were used to resuspend the peptides.

292 **LC-MS/MS**

293 4 μL of the digests were injected. Peptides were separated by a commercialized high-performance
294 chromatography column (Aurora, IonOpticks) at 100 nL/min on a nanoflow liquid chromatography
295 (Ultimate 3000 RSLCnano, ThermoFisher). For label-free pipeline, the effective gradient was 70 min,
296 generating throughput of 16 cells per day (full gradient of 90 min). For isobaric labeling, the effective
297 gradient was extended to 100 min (full gradient 120 min). Details of gradient can be found in SI Appendix.
298 3 compensation voltages and MS^3-based quantification were used, generating throughput of 100 cells per
299 day at our set-up with blank control channels (which may be improved to 140 cells per day with TMTpro
300 18-plex reagent). Details of gradients are described in SI Appendix Supplementary Information Text.
301 Peptides were analyzed by a tribrid mass spectrometer (Orbitrap Eclipse, ThermoFisher) with an ion
302 mobility interface (FAIMS Pro, ThermoFisher). For label-free pipeline, FAIMS compensation voltages of -
303 55 V and -70 V were used. Cycle time of 1 s was used for both CVs. MS spectra were acquired by
304 Orbitrap analyzer, and MS2 spectra were acquired by linear ion trap analyzer. Max ion injection time for
305 MS2 was 200 ms. For isobaric labeling pipeline, FAIMS compensation voltages of -40, -60, and -80 V
306 were used. Cycle time of 1 s was used for all CVs. MS spectra were acquired by Orbitrap analyzer. MS^2
307 spectra were acquired by linear ion trap analyzer. A real-time search (RTS) node was added to decide if
308 MS^2 spectra were from human tryptic peptides that had been chemically labeled by TMTpro. If a peptide
309 was identified, a synchronous precursor selection (SPS)-MS^3 scan was acquired to quantify that peptide
310 in different TMTpro channels using Orbitrap as analyzer.

311 **Proteomic data analysis**

312 MS raw files were searched against UniProt human protein database (20286 reviewed accessions,
313 downloaded on April 14, 2020, for searches of mouse oocytes, a FASTA database with 17015 reviewed
314 accessions downloaded on July 31, 2020 were used instead) and an in-house curated contamination
315 database (284 accessions) by Proteome Discoverer 2.4 (ThermoFisher). For label-free quantification,
316 peak intensity of MS feature was used as signal of quantification and match-between-runs (MBR) feature
317 was used. For isobaric labeling quantification, signal-to-noise ratio of MS^3 spectrum was used as signal of
318 quantification.

319 **Generation of correlated protein modules**

320 The abundances of all proteins in quantified cells were further filtered to 1) remove cells with extremely
321 high level of contamination; 2) remove protein contaminants; 3) remove proteins with more than 50%
322 missing values in all cells; 4) remove outlier cells (by fitting to a Gaussian distribution). Protein abundance
323 of single cells was normalized by total protein intensity. The missing values were filled with zeros.
324 Pearson correlation of proteins were calculated by numpy.corrcoef of NumPy (42). The correlation map
325 was reordered using unweighted pair group method with arithmetic means clustering using the SciPy (43)
326 package. One CPM is defined as the biggest protein groups of the correlation map, who has over 80
327 percent of protein pairs with correlation coefficient over 0.3. List of correlated protein modules can be
328 found in SI Appendix, Dataset S1. We define the abundance of a CPM in a cell by summing up the
329 abundances of proteins included in the very CPM.

330 **Bioinformatic analysis**

331 Bioinformatical analysis was conducted with R. Principle component analysis was conducted by “PCA”
332 function in “FactoMineR” package with default parameters. Proteins whose average abundance changes
333 over two folds between two subpopulations with p-value < 0.05 (Wilcoxon test) were defined as
334 differentially expressed proteins. Gene Ontology enrichment and KEGG pathway enrichment was done
335 by R package “clusterProfiler” (v.4.0.2). The adjusted p-value cutoff was set as 0.05, and only GO terms
336 with more than 2 genes were kept for further analysis. Reactome pathway enrichment was done by R
337 package “ReactomePA”. The p-value cutoff was set as 0.05, and only terms over 2 genes were kept.
338 Protein-protein interaction ratio was calculated based on R package “STRINGdb” (v.11). The combined
339 score cutoff was 250. The ratio value was calculated as the proportion of known PPI pairs to the total
340 number of protein pairs in a CPM. CORUM version 3.0 dataset was downloaded at “http://mips.helmholtz-
341 muenchen.de/corum/”. Result of enrichment analysis of correlated protein modules can be found in SI
342 Appendix, Dataset S1.

