Gui *et al. Genome Biology* (2024) 25:267
https://doi.org/10.1186/s13059-024-03382-9

Genome Biology

METHOD **Open Access**

zMAP toolset: model-based analysis of large-scale proteomic data via a variance stabilizing *z*-transformation

Xiuqi Gui[1†], Jing Huang[1†], Linjie Ruan[2†], Yanjun Wu[3†], Xuan Guo[1†], Ruifang Cao[1], Shuhan Zhou[1], Fengxiang Tan[1], Hongwen Zhu[4], Mushan Li[5], Guoqing Zhang[1], Hu Zhou[4], Lixing Zhan[3*], Xin Liu[2*], Shiqi Tu[1*] and Zhen Shao[1*]

†Xiuqi Gui, Jing Huang, Linjie Ruan, Yanjun Wu and Xuan Guo contributed equally to this work.

*Correspondence: lxzhan@sinh.ac.cn; xin.liu@sibcb.ac.cn; tushiqi@sinh.ac.cn; shaozhen@sinh.ac.cn

[1] CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2] Key Laboratory of Epigenetic Regulation and Intervention, Shanghai Institute of Biochemistry and Cell Biology, CAS Center for Excellence in Molecular Cell Science, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[3] CAS Key Laboratory of Nutrition, Metabolism and Food Safety, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
Full list of author information is available at the end of the article

Abstract

Isobaric labeling-based mass spectrometry (ILMS) has been widely used to quantify, on a proteome-wide scale, the relative protein abundance in different biological conditions. However, large-scale ILMS data sets typically involve multiple runs of mass spectrometry, bringing great computational difficulty to the integration of ILMS samples. We present zMAP, a toolset that makes ILMS intensities comparable across mass spectrometry runs by modeling the associated mean-variance dependence and accordingly applying a variance stabilizing z-transformation. The practical utility of zMAP is demonstrated in several case studies involving the dynamics of cell differentiation and the heterogeneity across cancer patients.

Background

Isobaric labeling-based mass spectrometry (ILMS) methods, such as tandem mass tagging (TMT) [1] and isobaric tags for relative and absolute quantitation (iTRAQ) [2, 3], have been widely used for quantitative proteomic profiling, showing unique advantages in a broad range of biological studies [4–9]. These methods can profile the abundance of thousands of proteins in parallel and allow, with the high multiplexing capability (up to 16), the evaluation of many different biological conditions in a single run of mass spectrometry (MS) [10, 11]. In recent years, large-scale proteomic studies have become prevalent in which tens or even hundreds of ILMS samples were generated. The large sample sizes have facilitated various studies, such as dissecting the dynamics during cell differentiation and the heterogeneity across a large cohort of cancer patients [12–15]. Meanwhile, they have brought new computational challenges to the analysis of ILMS data.

A fundamental problem is that, in a large-scale proteomic study, performing multiple MS runs is usually inevitable for achieving the required sample size. With the current

ILMS techniques, however, absolute signal intensities from different MS runs are far from comparable, posing a huge difficulty for quantitatively integrating ILMS data across MS runs [16–18].

Previous studies have developed statistical methods for the differential analysis of ILMS data that are generated from multiple MS runs [19–21]. For example, MSstatsTMT treats MS runs and biological conditions as random and fixed effects, respectively, and it identifies differentially expressed proteins between conditions by fitting linear mixed-effects models [19]; msTrawler adopts the same regression technique and further accounts for the signal-to-noise ratios of different measurements [21]. These methods, however, are specifically designed for the traditional differential analysis and cannot allow the application of other integration analyses across MS runs, such as principal component analysis (PCA) and unsupervised clustering analysis of samples/proteins.

Other studies tackled this problem by adding a biologically identical reference sample to each MS run, which was typically designed to be a uniform mixture of protein extracts from many different materials. Then, the $\log_2$-ratios (M-values) of ILMS intensities of each sample to those of the corresponding reference sample were calculated and were considered as comparable across MS runs. This strategy has been employed by many cancer studies to combine the proteomic profiles of a large number of patients, with the purpose of systematically identifying differentially expressed proteins between tumor tissues and normal tissues adjacent to the tumor (NATs), as well as proteins with hypervariable expression across tumors [12, 14, 22, 23]. However, these studies did not take the mean-variance dependence [18, 24–26] of ILMS data into account, which may compromise the reliability of downstream analysis results. For example, several studies used the standard deviation of M-values across samples to rank proteins. Top-ranked proteins were then selected as hypervariable ones for exploring the similarity structure among the samples [22, 27]. However, the M-values derived from low intensity levels could be much more unstable than those derived from high intensities, giving rise to a bias towards selecting low-abundance proteins [10, 28]. Besides, the exact number of proteins to be selected could only be determined based on practical experience, owing to the lack of a statistical model for assessing the associated significances.

We previously developed MAP for identifying differentially expressed proteins between a pair of ILMS samples generated from the same MS run [29]. In this study, we present zMAP, a computational toolset that significantly extends the capability of MAP to accommodate to features of large-scale data sets. The key step of zMAP is to calculate the M-values of each sample against the corresponding reference sample and scale the results by estimated signal variability, which is obtained by modeling the mean-variance dependence of ILMS intensities separately for each MS run. The primary rationale behind this variance stabilizing z-transformation is using relative abundance compared to a biological invariant to improve the comparability across MS runs, while accounting for the different reliability of M-values derived from different intensity levels.

Results

Workflow of zMAP module

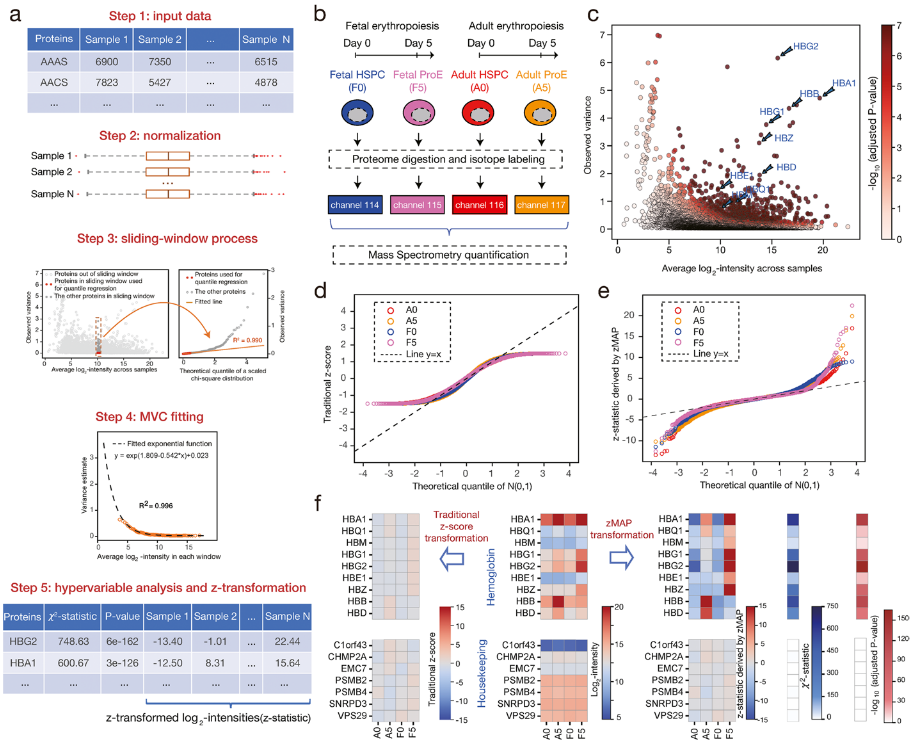
Given a large-scale ILMS data set, zMAP toolset transforms each ILMS intensity into a z-statistic that essentially assesses the statistical significance of the deviation of this measurement from that (of the same protein) in the corresponding reference sample. The final *p*-value for a certain protein is derived by integrating the associated statistical significances in all samples. For this *p*-value to make biological sense, the reference samples in all MS runs must be biologically identical, and the corresponding null hypothesis is no differential expression between each involved condition and the biological state represented by the reference samples. In practice, reference samples are almost always designed to be the average of all the involved conditions. In this case, the corresponding null hypothesis amounts to no differential expression between each pair of the conditions, and we refer to significant proteins as hypervariable ones.

Two computational modules, named zMAP and reverse-zMAP, have been implemented in the zMAP toolset for handling different scenarios. We first introduce the former, which is specifically developed for ILMS data sets without real reference samples but whose MS runs are all associated with the same biological composition (i.e., they are biological replication of each other).

The basic principle of zMAP module is considering the average of the ILMS samples from each MS run as a pseudo reference profile for these samples. Then, the M-values of each sample against the corresponding reference profile are scaled based on estimated signal variances, producing *z*-statistics for a final integration across MS runs. Unlike traditional z-score transformation, which is based on observed variance, zMAP module derives the variance estimates by borrowing strength between proteins with similar intensity levels and fitting a smooth mean-variance curve (MVC) for each MS run. Figure 1a depicts the workflow of zMAP module on a single MS run. In detail, after normalizing protein-level ILMS intensities across multiple samples generated by the same MS run, zMAP makes a $\log_2$ transformation and models the results as following normal distributions, with the mean and variance parameters linked by an unknown MVC. To fit this MVC, a mean-variance scatter plot for all proteins is drawn, and zMAP uses a sliding window to scan this plot. In this process, the proteins with similar average $\log_2$-intensities across samples are grouped together, and a variance estimate is derived for each individual window. For the latter part, zMAP performs a quantile regression that regresses the observed variances of included proteins against a scaled chi-square distribution by fitting a straight line through the origin, and the slope is taken as the variance estimate. To avoid the influence of potential hypervariable proteins, which can lead to an overestimated variance, only a certain proportion of the proteins with the smallest observed variances are used for the quantile regression. Finally, the MVC is fitted by regressing the variance estimates from all windows against the corresponding intensity levels. Notably, this sliding-window process equipped with the robust quantile regression can avoid the requirement of biological replicates within each single MS run for MVC fitting.