343 **Simulation of effect of detectability on measured correlation and real correlation**

344 The simulation was conducted with R. Random RNA abundances were generated by Poisson
345 distributions with median of 20 copies. Random protein abundances were generated by negative bimodal
346 distributions with median of 10^5 and CV of 200. Then a normal-distributed random noise was added to
347 RNA and protein. Finally, the binominal distributed detectability was multiplied to original data. All
348 simulations were repeated 100 times to reduce random noise.
349 To assess the measurement noise of single-cell RNA analysis and single-cell proteins analysis,
350 simulation was conducted with R. Random RNA abundances of 1000 cells were generated by Poisson

351 distributions with median of 17 copies. Random protein abundances of 100 cells were generated by

352 negative bimodal distributions with median of $5*10^4$ and CV of 0.01. The binominal distributed

353 detectability was multiplied to original data.

354 To assess the accuracy of correlation measurement of RNA and proteins, simulation was conducted with

355 R. Random RNA abundances of 1000 cells were generated by Poisson distributions with median of 20

356 copies. Random protein abundances of 100 cells were generated by negative bimodal distributions with

357 median of 10^5 and CV of 0.01. Then a normally distributed random noise was added to RNA and protein.

358 Finally, the binominal distributed detectability was multiplied to original data. All simulations were

359 repeated 100 times to reduce random noise.

360 **Comparison of CPMs with CGMs and ProteomeHD**

361 The correlation coefficient and CGMs of 12,660 genes in 2,302 K562 cells is analyzed as described in (6).

362 The K562 dataset can be found at [data repository to be added]. When comparing CPMs and CGMs,

363 because there are multiple transcripts corresponding to the same protein, the number of proteins and

364 genes in **Fig. 4a** are not the same. ProteomeHD have correlation score no more than 0.3. To compare

365 with our Pearson correlation coefficient, ProteomeHD data was scaled by making its maximum correlation

366 score equal to 1.

367

368 **Acknowledgements**

369 This work was supported by Beijing Advanced Innovation Center for Genomics at Peking University and

370 Beijing Municipal Commission of Science and Technology (Z201100005320015). We thank Alec R.

Chapman, David F. Lee, Fuchou Tang and Ge Gao for their inspiring work in gene modules by high-precision single-cell transcriptomics. We thank Alec R. Chapman and Jiekun Yang for their involvement in the early attempt on single-cell proteomics of our group. We thank Bogdan Budnik and Nikolai Slavov for their help on our implementation of SCoPE-MS. We thank Xiaoran Chai and Wenjie Sun for helpful discussions of the manuscript.

Author Contributions: M.H. and X.S.X. designed research; M.H., Y. Zhang, Y.Y. and Y. Zheng performed research; M.H., Y. Zhang, Y.Y. and X.S.X. analyzed data; and M.H., Y. Zhang, Q.G. and X.S.X. wrote the paper with the help of other authors.

Competing Interest Statement: The authors do not disclose any competing interests.

383 **References**

384 1. L. Cai, C. K. Dalal, M. B. Elowitz, Frequency-modulated nuclear localization bursts coordinate gene
385 regulation. *Nature* **455**, 485–490 (2008).

386 2. A. J. M. Larsson, *et al.*, Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254
387 (2019).

388 3. Y. Taniguchi, *et al.*, Quantifying *E. coli* Proteome and Transcriptome with Single-Molecule
389 Sensitivity in Single Cells. *Science* **329**, 533–538 (2010).

390 4. S. Chong, C. Chen, H. Ge, X. S. Xie, Mechanism of Transcriptional Bursting in Bacteria. *Cell* **158**,
391 314–326 (2014).

392 5. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC*
393 *Bioinformatics* **9**, 559 (2008).

394 6. A. R. Chapman, *et al.*, Correlated Gene Modules Uncovered by Single-Cell Transcriptomics with
395 High Detectability and Accuracy. *bioRxiv*, 2019.12.31.892190 (2020).

396 7. F. Tian, *et al.*, Genomic Architecture of Cells in Tissues (GeACT): Study of Human Mid-gestation
397 Fetus. *bioRxiv*, 2020.04.12.038000 (2020).

398 8. R. Aebersold, M. Mann, Mass-spectrometric exploration of proteome structure and function. *Nature*
399 **537**, 347–355 (2016).

400 9. Y. Liu, A. Beyer, R. Aebersold, On the Dependency of Cellular Protein Levels on mRNA
401 Abundance. *Cell* **165**, 535–550 (2016).

402 10. F. Meissner, R. A. Scheltema, H.-J. Mollenkopf, M. Mann, Direct Proteomic Quantification of the
403 Secretome of Activated Immune Cells. *Science* **340**, 475–478 (2013).

404 11. Y. Taniguchi, *et al.*, Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity
405 in single cells. *Science* **329**, 533–538 (2010).

406 12. H. Specht, N. Slavov, Transformative Opportunities for Single-Cell Proteomics. *J. Proteome Res.*
407 **17**, 2565–2571 (2018).

408 13. M. Mann, The Origins of Organellar Mapping by Protein Correlation Profiling. *Proteomics* **20**,
409 1900330 (2020).

410 14. J. S. Andersen, *et al.*, Proteomic characterization of the human centrosome by protein correlation
411 profiling. *Nature* **426**, 570–574 (2003).

412 15. L. J. Foster, *et al.*, A Mammalian Organelle Map by Protein Correlation Profiling. *Cell* **125**, 187–199
413 (2006).

414 16. M. A. Skinnider, *et al.*, An atlas of protein-protein interactions across mouse tissues. *Cell* **184**, 4073-
415 4089.e17 (2021).

416 17. G. Kustatscher, *et al.*, Co-regulation map of the human proteome enables identification of protein
417 functions. *Nat Biotechnol* **37**, 1361–1371 (2019).

418 18. B. Budnik, E. Levy, G. Harmange, N. Slavov, SCoPE-MS: mass spectrometry of single mammalian
419 cells quantifies proteome heterogeneity during cell differentiation. *Genome Biology* **19**, 161 (2018).

420 19. H. Specht, *et al.*, Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity
421 using SCoPE2. *Genome Biol* **22**, 50 (2021).

422 20. Y. Cong, *et al.*, Ultrasensitive single-cell proteomics workflow identifies >1000 protein groups per
423 mammalian cell. *Chem. Sci.* **12**, 1001–1006 (2021).

424 21. J. Woo, *et al.*, Three-dimensional feature matching improves coverage for single-cell proteomics
425 based on ion mobility filtering. *Cell Systems* **13**, 426-434.e4 (2022).

426 22. M. Giurgiu, *et al.*, CORUM: the comprehensive resource of mammalian protein complexes—2019.
427 *Nucleic Acids Research* **47**, D559–D563 (2019).

428 23. M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*
429 **28**, 27–30 (2000).

430 24. A. Fabregat, *et al.*, The Reactome Pathway Knowledgebase. *Nucleic Acids Research* **46**, D649–
431 D655 (2018).

432 25. X. Chen, *et al.*, DNA methylation directly downregulates human cathelicidin antimicrobial peptide
433 gene (CAMP) promoter activity. *Oncotarget* **8**, 27943–27952 (2017).

434 26. D. Szklarczyk, *et al.*, STRING v10: protein-protein interaction networks, integrated over the tree of
435 life. *Nucleic Acids Res.* **43**, D447-452 (2015).

436 27. A. Murayama, *et al.*, Epigenetic Control of rDNA Loci in Response to Intracellular Energy Status.
437 *Cell* **133**, 627–639 (2008).

438 28. F. Lévi, A. Okyar, S. Dulong, P. F. Innominato, J. Clairambault, Circadian Timing in Cancer
439 Treatments. *Annual Review of Pharmacology and Toxicology* **50**, 377–421 (2010).

440 29. A. Regev, *et al.*, The Human Cell Atlas. *Elife* **6** (2017).