To demonstrate the practical utility of zMAP module, we applied it to an iTRAQ data set regarding human erythropoiesis at both fetal and adult stages [30]. This data set was comprised of five replicate MS runs, each of which generated four proteomic profiles corresponding to hematopoietic stem/progenitor cells (HSPCs) and proerythroblasts (ProEs) at fetal and adult stages (labeled as F0, F5, A0, and A5; Fig. 1b).

We first applied zMAP to the first MS run. In practical analysis, an essential parameter for the sliding-window process, denoted by W, is the exact proportion of proteins in



Proteins	Sample 1	Sample 2	...	Sample N
AAAS	6900	7350	...	6515
AACS	7823	5427	...	4878
...	...	...	...	...

Proteins	χ^2-statistic	P-value	Sample 1	Sample 2	...	Sample N
HBG2	748.63	6e-162	-13.40	-1.01	...	22.44
HBA1	600.67	3e-126	-12.50	8.31	...	15.64
...	...	...	...	...	...	...

Fig. 1 Applying zMAP module to ILMS samples generated by a single MS run. **a** The workflow of zMAP module. **b** The experiment design of each MS run for the iTRAQ data set regarding human erythropoiesis at fetal and adult stages. HSPC, hematopoietic stem and progenitor cell; ProE, proerythroblast. **c** Mean-variance scatter plot of all proteins, with colors indicating the BH-adjusted *p*-values derived for identifying hypervariable proteins. All detected hemoglobin subunits are explicitly marked. **d**, **e** For each sample, the associated (**d**) traditional *z*-scores or (**e**) *z*-statistics derived by zMAP are plotted against the corresponding theoretical quantiles of the standard normal distribution. **f** For the hemoglobin subunits and a list of housekeeping proteins, heat maps showing their traditional *z*-scores and *z*-statistics derived by zMAP in all samples. The associated chi-square statistics and BH-adjusted *p*-values are also displayed

each window that are used to perform the quantile regression. The setting of W depends directly on the abundance of differentially expressed proteins across samples (Additional file 2: Note S1). Here, we tried the default setting (W = 30%) and assessed the goodness of fit for each window by calculating the R-squared (R^2) statistic. All the R^2 values were above 0.95 (Additional file 1: Fig. S1a). Besides, zMAP also generated a diagnostic plot to summarize the quantile-quantile plots of all windows. Specifically, the observed variances associated with each window were first ordered and scaled by the corresponding variance estimate (i.e., the slope of the fitted line). Then, the scaled variances were averaged across all windows (note that all windows covered the same number of proteins). Finally, the results, together with error bars to indicate the variability across windows, were plotted against the theoretical quantiles (Additional file 1: Fig. S1b). We found that the scaled variances that had been used for the quantile regressions matched well with the line $y = x$, while the other scaled variances, especially the largest ones, deviated clearly from the line, suggesting the default setting of W was suitable in this case. In practice, this diagnostic plot can be useful for fine-tuning W.

After the sliding-window process, the variance estimates and average $\log_2$-intensities of all windows were subject to a regression procedure to fit the MVC. Based on previous studies of iTRAQ data [24, 31, 32], an exponential decay function was fitted, with an R^2 of 0.996. zMAP next identified hypervariable proteins among the four conditions by calculating the ratio of the observed variance of each protein to the corresponding variance indicated by the MVC, which was obtained by applying the exponential function to the average $\log_2$-intensity of the protein. This ratio followed a scaled chi-square distribution under the null hypothesis of no differential expression between each pair of conditions, and a p-value was accordingly derived. As an evaluation of zMAP, we specifically examined the results regarding all known subunits of hemoglobin (the *HBA2* subunit was excluded because none of the five MS runs had detected it). It was found that zMAP deemed all these subunits as significant hypervariable proteins (BH-adjusted p-value $< 1\mathrm{e}-7$ for each of them; Fig. 1c).

Finally, zMAP transformed the $\log_2$-intensities of each protein by subtracting its average $\log_2$-intensity and dividing the results by the standard deviation implied by the MVC. The only difference between this z-transformation and the traditional z-score transformation is that the latter uses observed standard deviation as the scaling factor, which may cause a compression of biologically meaningful signal changes across samples. Here, we tried both transformations and aligned the results with the standard normal distribution separately for each sample (Fig. 1d, e). The empirical distributions of traditional z-scores were even more concentrated around zero than the standard normal distribution, while those of the z-statistics derived by zMAP were relatively more long-tailed, suggesting an improved statistical power for identifying hypervariable proteins. We further examined the transformation results of the hemoglobin proteins and a list of housekeeping proteins (Fig. 1f). For the traditional transformation, the intensity differences among samples were considerably compressed for both classes of proteins. In comparison, the zMAP transformation dramatically increased the sensitivity to biologically meaningful intensity differences without sacrificing the specificity. Specifically, all the hemoglobin proteins were upregulated during human erythropoiesis at fetal and/or adult stage. Consistent with previous research [33, 34], *HBE1*, *HBG1*, *HBG2*, and *HBZ* were mainly expressed at the fetal stage; *HBB* and *HBD* were mainly expressed at the adult stage; *HBA1* was highly expressed at both stages. We applied zMAP separately to each of the other four MS runs and found similar results (Fig. 2a).

Integrating proteomic profiles across MS runs and detecting proteins important for human fetal erythropoiesis

A hierarchical clustering of all samples from the total five MS runs was performed based on the Pearson correlation coefficient (PCC) between each pair of them. When the PCCs were calculated by using original (untransformed) $\log_2$-intensities, the samples were perfectly clustered by their MS runs of origin, indicating severe batch effects (Additional file 1: Fig. S2a). By contrast, the samples were clustered by their biological labels when the z-statistics derived by zMAP were used (Additional file 1: Fig. S2b).

We next benchmarked zMAP against two other methods for integrating these samples, which were respectively based on traditional z-scores and centered $\log_2$-intensities. The latter was equivalent to the M-values of each sample against the corresponding pseudo reference profile. Similar to the z-statistics of zMAP, both traditional z-scores and centered $\log_2$-intensities have led to a hierarchical clustering result that was perfectly

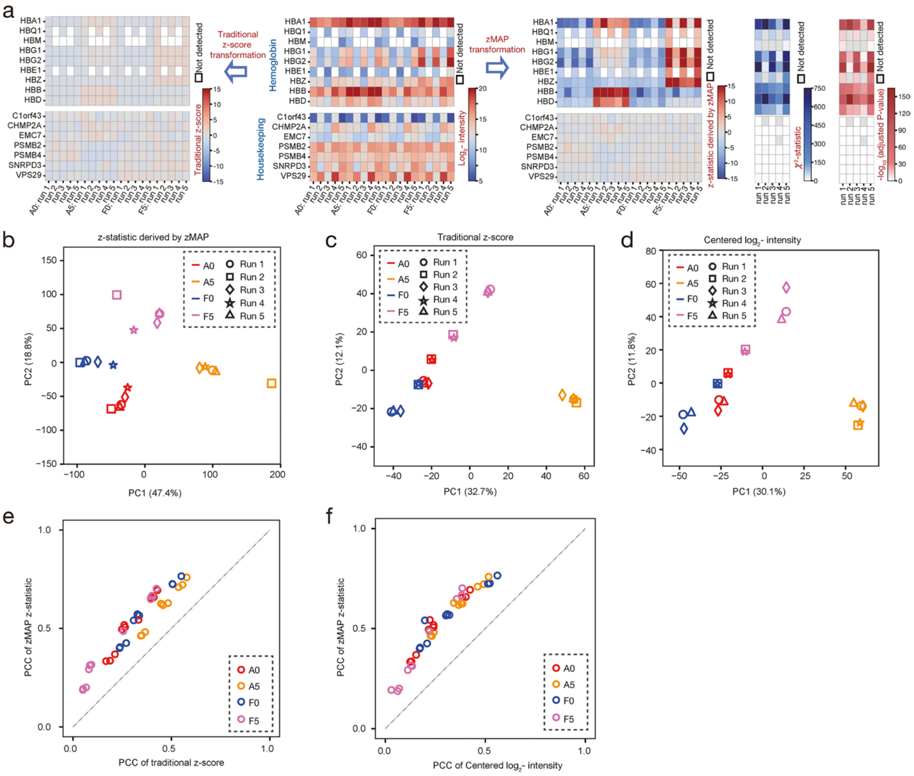


Fig. 2 Using different statistics to integrate ILMS samples across MS runs. **a** For the hemoglobin subunits and the housekeeping proteins, heat maps showing their traditional *z*-scores and *z*-statistics derived by zMAP for all the five MS runs. The associated chi-square statistics and BH-adjusted *p*-values are also displayed. **b**–**d** Two-dimensional PC plots generated by performing PCA on all samples from all MS runs. Each plot corresponds to a kind of measurements for integrating ILMS samples across MS runs. **e**, **f** For each kind of measurements, the PCC between each pair of samples with the same biological label was calculated. The PCCs derived from the *z*-statistics of zMAP are compared to those from (**e**) traditional *z*-scores and (**f**) centered $\log_2$-intensities

consistent with the biological labels of samples, suggesting an effective removal of batch effects (Additional file 1: Fig. S2c, d).

We performed PCA on all the samples by using separately the measurements provided by each method. It was found that the samples were well clustered by their biological labels in the two-dimensional PC plot generated by zMAP (Fig. 2b). Moreover, the F0 and A0 clusters were relatively close to each other, and two separate differentiation trajectories starting from them could be depicted in the plot, corresponding to HSPC-to-ProE differentiation at fetal and adult stages respectively. In comparison, the PC plots produced by the other two methods did not distinguish between different biological labels as clearly as did the zMAP plot (Fig. 2c, d). In particular, the F0 and A0 samples tended to mix with each other.