441 30. L. Yan, *et al.*, Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem
442 cells. *Nat Struct Mol Biol* **20**, 1131–1139 (2013).

443 31. Q. Deng, D. Ramsköld, B. Reinius, R. Sandberg, Single-Cell RNA-Seq Reveals Dynamic, Random
444 Monoallelic Gene Expression in Mammalian Cells. *Science* **343**, 193–196 (2014).

445 32. D. Kotliar, *et al.*, Identifying gene expression programs of cell-type identity and cellular activity with
446 single-cell RNA-Seq. *eLife* **8**, e43803 (2019).

447 33. A. Brunner, *et al.*, Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes
448 upon perturbation. *Molecular Systems Biology* **18** (2022).

449 34. Y. Zhu, *et al.*, Nanodroplet processing platform for deep and quantitative proteome profiling of 10-
450 100 mammalian cells. *Nat Commun* **9**, 882 (2018).

451 35. Z.-Y. Li, *et al.*, Nanoliter-Scale Oil-Air-Droplet Chip-Based Single Cell Proteomic Analysis. *Anal.*
452 *Chem.* **90**, 5430–5438 (2018).

453 36. J. Woo, *et al.*, High-throughput and high-efficiency sample preparation for single-cell proteomics
454 using a nested nanowell chip. *Nat Commun* **12**, 6246 (2021).

455 37. Y. Cong, *et al.*, Improved Single-Cell Proteome Coverage Using Narrow-Bore Packed NanoLC
456 Columns and Ultrasensitive Mass Spectrometry. *Anal. Chem.* **92**, 2665–2671 (2020).

457 38. J. Stadlmann, *et al.*, Improved Sensitivity in Low-Input Proteomics Using Micropillar Array-Based
458 Chromatography. *Anal. Chem.* **91**, 14203–14207 (2019).

459 39. C. Wichmann, *et al.*, MaxQuant.Live Enables Global Targeting of More Than 25,000 Peptides.
460 *Molecular & Cellular Proteomics* **18**, 982–994 (2019).

461 40. A. T. Chen, A. Franks, N. Slavov, DART-ID increases single-cell proteome coverage. *PLOS*
462 *Computational Biology* **15**, e1007082 (2019).

463 41. S.-H. Yu, P. Kyriakidou, J. Cox, Isobaric Matching between Runs and Novel PSM-Level
464 Normalization in MaxQuant Strongly Improve Reporter Ion-Based Quantification. *J. Proteome Res.*
465 **19**, 3945–3954 (2020).

466 42. C. R. Harris, *et al.*, Array programming with NumPy. *Nature* **585**, 357–362 (2020).

467 43. P. Virtanen, *et al.*, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat*
468 *Methods* **17**, 261–272 (2020).

469
470

Figures and Tables

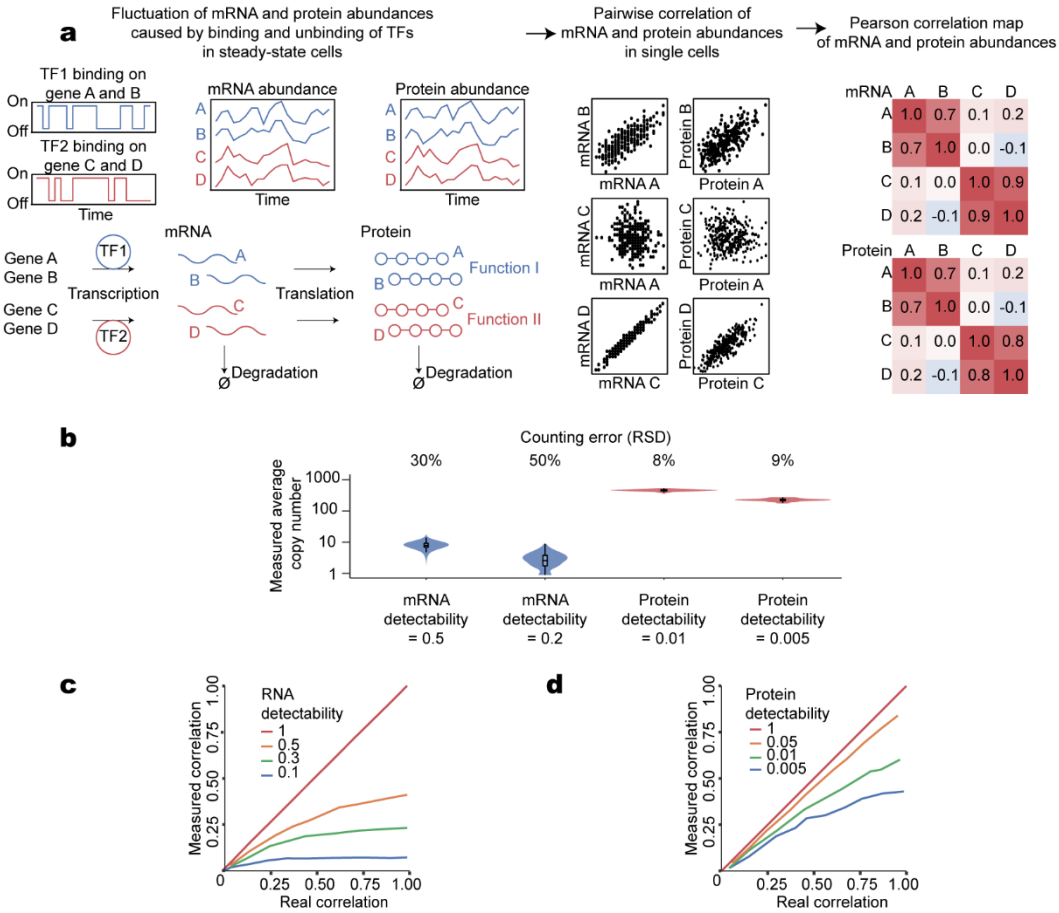

Figure 1. Direct measurement of protein correlations at single-cell level reveals the intrinsic dynamics of the cell. a. Measurement of correlations of protein abundances in single cells reveals coregulation in steady-state gene expression. Level of mRNAs and proteins fluctuate stochastically in steady-state cells, mostly caused by fluctuation of transcription factors (TFs). The expression of genes with similar biological functions tend to be coregulated, resulting in highly positive correlations of the abundances of corresponding mRNAs and proteins. Independently regulated proteins, on the other hand, will not exhibit high correlations. These correlations can be assessed by measuring abundances of mRNAs and proteins in many single cells. Level of mRNAs and proteins and their correlations are calculated from simulated

481 theoretical distributions without considering detectability. b. The copy number and detectability affect the

482 distribution of the measured copy number of protein and mRNA, which results in different counting errors.

483 c. Simulation of the effect of detectability on the measured correlation and real correlation in single cells

484 for mRNA. d. Simulation of the effect of detectability on the measured correlation and real correlation in

485 single cells for protein.