For a more quantitative evaluation of the methods, we assessed the consistency between each pair of samples with the same biological label, based on the measurements provided by each method. Each biological label was associated with five samples from five different MS runs, and we calculated the PCC between each pair of them. In each case, the PCC derived based on the *z*-statistics of zMAP was higher than those from the other two methods (Fig. 2e, f).

As a general technical problem of iTRAQ experiments, ratio compression has been suggested to arise from contamination during precursor ion selection by co-eluting peptides [24, 35]. These peptides contribute a background value equally to each reporter ion signal, leading to an increase in the intensity levels and a shrinkage of observed intensity fold changes across samples. For the same protein, the influence of ratio compression can be different across MS runs, resulting in different degrees to which the observed variance of $\log_2$-intensities is diminished. In the data set about human erythropoiesis, we observed, for specific proteins, that the observed variances in different MS runs were strongly negatively correlated with the average $\log_2$-intensities, and that such correlation was effectively eliminated when the observed variances were scaled based on the corresponding MVCs (Additional file 1: Fig. S3a, b). Globally, the median PCC (for all proteins) between observed variances and average $\log_2$-intensities was − 0.34, and it became 0.09 when the scaled variances were used (Additional file 1: Fig. S3c).

We next used zMAP to identify hypervariable proteins based on all samples generated by the total five MS runs. This analysis could be performed in either an unsupervised or a supervised manner. For the former, the chi-square statistics derived by zMAP along with the associated numbers of degrees of freedom were summed across MS runs, producing p-values that assessed the overall expression variability of each protein (see "Methods"). In total, we identified 2290 significant hypervariable proteins (BH-adjusted p-value < 0.05; Additional file 2: Note S2). A hierarchical clustering of these proteins revealed quite a few meaningful protein expression patterns (Fig. 3a). For example, the largest cluster comprised 591 proteins whose expression was decreased during human erythropoiesis at both fetal and adult stages. GO enrichment analysis showed that these proteins were significantly enriched in several biological processes, including regulation of actin filament length, regulation of cellular component size, and glycolytic process. All of them were associated with stem cell maintenance and self-renewal [36, 37]. Another example cluster consisted of 316 proteins that were specifically upregulated during adult erythropoiesis. These proteins were enriched in biological processes related to ATP and nucleoside synthesis. Consistently, it has been reported that many ATP synthesis genes are subject to post-transcriptional regulation in adult ProEs [30].

Alternatively, the hypervariable analysis could be conducted in a supervised manner in which the samples were grouped by their biological labels. Here, we tried a simple computational pipeline for selecting hypervariable proteins across the four conditions. First, the z-statistics associated with each protein were averaged separately within each group. Then, the standard deviation of the average z-statistics was used to rank all proteins, and we selected top 0.5% as hypervariable ones (Fig. 3b). We further clustered these proteins and found that the vast majority of them had elevated expression specifically in one or two conditions (Fig. 3c). Similar results were observed when different cutoffs were applied to the selection of hypervariable proteins (Additional file 1: Fig. S4). Previously, we had studied the dynamics of protein expression during human erythroid differentiation at adult stage [30]. Here, we focused on fetal erythropoiesis and selected four proteins for further exploration, which were *PNMT*, *CHI3L1*, *S100A9*, and *S100A8*. All of them showed a potential to promote erythroid differentiation at fetal stage: the expression of *PNMT* and *CHI3L1* was concentrated in F5; *S100A9* and *S100A8* were mainly expressed in F0 and F5, with the expression in the latter being even higher than in the

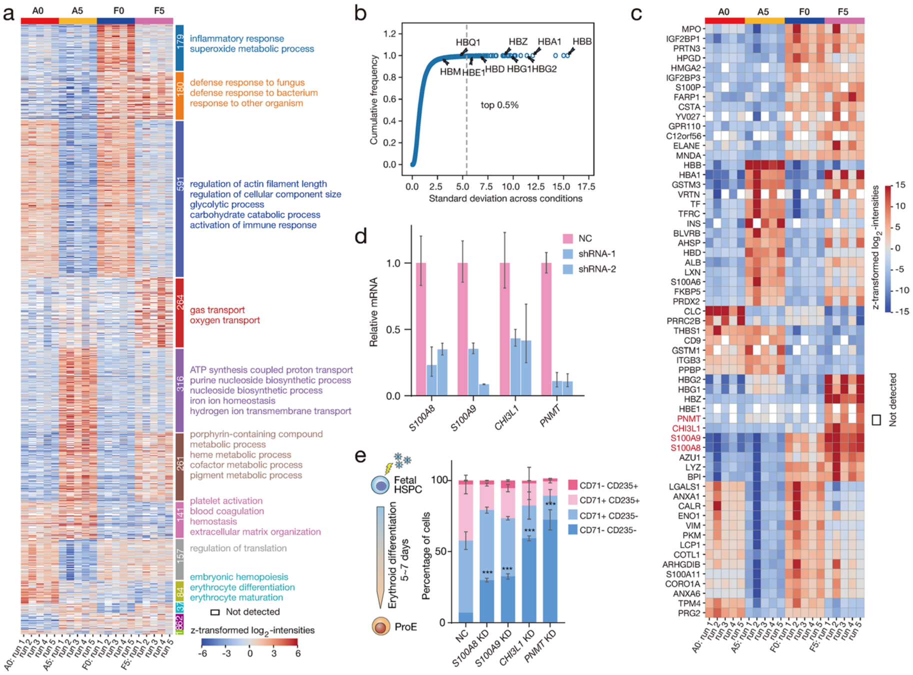


Fig. 3 Simultaneously comparing ILMS samples from all MS runs. **a** Heat map showing the *z*-statistics of significant hypervariable proteins identified from the unsupervised comparison analysis. These proteins were clustered into 12 groups. Representative biological processes for most groups (from GO enrichment analysis) are displayed. **b** In the supervised comparison analysis, ranking all proteins by the standard deviation of average *z*-statistics in the four biological conditions. Top 0.5% were selected as hypervariable proteins. **c** Heat map showing the *z*-statistics of these hypervariable proteins. Proteins that were detected in only one or two MS runs are not displayed. **d** Using qRT-PCR to measure the mRNA expression levels of *S100A8*, *S100A9*, *CHI3L1*, and *PNMT* in human fetal HSPCs under different conditions. NC, negative control. **e** Assessing the progress of erythroid differentiation under different conditions based on the expression of *CD71* and *CD235* (accessed via flow cytometry). Two biological replicates using different shRNAs were incorporated. KD, knock-down

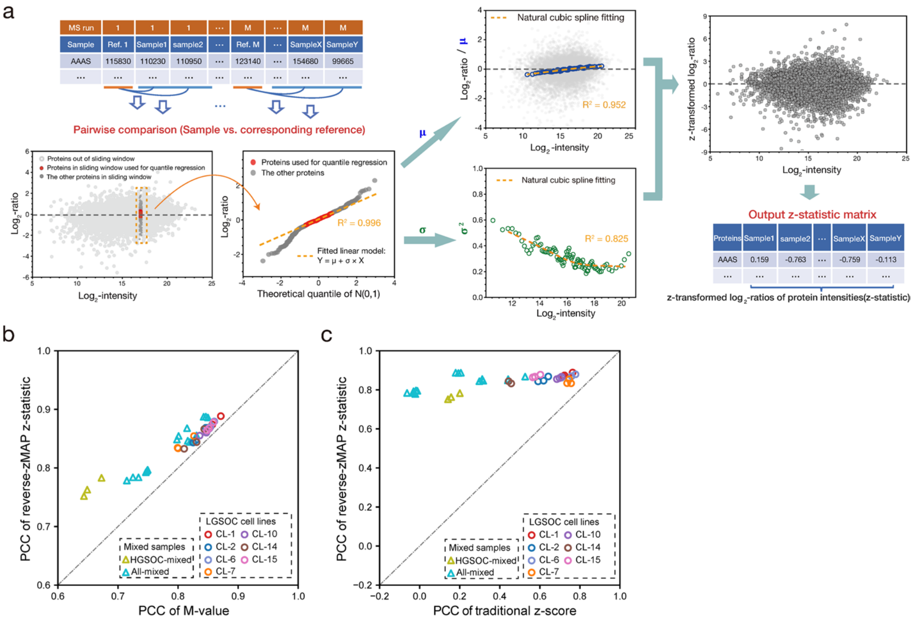
former (Fig. 3c). To assess whether these proteins were required for fetal erythropoiesis, we employed an in vitro differentiation assay of human fetal HSPCs into ProEs, and utilized short hairpin RNA (shRNA) to separately knock down the four proteins in fetal HSPCs on day 0 (Fig. 3d). Upon the depletion of each protein, we observed significantly suppressed erythroid differentiation. Specifically, fetal HSPCs with *PNMT*, *CHI3L1*, *S100A9*, or *S100A8* knock-down generated much fewer CD71$^+$CD235$^+$ ProEs on day 6 (Fig. 3e; Additional file 1: Fig. S5), suggesting these genes play an indispensable role in the maturation of human fetal erythroid cells.

Workflow of reverse-zMAP module

The practical applicability of zMAP module is limited. For example, in cancer studies with a large cohort of patients, different MS runs are typically designed to handle different individuals. The zMAP module is not applicable in such cases because the pseudo reference profiles constructed by it for different MS runs are not biologically identical, owing to the heterogeneity across patients. Another practical concern is that, when the number of conditions involved in each single MS run is large (such as in 8-plex iTRAQ and 11-plex TMT platforms), the differential proteins between samples can be abundant and the proportion of proteins that are suitable to use for the quantile regression in each sliding window can be very small, leading to unreliable variance estimates. Besides, fitting a single MVC for a large number of samples may not be flexible enough to allow for the variation of mean–variance trend across samples.