486

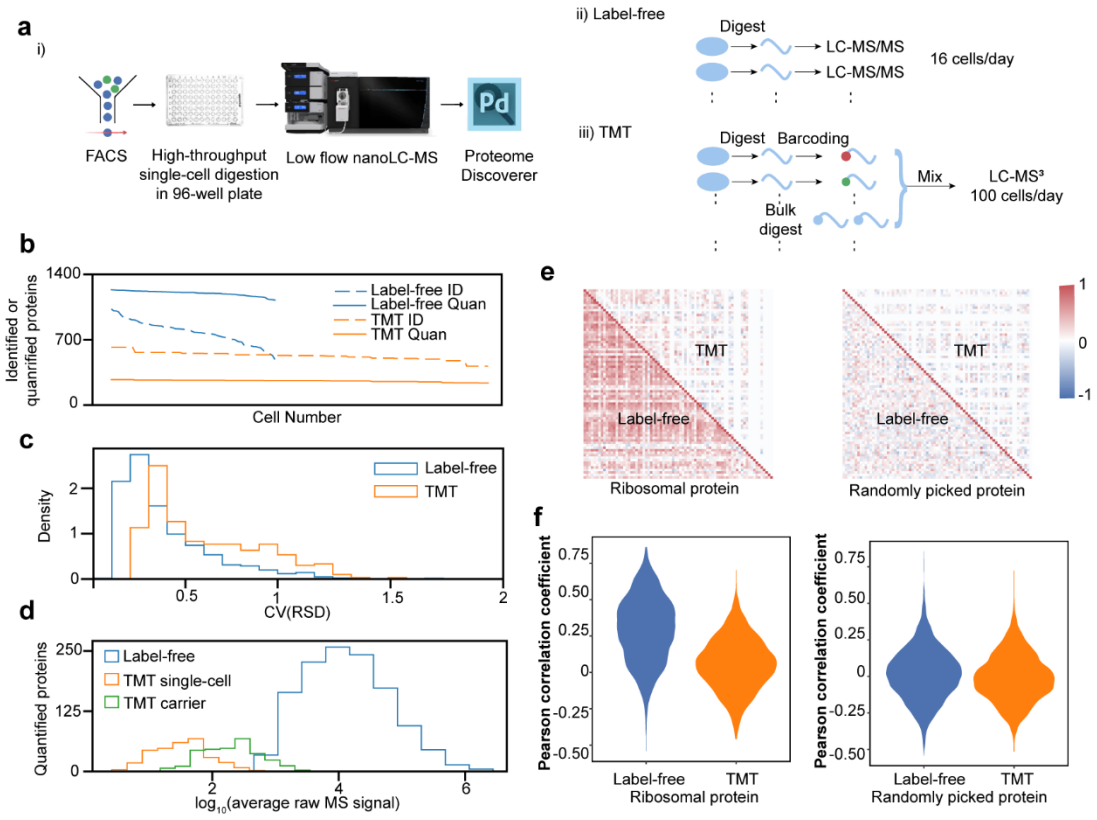

Figure 2. A single-cell proteomics pipeline for accurate protein correlation measurement. a. Flowchart of pipelines we used to measure protein-protein correlations in single-cells. label-free: label-free pipeline. TMT: isobaric labeling pipeline. b. Identified and quantified proteins from single cells by the label-free (*N* = 71) and TMT (*N* = 161) pipelines. Cells were sorted by their numbers of identified and quantified proteins. c. Distribution of CV (in RSD) of proteins by the label-free (*N* = 1,257) and TMT (*N* = 300) pipelines. d. Distribution of raw MS signals of quantified proteins by the label-free (*N* = 1,257) and TMT (*N* = 300) pipelines. e. Comparison of correlation coefficients of ribosomal proteins and randomly picked proteins of the same size. Number of proteins: label-free *N* = 77, TMT *N* = 53. f. Violin plot of correlation coefficients of ribosomal proteins and randomly picked proteins. Number of proteins: label-free *N* = 77, TMT *N* = 53.

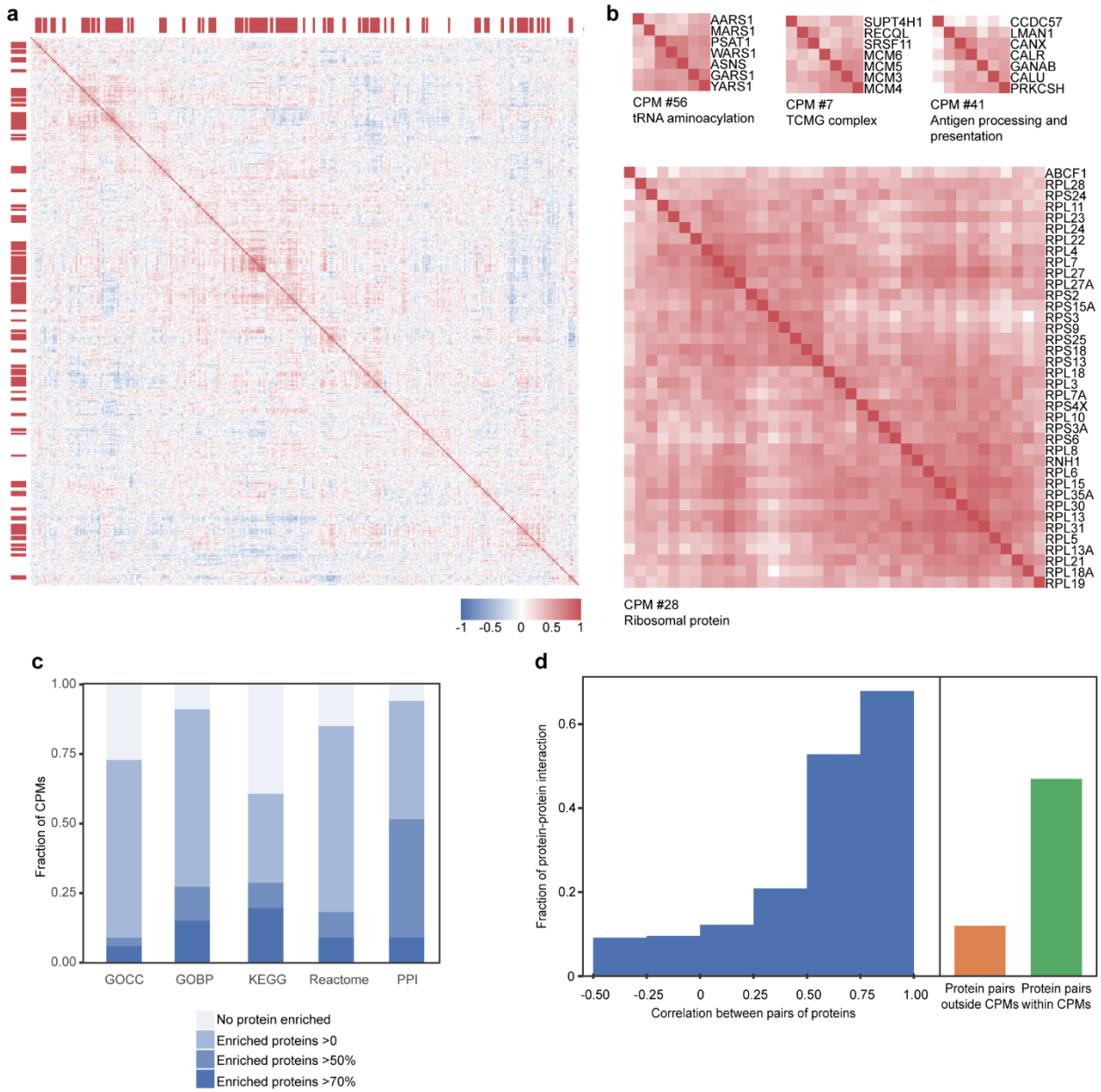

Figure 3. Correlated protein modules of K562 cell line by single-cell proteomics. a. Correlation map of 1,249 proteins in 69 K562 single cells. Red blocks indicate that corresponding proteins were defined as a CPM. b. Representation of four CPMs and enriched function terms of them. c. Fraction of CPMs with different enrichment results of Gene Ontology cellular component (GOCC), Gene Ontology biological processes (GOBP), KEGG, Reactome and protein-protein interaction (PPI) of proteins in each CPM.

504 CPMs were categorized by their percentage of proteins in enriched functional term (enriched proteins)

505 over proteins in corresponding CPM. d. Fraction of protein pairs with known PPIs in the STRING

506 database at different levels of correlation coefficients (blue), outside CPMs (orange), and within CPMs

507 (green).