In the above scenarios, we recommend adopting the experiment design of adding a true reference sample to each MS run, which has been widely adopted by many cancer proteomic studies [23, 38] and is also the only requirement for applying the reverse-zMAP module. The basic principle of reverse-zMAP module is to fit sample-specific MVCs by separately comparing each sample to the corresponding reference sample, for which the M-values of all proteins are calculated and a sliding window is used to group proteins with close intensity levels (Fig. 4a). In each window, the M-values of the enclosed proteins are approximately considered as following the same normal distribution, and the associated parameters are estimated by applying a quantile regression against the standard normal distribution. In detail, this regression is achieved by fitting a straight line, and the intercept and slope of it are taken as the mean and standard deviation estimates, respectively (similar to the zMAP module, the M-values are ordered and only the middle 50%, by default, are used to fit the line). Next, the standard deviation estimates from all windows are gathered to fit an MVC, and the mean estimates are used to model the trend of M-values along the range of intensity levels, producing an M-A curve that essentially serves as a baseline for correcting for normalization biases. Finally, the M-values of all proteins are transformed into *z*-statistics, with the M-A curve and MVC used for centering and scaling them, respectively (see "Methods").

In order to benchmark reverse-zMAP, we collected a TMT data set that comprised three replicate MS runs [39], such that we could compare different methods for



MS run	1	1	1	···	M	···	M	M
Sample	Ref. 1	Sample1	sample2	···	Ref. M	···	SampleX	SampleY
AAAS	115830	110230	110950	···	123140	···	154680	99665
···	···	···	···	···	···	···	···	···

Output z-statistic matrix

Proteins	Sample1	sample2	···	SampleX	SampleY
AAAS	0.159	-0.763	···	-0.759	-0.113
···	···	···	···	···	···

Fig. 4 Benchmarking for the reverse-zMAP module. **a** The workflow of reverse-zMAP module. **b**, **c** For the ovarian carcinoma data set, the PCC between each pair of samples with the same biological context was calculated. The PCCs derived from the *z*-statistics of reverse-zMAP are compared to those from (**b**) M-values and (**c**) traditional *z*-scores

integrating samples across MS runs by evaluating the consistency between replicate samples after applying the corresponding transformations. Each MS run generated 11 samples, profiling the proteomes of 7 low-grade serous ovarian carcinoma (LGSOC) cell lines, a uniform mixture of protein extracts from these LGSOC cell lines, a uniform mixture of 5 high-grade serous ovarian carcinoma (HGSOC) cell lines, and a uniform mixture of all the 12 ovarian carcinoma cell lines (two replicates for this mixture).

We then considered the LGSOC-mixed sample in each MS run as internal reference and accordingly applied the reverse-zMAP module. Similar to the previous benchmarking analysis, we have also used M-values and traditional *z*-scores to integrate samples across MS runs. The former was derived by comparing each sample to the corresponding reference sample, and the latter was derived by applying *z*-score transformation to all the M-values (from all MS runs) associated with each protein. Finally, we calculated the PCC between each pair of samples with the same biological context, based on the different kinds of signal measurements. It was found that the PCC derived based on the *z*-statistics of reverse-zMAP was always higher than those from M-values and traditional *z*-scores (Fig. 4b, c).

Applying reverse-zMAP to a TMT data set about human hepatocellular carcinoma (HCC)

We applied reverse-zMAP to a TMT data set that profiled the proteomes of 159 hepatitis B virus (HBV)-related HCC patients [23]. This data set comprised 33 MS runs, each of which generated 11 samples corresponding to the tumor tissues and NATs of 5 patients plus a reference sample (samples of 6 patients were later excluded in the original study because of low quality; Additional file 1: Fig. S6). A mixture of equal amounts of protein extracts from the tumor tissues and NATs of 50 patients was used to generate the reference sample in each MS run.

When applying reverse-zMAP, we examined the goodness of fit for the associated regressions. For all pairwise comparisons, the quantile regressions performed in the sliding-window process all achieved an R^2 value above 0.99, and the median R^2 values for the associated fitting of M-A curves and MVCs were 0.87 and 0.72, respectively (Additional file 1: Fig. S7). Note that the observed mean–variance trend varied considerably across samples and was typically not as regular as observed on the previous iTRAQ data set (Additional file 1: Fig. S8; Additional file 2: Note S3). We therefore used natural cubic spline interpolation for the fitting of all MVCs (and also the fitting of all M-A curves). After transforming all M-values into *z*-statistics, we examined their distribution separately for each sample. Specifically, the *z*-statistics associated with each sample were ordered and were plotted against the corresponding theoretical quantiles of the standard normal distribution. It was found that the middle 50% of the *z*-statistics matched very well with corresponding theoretical quantiles, while the *z*-statistics at the two ends had even larger absolute values than corresponding theoretical quantiles (Additional file 1: Fig. S9).

We next identified hypervariable proteins for this data set. For each protein, the sum of squares of all the associated *z*-statistics was compared to a chi-square distribution, producing a *p*-value that assessed the overall expression variability of the protein. In total, 3097 significant hypervariable proteins were identified (Bonferroni-adjusted *p*-value < 0.01), and most of them showed consistently increased/decreased expression in tumor tissues compared to NATs (Additional file 1: Fig. S10). We further performed PCA on all the 159 pairs of samples by using *z*-statistics of the hypervariable proteins as features (Fig. 5a). In the two-dimensional PC plot, samples originated from different MS runs were mixed together, indicating the associated batch effects were effectively removed.

It was also observed that the tumor samples and NAT samples were largely separated from each other along the PC1 dimension, suggesting the PC1 score might have the

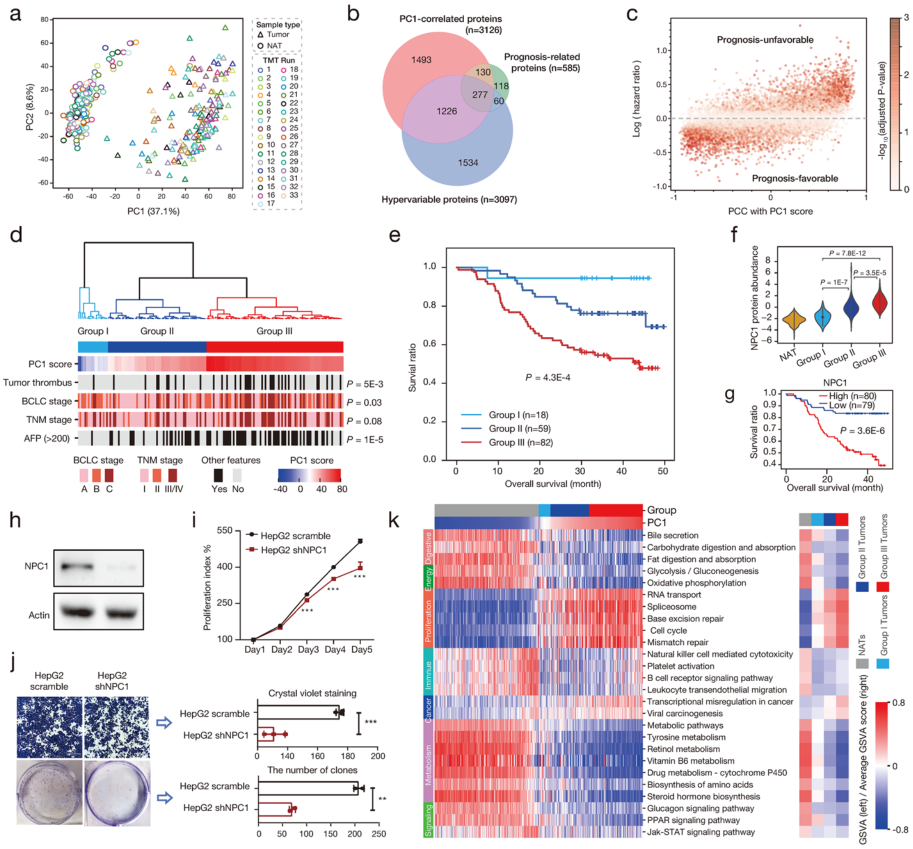


Fig. 5 Applying reverse-zMAP to an HCC TMT data set. **a** Performing PCA on all samples of the HCC data set, with the *z*-statistics (derived by reverse-zMAP) of hypervariable proteins as features. **b** Venn diagram showing the overlap among PC1-correlated proteins, prognosis-related ones, and hypervariable ones. **c** Plotting the log-hazard ratio of each protein against the PCC between its *z*-statistics and the PC1 scores of samples. Colors indicate the BH-adjusted *p*-values for identifying prognosis-related proteins. **d** Hierarchical clustering of all patients based on the PC1 scores of their tumor samples. The patients were accordingly classified into three groups. The *p*-value associated with each clinical feature was derived by applying ANOVA to comparing PC1 scores. **e** Kaplan–Meier curves for the three groups of HCC patients. The *p*-value assessed the survival difference across the groups and was derived by applying log-rank test. **f** Violin plots showing the *z*-statistics of *NPC1* in all NAT samples and different groups of tumor samples. The *p*-values were derived by applying *t*-test. **g** Dividing all patients into two groups based on the median expression of *NPC1* in their tumor tissues and assessing the survival difference between the groups. **h** Using western blot to measure the expression of *NPC1* in HepG2 cell line under different conditions. **i, j** Using MTT assay, transwell migration assay, and plate clone formation experiment to quantitatively assess the influence of *NPC1* knock-down on the cell proliferation, migration, and colony forming ability of HepG2 cell line, respectively. Three biological replicates were generated for each experiment. **k** Heat map showing the GSVA scores of biological pathways associated with differential activity across sample groups. The average GSVA score of each pathway in each sample group is also displayed. The differential pathways were identified by applying limma to comparing GSVA scores (BH-adjusted *p*-value < 0.05)

potential to quantitatively assess tumorigenesis and even the stage of tumor progression, with respect to the proteomic landscape. To further explore the role of PC1, we defined PC1-correlated proteins based on the PCC between *z*-statistics and PC1 scores. In total, we found 1638 positively correlated proteins ($\text{PCC} > 0.5$) and 1488 negatively correlated ones ($\text{PCC} < -0.5$), which corresponded very well to upregulated and downregulated proteins in tumor tissues, respectively (Additional file 1: Fig. S11a). Pathway enrichment analysis showed that the positively correlated proteins were enriched in several pathways related to genetic information processing, such as spliceosome, cell cycle, and mismatch repair, while the negatively correlated proteins were enriched in metabolic pathways associated with normal liver function, such as retinol metabolism [40, 41], drug metabolism [42], and fatty acid degradation [43] (Additional file 1: Fig. S11b). These results suggested excessive cell proliferation and the disorder of liver metabolism in HCC tumor.

We next evaluated the prognostic association of the PC1-correlated proteins. We first defined prognosis-related proteins by performing a regression of the overall survival time of the patients on the expression intensities of each protein. More specifically, a Cox proportional hazards model was separately fitted for each protein, with its *z*-statistics in tumor samples as the only predictor [44]. In total, we defined 585 prognosis-related proteins (BH-adjusted $p\text{-value} < 0.05$), including 335 prognosis-favorable proteins ($\text{hazard ratio} < 1$) and 250 prognosis-unfavorable ones ($\text{hazard ratio} > 1$). A significant overlap was observed between these prognosis-related proteins and the PC1-correlated ones: 69.6% of the prognosis-related proteins were also PC1-correlated (Fig. 5b). As a comparison, only 3.9 and 0.8% of the prognosis-related proteins were identified as PC2 and PC3-correlated ones, respectively (Additional file 1: Fig. S12a). Further examination revealed that the prognosis-favorable proteins remarkably overlapped with the PC1-negatively correlated proteins, while most of the prognosis-unfavorable proteins were PC1-positively correlated ones (Additional file 1: Fig. S12b). We also globally examined this relationship in a more quantitative manner: the log-hazard ratio of each protein was plotted against the PCC between its *z*-statistics and the PC1 scores. A strong positive correlation between these two statistics was observed (Fig. 5c).

Intriguingly, the PC1 scores of the tumor samples were significantly associated with several key clinical features of the HCC patients (Fig. 5d). For example, patients with tumor thrombus got significantly larger PC1 scores than the others ($p\text{-value} = 5\text{e} - 3$). Similar results were also observed on patients with high alpha-fetoprotein (AFP) level. These observations strongly implied the clinical implication of PC1. We next made a classification of all the HCC patients based on the PC1 scores of their tumor samples, producing three subgroups of patients (Fig. 5d). A clear survival difference was observed between these subgroups (Fig. 5e), with the corresponding *p*-value being even more significant than those resulting from the classifications based on TNM or BCLC stage (Additional file 1: Fig. S13). Specifically, the patients in group III showed clearly worse prognosis and also much higher frequency of tumor thrombus than those in groups I and II (Fig. 5d, e), suggesting PC1 could contribute to elucidating the molecular events underlying HCC progression.

Following this speculation, we tried identifying HCC progression-related proteins based on the protein expression profile across the subgroups of patients as well as the prognostic association. We selected four proteins for further exploration, including two potentially oncogenic ones (*NPC1* and *UBE2C*) and two potential tumor suppressors (*PIPOX* and *MMAA*) (Fig. 5f, g; Additional file 1: Fig. S14). As an independent verification, we collected RNA expression data of 363 HCC patients from the TCGA database [45]. For each of the four proteins, a significant survival difference was observed when the patients were divided into two groups based on the median RNA expression level of the corresponding gene (Additional file 1: Fig. S15). We also experimentally explored the roles of *NPC1* and *UBE2C* in two HCC cell lines (HepG2 and Huh7). For *NPC1*, shRNA was used to knock down its expression in both cell lines, which resulted in significant suppression of cell proliferation, migration, and colony forming ability (Fig. 5h–j; Additional file 1: Fig. S16a-c). For *UBE2C*, we established its overexpression in the two cell lines and observed significant increases of all the three indexes (Additional file 1: Fig. S16d-i). These results confirmed that elevated expression of the two genes contributed to the proliferation and invasion of HCC cells.

Applying various downstream analyses on the z-statistics of reverse-zMAP

In practice, a common downstream analysis is to assess the overall enrichments of protein sets based on the abundance of individual proteins. Here, we used biological pathways collected from the KEGG database [46] to define protein sets of interest. The GSVA method [47] was then applied to the *z*-statistic matrix produced by reverse-zMAP for the HCC TMT data set, which quantified the activity of each biological pathway in each sample. Finally, the pathways exhibiting differential activity across sample groups were identified and clustered (Fig. 5k).

It was observed that many of the identified pathways had stepwise increased/decreased activity from NAT samples to tumor samples in groups I–III, showing a good correlation with the PC1 scores. Example pathways with decreased activity across the sample groups included many liver function-related ones, such as bile secretion, retinol metabolism, and vitamin B6 metabolism; examples with increased activity included those related to cell proliferation or cancer, such as cell cycle, mismatch repair, transcriptional misregulation in cancer, and viral carcinogenesis (Fig. 5k). Notably, a few NAT samples were associated with even larger PC1 scores than some tumor samples in group I. Concordantly, compared to the other NAT samples, these samples showed clearly lower activity of many liver function-related pathways (e.g., bile secretion and retinol metabolism) and much higher activity of several cancer-related pathways (e.g., viral carcinogenesis) (Additional file 1: Fig. S17).

We further dug into the NAT samples with excessively large PC1 scores. It was noted that these samples, compared to the other NAT samples, displayed clearly higher activity of Jak-STAT signaling pathway and leukocyte transendothelial migration (Additional file 1: Fig. S18a). The former has been implicated in the pathogenesis of inflammation [48], and the latter is vital for innate immunity and inflammation response [49]. Inspired by these observations, we examined the serum gamma-glutamyl transferase (GGT) concentrations, a commonly used biomarker for hepatitis, of the HCC patients. Intriguingly, all the patients whose NAT samples had excessively large PC1 scores were associated with abnormally high levels of serum GGT (Additional file 1: Fig. S18b). Together, these findings suggested the high-PC1 NATs could be in a state of severe inflammation, providing proteome-level insights into the progression from HBV-infected liver tissue to HCC.

We concluded here that the variation of PC1 score across NAT samples made biological sense as well as that across the tumor samples. Since the PC1 scores were derived from the *z*-statistics of the hypervariable proteins, this conclusion implied a potential advantage of unsupervised hypervariable analysis over traditional differential analysis: in addition to the differences between sample groups, hypervariable analysis may also capture the heterogeneity within each group. In the original study of the HCC patients, 1274 differentially expressed proteins between the tumor and NAT samples were identified, 1113 (87.4%) of which were also identified as hypervariable proteins in this study (Additional file 1: Fig. S19a). We then clustered the proteins uniquely identified as hypervariable ones and found that two of the resulting clusters were associated with clear expression heterogeneity across the NAT samples (these two clusters were referred to as C4 and C5, which comprised 176 and 241 proteins, respectively). Consistent with the above speculation of hepatitis, both C4 and C5 showed elevated expression levels specifically in the NAT samples with excessively large PC1 scores and were also significantly enriched in biological pathways that suggest activated immune and inflammatory responses (Additional file 1: Fig. S19b). For example, the C4 proteins were enriched in the pathway of ECM-receptor interaction, and it has been suggested that the deposition and remodeling of ECM can enhance local immune response to chronic hepatitis tissue [50]; the C5 proteins were enriched in neutrophil extracellular trap formation and leukocyte transendothelial migration, which are crucial for innate immunity and inflammation response [51, 52]. In order to more quantitatively dissect the expression heterogeneity of the two protein clusters, we hierarchically clustered all NAT samples into two subgroups based on their PC1 scores, which successfully isolated the NAT samples with excessively large PC1 scores (NAT II group, 12 samples) from the others (NAT I group, 147 samples). We then calculated for each sample the average *z*-statistic across C4/C5 proteins, as an overall expression evaluation for the protein cluster. It was found that the expression of both C4 and C5 was considerably higher in NAT II than in NAT I and the three subgroups of tumor samples, while the differences among NAT I and the tumor subgroups were not as distinct (Additional file 1: Fig. S19c, d). Since NAT II only accounted for a small proportion (7.5%) of all NAT samples, the overall expression difference between all NAT and tumor samples was also not distinct (especially for the C4 cluster, to which the corresponding *t*-test *p*-value was only 0.31 even with such a large sample size (159 vs. 159); Additional file 1: Fig. S19c, d), which explained why the associated proteins had not been identified as differential ones in the original study.

As another demonstration of the utility of the *z*-statistics derived by reverse-zMAP, we constructed a protein co-expression network for the HCC patients based on the *z*-statistics of their tumor samples. In this network, each protein pair associated with a positive and significant partial correlation coefficient (PTCC) was considered to be co-expressed (see “Methods”). We then specifically took out the sub-network consisting of the hypervariable proteins and identified co-expression modules from it. Finally, we performed functional annotation for each module by identifying enriched biological pathways.

It was found that most of the modules were significantly enriched within one or more pathways, suggesting the proteins belonging to the same module had coordinated functions (Additional file 1: Fig. S20a). We also noticed that the expression of the proteins in the same module tended to have consistent correlations with the PC1 scores (Additional file 1: Fig. S20b). For example, there were two modules enriched within the cell cycle pathway, and all their members had previously been identified as PC1-positively correlated proteins.