508

509

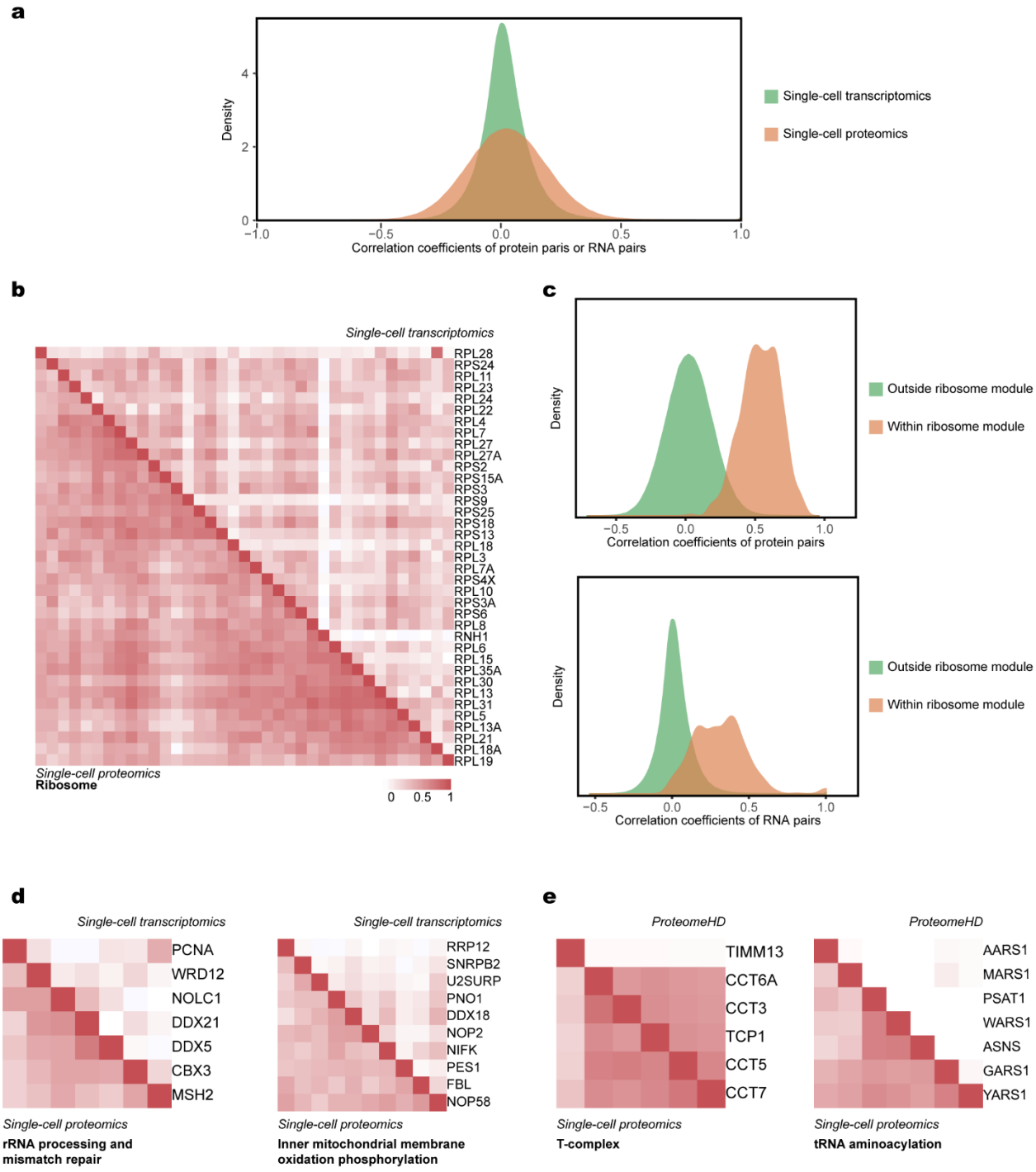

Figure 4. Correlated protein modules by single-cell proteomics shows unique characteristics. a. Distribution of pairwise correlation coefficients of single-cell transcriptomics (N = 1,275) and single-cell proteomics (N = 1,249). b. Comparison of protein (bottom-left) and mRNA (top-right) correlation

514 coefficients for the ribosome modules of K562. Correlation coefficients between proteins are generally
515 stronger than those between corresponding mRNAs. c. Distribution of correlation coefficients of protein
516 pairs (top) or RNA pairs (bottom) that within the ribosome module (*N* = 1,332 for proteins, *N* = 1,351 for
517 RNAs) and outside the ribosome module (*N* = 1,558,752 for proteins, *N* = 1,624,981 for RNAs). The
518 correlation coefficients of functional protein pairs are more distinguishable from the overall pairwise
519 correlations. d. Comparison of protein (bottom-left) and mRNA (top-right) correlation coefficients for the
520 rRNA processing and mismatch repairment and inner mitochondrial membrane oxidation phosphorylation
521 modules of K562. Correlation coefficients between proteins are generally stronger than those between
522 corresponding mRNAs. e. Comparison of correlation coefficients in CPM (bottom-left) and scaled
523 correlation scores from ProteomeHD (top-right) for the T-complex composition and tRNA aminoacylation
524 modules. CPM clearly exhibits correlations that cannot be seen in the bulk data obtained via perturbation.

525

526

527