Associating z-statistics with the mutation landscapes of the HCC patients

To search for potential driver mutations of proteomic variation across the HCC patients, we performed a quantitative trait locus (QTL) analysis that was aimed at identifying significant gene-protein associations, in the sense that the abundance of the protein was significantly associated with the mutation status of the gene. For each candidate gene-protein pair, the *z*-statistics of the protein in tumor samples were linearly regressed against the (non-silent somatic) mutation indicators of the gene, with the age and gender variables properly accounted for by treating them as covariates. One hundred twelve genes with non-silent somatic mutations in at least 10 HCC patients were used for this analysis, resulting in 2031 significant gene-protein associations in total (BH-adjusted p-value < 0.05). Notably, the vast majority (98.3%) of these associations were attributed to 5 genes, which were *CTNNB1*, *TP53*, *AXIN1*, *TSC2*, and *RB1* (Fig. 6a). For these 5 hotspot genes, we evaluated their associations with the PC1 scores. Both *AXIN1* and *TSC2* showed a tendency to mutate at tumor samples with relatively large PC1 scores, as suggested by two analysis results: (i) the tumor samples harboring *AXIN1/TSC2* mutations had significantly larger PC1 scores than the other tumor samples (Fig. 6b); (ii) the mutation rate of *AXIN1/TSC2* among the tumor samples in group III was significantly higher than in the other two groups (Additional file 1: Fig. S21). No significant PC1 association was observed on the other 3 genes. We also examined the prognostic associations of the 5 genes. It was found that only *TP53* and *TSC2* were linked with a significant survival difference between mutated and non-mutated patients (Fig. 6c; Additional file 1: Fig. S22). Together, these observations implied an important role of *TSC2* mutation in HCC.

The HCC patients harboring *TSC2* mutations had worse survival than the others. Moreover, the protein expression of *TSC2* itself significantly decreased in those *TSC2*-mutated tumor samples (Fig. 6d), suggesting *TSC2* could be a tumor suppressor for HCC. To further explore this speculation, we assessed the prognostic associations of the proteins whose expression was significantly associated with *TSC2* mutation as identified in the previous QTL analysis. By examining the fitted Cox models (Fig. 5c), we found that almost all the proteins with increased expression from the mutation of *TSC2* (referred to as class I proteins) were prognosis-unfavorable, and almost all the proteins with decreased expression (referred to as class II proteins) were prognosis-favorable (Fig. 6e). GO enrichment analysis indicated the class I proteins were enriched within the biological process of response to amino acid starvation. Related proteins included quite a few belonging to the family of V-ATPases (Fig. 6f), which plays a vital role in modulating autophagy, cell invasion, and cell death [53]. Previous studies have also shown that the inhibition of V-ATPases considerably restrained the growth, migration, and invasion of cancer cells [54]. The class II proteins were enriched within the process of cell–cell junction organization. The dysregulation of this process can lead to the loss of cell–cell

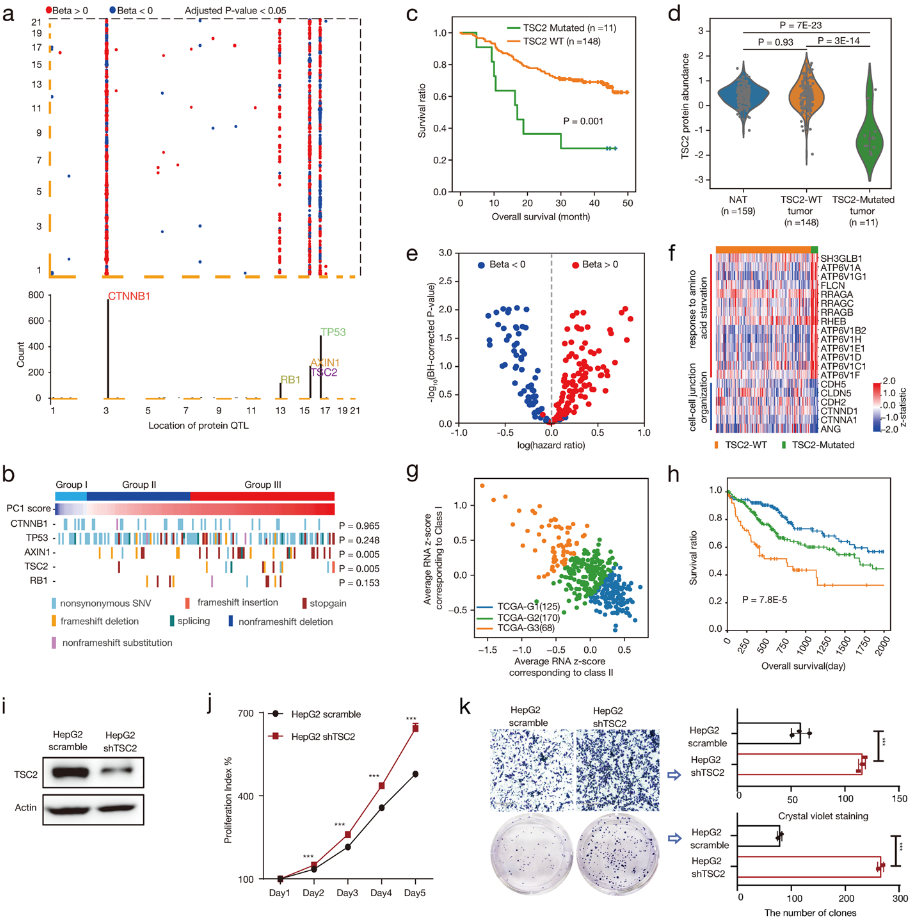


Fig. 6 Associating *z*-statistics of reverse-zMAP with mutation landscapes of the HCC patients. **a** Two-dimensional plot displaying the significant gene-protein associations, with *Y* and *X* axes representing the locations of proteins and mutated genes in the genome, respectively. The total number of proteins associated with each mutated gene is also displayed. For a specific gene-protein association, Beta > 0 suggests the protein has increased expression with the mutation of the gene, and vice versa. **b** The (non-silent somatic) mutation states of each hotspot gene in all the HCC patients. The *p*-values were derived by applying Wilcoxon rank sum test to comparing PC1 scores. **c** Assessing the survival difference between the patients with and without *TSC2* mutation. The *p*-value was derived by applying log-rank test. **d** Violin plots showing the *z*-statistics of *TSC2* in different groups of samples. The *p*-values were derived by applying *t*-test. **e** Scatter plot of BH-adjusted *p*-values against hazard ratios for the proteins whose expression is significantly associated with *TSC2* mutation (both the *p*-values and the hazard ratios are from the previous fitting of Cox models). These proteins were divided into class I (Beta > 0) and class II (Beta < 0) proteins. **f** Heat map showing the *z*-statistics of selected class I/II proteins in tumor samples. **g** Classifying HCC patients of the TCGA cohort into three subgroups based on the average RNA expression levels corresponding to the two classes of proteins. *K*-means method was used for this classification. The *z*-scores associated with each RNA were calculated based on $\log_2$-FPKM values derived from RNA-seq data. **h** Assessing the survival difference among the three subgroups of TCGA HCC patients by applying log-rank test. **i** Using western blot to measure the expression of *TSC2* in HepG2 cell line under different conditions. **j**, **k** Using MTT assay, transwell migration assay, and plate clone formation experiment to quantitatively assess the influence of *TSC2* knock-down on the cell proliferation, migration, and colony-forming ability of HepG2 cell line, respectively. Three (MTT assay, transwell migration assay) or two (plate clone formation experiment) biological replicates were generated

adhesion and the onset of epithelial-mesenchymal transition, a crucial event promoting cancer cell invasion and metastasis in various types of solid tumors [55–59].

For an independent verification, we again turned to the RNA expression data of the TCGA HCC cohort. For each of the two classes of proteins, the corresponding RNA expression levels in the tumor tissue of each patient were averaged. The resulting two average values were then used as features to classify the patients, producing three subgroups of patients (referred to as G1, G2, and G3; Fig. 6g). For these subgroups, G1 was associated with the lowest and highest RNA expression for the proteins of class I and II respectively, G3 was associated with the highest and lowest RNA expression for the proteins of class I and II respectively, and G2 was associated with intermediate RNA expression for both classes of proteins. Consistently, we found a significant survival difference between these subgroups, with G1 and G3 having the best and worst survival curves, respectively (Fig. 6h).

We have also used shRNA to knock down *TSC2* in two HCC cell lines (HepG2 and Huh7) that expressed wild-type *TSC2* (Fig. 6i; Additional file 1: Fig. S23a). For both cell lines, the knock-down of *TSC2* has led to significant improvements of cell proliferation, migration, and colony-forming ability (Fig. 6j, k; Additional file 1: Fig. S23b, c), confirming the tumor suppressor role of *TSC2* in HCC cells.

Identifying hypervariable proteins across tumor samples only

In large-scale cancer studies, researchers may be interested specifically in the proteomic heterogeneity across tumor tissues from different patients. In this case, reverse-zMAP is applicable as long as the reference samples are designed to be a mixture of protein extracts from tumor tissues only. As an illustration, we have applied reverse-zMAP to a TMT data set about pediatric brain cancer (PBC) [38]. This data set profiled the proteomes of 218 tumor tissue samples from 199 patients representing 7 different histological diagnoses of PBC (Fig. 7a). A reference TMT sample was generated in each MS run by using a uniform mixture of representative tumor tissue samples.

After identifying significant hypervariable proteins (3130 proteins in total, with BH-adjusted p-value < 0.05), we picked out the ones with detected expression in at least half of the 218 samples and used the z-statistics of these proteins to classify the samples. A consensus clustering method [60] was applied, which resulted in 7 subgroups of PBC samples associated with distinct molecular characteristics and significantly different survival outcomes (Fig. 7a, b). Notably, this proteome-based classification showed a statistically significant consistency with the histological diagnoses (chi-square test p-value $= 8.4e-45$), which was also implied by the fact that 72.8% of the hypervariable proteins were covered by differentially expressed proteins (identified by msTrawler) between any pair of histological types (Additional file 1: Fig. S24). For instance, subgroup 7 had the smallest group size and was associated with the worst survival curve among all the subgroups. Consistently, all the 5 PBC samples in this subgroup were histologically diagnosed as high-grade glioma (HGG), which is generally associated with rather bad prognosis [61]. On the other hand, PBC samples of the same histological type could still have distinct protein expression characteristics and be classified into different subgroups. For example, the samples diagnosed as ependymoma were primarily distributed in subgroups 3 and 4. By performing a hierarchical clustering for the hypervariable

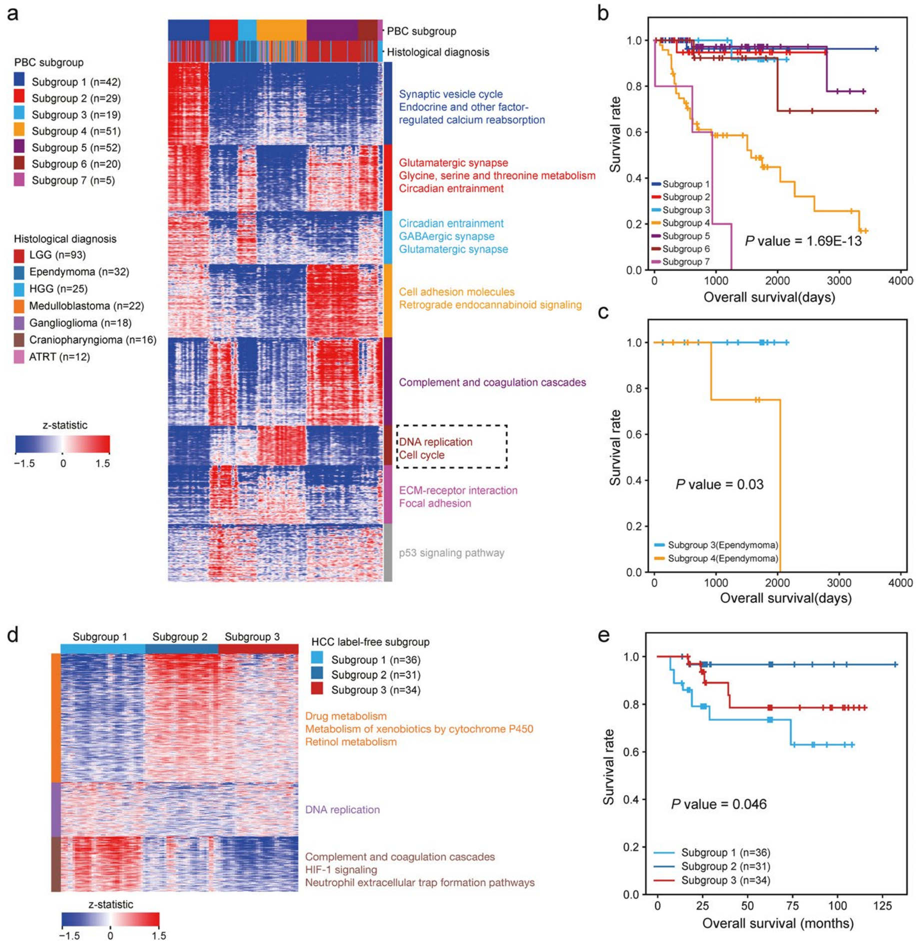


Fig. 7 Identifying hypervariable proteins across tumor samples only. **a** Heat map showing the *z*-statistics of significant hypervariable proteins across the PBC tumor samples. Only the proteins with detected expression in at least half the samples are displayed. Representative biological pathways for the protein clusters (from KEGG enrichment analysis) are also shown. **b** Kaplan–Meier curves showing overall survival (OS) for patients associated with the 7 subgroups of the PBC tumor samples. The *p*-value was derived by applying log-rank test. **c** Kaplan–Meier curves for the PBC tumors diagnosed as ependymoma in subgroups 3 and 4. **d** Heat map showing the *z*-statistics of significant hypervariable proteins identified across the tumor samples of the HCC label-free data set. **e** Kaplan–Meier OS curves for the 3 subgroups of the HCC tumors

proteins, we observed several protein clusters with distinct expression levels between the two subgroups of samples, including a protein cluster that was significantly enriched in the biological pathways of DNA replication and cell cycle (Fig. 7a). The overall expression of this protein cluster was clearly higher in subgroup 4 than that in subgroup 3, which was consistent with the survival difference between the two subgroups (Fig. 7b). Intriguingly, this survival difference was still significant when we specifically picked out the ependymoma samples (Fig. 7c), suggesting different molecular characteristics of PBC samples with the same histological diagnosis could be biologically meaningful.

In practice, label-free proteomic quantification methods are also frequently used in large-scale cancer studies [13, 62]. This technique, unlike ILMS, enables the generation of unlimited samples without introducing any labels [63]. In principle, reverse-zMAP can be applied to label-free proteomic data sets by treating all the included samples as from a single MS run (the zMAP module is not applicable because the distribution of missing values is not consistent across label-free samples, resulting in different numbers of degrees of freedom associated with the observed variances of different proteins). Here, we collected a label-free data set that profiled the proteomes of 101 early-stage HCC tumor samples [13]. For identifying hypervariable proteins across these samples, reverse-zMAP required a reference profile representing the "average" proteomic landscape of them. Such a proteomic sample, however, had not been designed for the data set. We therefore created a pseudo reference profile by averaging the 101 samples, in the same manner as achieved by the zMAP module (see also "Discussion"). Next, reverse-zMAP was applied to transforming the M-values of each sample against the pseudo reference profile into z-statistics. It was found that the signal measurements of this label-free data set were associated with clear mean–variance dependence, as suggested by large variability of M-values derived from low intensity levels (Additional file 1: Fig. S25a). This dependence was effectively diminished by the z-transformation of reverse-zMAP (Additional file 1: Fig. S25b).

In total, 2567 significant hypervariable proteins were identified by reverse-zMAP (BH-adjusted p-value < 0.05). We next picked out the ones with detected expression in all the 101 samples and used the z-statistics of these proteins to classify the samples, which produced 3 subgroups of samples with distinct protein expression signatures and significantly different survival outcomes (Fig. 7d, e). We also performed a hierarchical clustering for the significant hypervariable proteins, resulting in 3 protein clusters enriched in different biological pathways (Fig. 7d). One of the clusters was enriched in metabolic pathways associated with normal liver function, including drug metabolism [42], metabolism of xenobiotics by cytochrome P450 [64], and retinol metabolism [40]. Consistently, this protein cluster exhibited the highest and lowest overall expression in the sample subgroups with the best (subgroup 2) and worst (subgroup 1) survival curves, respectively. Another protein cluster was enriched in the pathway of DNA replication and showed, consistently, the lowest overall expression in subgroup 2.

Discussion

In the study, we developed zMAP toolset to facilitate the analysis of large-scale ILMS data sets. It calculates the M-values of each sample against the corresponding (pseudo) reference profile and accounts for the different uncertainty of these M-values by modeling the associated mean–variance dependence, from which a variance stabilizing z-transformation is devised to improve the comparability of ILMS intensities across MS runs. The transformed z-statistics, as a new kind of measurements of protein abundance, can be directly used to integrate samples from multiple MS runs and perform a variety of downstream analyses on them. Besides the identification of hypervariable proteins, we have shown in the study that the z-statistics can be effectively used for PCA, clustering of proteins/samples, association analysis with survival/mutation data of cancer patients, GSVA, and construction of co-expression network.

A Web-based application of zMAP toolset is provided at http://bioinfo.cemps.ac.cn/shaolab/zMAP, with many of the downstream analyses available for users to choose from (Fig. 8). Both the zMAP and reverse-zMAP modules have been implemented in

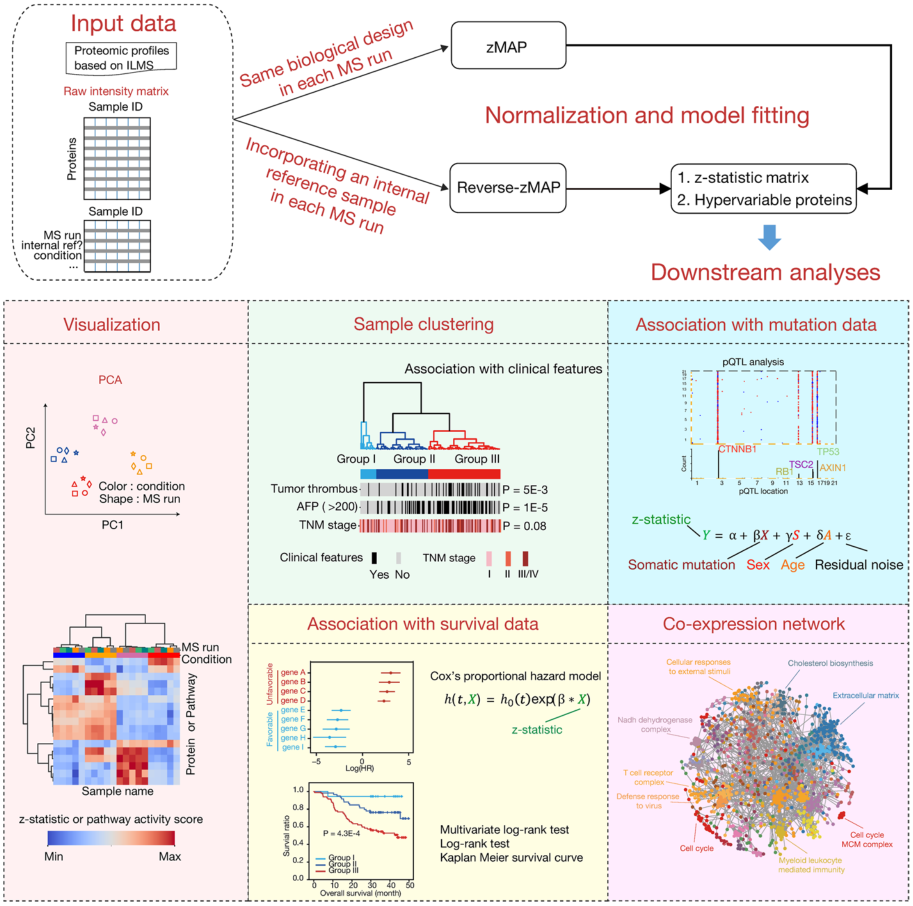


Fig. 8 The overall structure of a Web server based on the zMAP toolset

this application. In practice, these two modules are expected to be sufficient for handling most cases, yet there are indeed scenarios where neither of them is applicable.

Firstly, the reverse-zMAP module can only be used to identify hypervariable proteins around the biological state represented by the internal reference samples (since the reference profiles employed by zMAP module are created by averaging real samples, users can select only the samples across which the hypervariable proteins are of interest as the input of zMAP module, provided the requirement for applying this module is satisfied). For example, if the internal reference samples of a cancer data set are generated from a mixture of both NATs and tumor tissues of different patients (just as those of the HCC TMT data set), then reverse-zMAP is not able to directly identify the hypervariable proteins specifically across the tumor samples. The second point regards the applicability to label-free proteomic data sets. As previously mentioned, our toolset can be applied to such a data set by treating all the included samples as from a single MS run. The application of zMAP module, however, is hindered by the distribution of missing values, which is not consistent between different samples. Applying reverse-zMAP, on the other hand, requires a real reference sample (better represent the average of the other samples) for its statistical model to rigorously hold. Such a reference sample is not always available for a practical label-free data set. If that is the case, reverse-zMAP is still applicable by averaging the included samples to create a pseudo reference profile (just as for the HCC label-free data set), only that it will cause a flaw in the theoretical derivation of the exact null distribution of the final chi-square statistic associated with each protein. This flaw might bias the resulting p-values, though the effect should be very weak when the sample size is large. For future studies, we shall put effort into addressing the above concerns and widening the application scope of zMAP toolset.

Conclusions

zMAP has been presented as a computational toolset for analyzing large-scale ILMS data sets that involve multiple MS runs. The z-statistics derived by zMAP greatly improve the comparability between different MS runs and can be effectively used for a variety of downstream analyses. The wide applicability of zMAP as well as its advantages over existing methods has been demonstrated on several real proteomic data sets.

Methods

The zMAP module

We first detail how the zMAP module handles ILMS data generated by a single MS run. Given protein-level intensities of a set of n samples generated from the same MS run, zMAP first normalizes them based on trimmed total intensities [29]. Let S_i^j denote $\log_2$-normalized intensity of protein i in sample j. zMAP assumes all S_i^j independently follow normal distributions, with the mean and variance parameters linked by an unknown MVC (denoted by $f(\cdot)$). Formally, for each protein i that is not associated with hypervariable expression across the n samples, we have

$$S_i^j \sim N\left(\mu_i, \sigma_i^2\right), j = 1,2,\cdots,n, \tag{1}$$

$$\sigma_i^2 = f(\mu_i), \tag{2}$$

where μ_i and σ_i^2 are unknown parameters representing the mean expression of protein i and the associated intensity variability, respectively.

To fit the MVC, zMAP calculates the sample mean intensity and sample variance for each protein i by

$$\overline{S}_i = \frac{\sum_{j=1}^{n} S_i^j}{n}, \tag{3}$$

$$V_i = \frac{\sum_{j=1}^{n}\left(S_i^j - \overline{S}_i\right)^2}{n}, \tag{4}$$

Then, the V_i of all proteins are plotted against the corresponding $\overline{S}_i$, and zMAP uses a sliding window to scan this plot from left to right. By default, the number of proteins covered by each window is fixed to 400, and the step size for moving the window is 100 proteins (the default settings were adopted throughout the whole study). Since the proteins grouped in the same window are expected to have similar mean expression, zMAP makes an approximation by deriving a single variance estimate for each window. For each non-hypervariable protein i,

$$V_i \sim \sigma_i^2 \cdot \frac{\chi_{n-1}^2}{n-1}, \tag{5}$$

we havewhere χ_{n-1}^2 refers to the chi-square distribution with $n-1$ degrees of freedom. Accordingly, zMAP derives the variance estimate for a specific window by performing a quantile regression, in which the enclosed V_i are ordered and are linearly regressed against the corresponding theoretical quantiles of $\chi_{n-1}^2/(n-1)$. To avoid the influence of underlying hypervariable proteins, only a certain proportion (30% by default) of the smallest V_i are used for this regression. The method of least squares is used to fit a straight line through the origin, and the fitted slope is taken as the variance estimate.

zMAP next fits the global MVC by regressing the variance estimates from all windows against the corresponding average $\overline{S}_i$ (the average $\overline{S}_i$ associated with each window is calculated over the proteins used for the quantile regression). Two optional methods have been designed for the MVC fitting, which are suited to iTRAQ and TMT data respectively. The first method fits an exponential decay function in the form of $\sigma^2 = \exp(\theta_1 + \theta_2 \cdot \mu) + \theta_3$, based on the least squares criterion. If the fitted θ_3 is negative, it is set to 0 and the other two parameters are re-estimated. The second method employs natural cubic spline interpolation to fit the MVC. The number of degrees of freedom is set to 3, and the number of knots is set to 3 or 4, depending on which setting leads to the better R^2.

Finally, zMAP performs a z-transformation of all S_i^j based on the fitted MVC:

$$Z_i^j = \frac{S_i^j - \overline{S}_i}{\sqrt{f\left(\overline{S}_i\right)}}. \tag{6}$$

It also identifies potential hypervariable proteins by comparing the sample variance of each protein with the corresponding variance implied by the MVC. Formally, the key statistic for protein i, named χ^2-statistic, is calculated as $Q_i = (n-1)V_i/f\left(\overline{S}_i\right)$, which is subsequently compared to the χ_{n-1}^2 distribution for deriving a p-value (i.e., the upper-tailed probability).

For integrating zMAP results across MS runs, the z-statistic matrices are simply concatenated. Besides, the key statistics for each protein, along with the associated numbers of degrees of freedom, are summed, producing a p-value that assesses the overall expression variability of the protein. The results of applying zMAP to the iTRAQ data set regarding human erythropoiesis are given in Table S1 in Additional file 3.

In vitro erythroid differentiation of human fetal HSPCs

Primary human fetal CD34$^+$ HSPCs were purchased from iXCells Biotechnologies. Fetal erythroid lineages were generated in vitro using a two-phase suspension culture system as described previously [65]. Briefly, human fetal CD34$^+$ HSPCs were firstly expanded in StemSpan SFEM medium (StemCell Technologies Inc.) with 1 × CC100 cytokine mix (StemCell Technologies Inc.) and 2% penicillin/streptomycin for 6 days. Cells were maintained at a density of ~ 5×10^5 cells/ml with media changes every 2 days during the initial expansion stage. Expanded HSPCs were collected on day 6 and differentiated in StemSpan SFEM medium supplied with 1 U/ml erythropoietin, 5 ng/ml IL-3, 20 ng/ml SCF, 2 μM dexamethasone, and 1 Mμ estradiol and 2% penicillin/streptomycin for additional 6 days. Cells were maintained at a density of less than 1×10^6 cells/ml by supplementing cultures with fresh media every 2 days.

shRNA targeting human *S100A8*, *S100A9*, *CHI3L1*, and *PNMT* were subcloned into the lentiviral pLKO.1-puro vector by EcoRI and AgeI sites. Insertion of shRNA was validated by Sanger sequencing. Cells transduced with empty vector were used as control.

Erythroid differentiation of fetal HSPCs on day 6 was analyzed on FACSfortessa (BD Biosciences) for *CD71* and *CD235a* expression using the antibodies conjugated to phycoerythrin (PE) and fluorescein isothiocyanate (FITC), respectively. Dead cells were excluded by DAPI staining. Erythroid lineages were defined by cell surface markers for CFU-E ($CD71^{+}CD235a^{-}$), Erythroblast ($CD71^{+}CD235a^{+}$), and mature erythroid cells ($CD71^{-}CD235a^{+}$). Data were analyzed by using FlowJo 7.6.1.

The reverse-zMAP module

We detail here how reverse-zMAP integrates ILMS samples across MS runs. Suppose there is a biologically identical reference sample in each MS run. Reverse-zMAP achieves the integration by separately comparing each sample to the corresponding reference sample and transforming the resulting M-values into *z*-statistics.

Given protein-level intensities of an ILMS sample generated from some MS run and the corresponding reference sample, reverse-zMAP first normalizes these two samples based on trimmed total intensities[33]. Let r_i and S_i denote $\log_2$-normalized intensities of protein i in the reference sample and the other one, respectively. Let $M_i = S_i - r_i$ be the M-value of protein i. Reverse-zMAP assumes the M_i of all non-differential proteins independently follow normal distributions, with the mean and variance parameters modeled as functions of r_i (all r_i are treated as non-stochastic in the statistical framework of reverse-zMAP). Formally, for each non-differential protein i, we have

$$M_i \sim N\big(g(r_i), f(r_i)\big) \tag{7}$$

Here, $f(\cdot)$ is an unknown MVC; $g(\cdot)$ is referred to as an M-A curve and is essentially used to account for normalization biases.

To fit the two curves, the M_i of all proteins are plotted against the corresponding r_i, and reverse-zMAP applies the same sliding-window procedure as used by the zMAP module to scanning this plot. For each specific window, the enclosed proteins are associated with similar r_i, and reverse-zMAP makes an approximation by deriving a single estimate of the corresponding $g(r_i)$ and $f(r_i)$. For this estimation, the M_i of the enclosed proteins are ordered and are linearly regressed against the corresponding theoretical quantiles of the standard normal distribution. To avoid the influence of differential proteins, only the middle 50% (by default) of the M_i are used for this regression. The least squares method is adopted to fit a straight line. The fitted intercept is taken as the estimate of $g(r_i)$, and the square of the slope is the estimate of $f(r_i)$.

Next, reverse-zMAP pools the $f(r_i)$ estimates from all windows and regresses them against the corresponding average r_i. The same MVC fitting procedure as used by the zMAP module is adopted for this regression. For fitting $g(\cdot)$, the $g(r_i)$ estimates are pooled and the natural cubic spline interpolation method is applied. Finally, reverse-zMAP applies a z-transformation to each M_i by

$$Z_i = \frac{M_i - g(r_i)}{\sqrt{f(r_i)}}. \tag{8}$$

To integrate ILMS samples across MS runs, reverse-zMAP repeatedly applies the above procedure to each (non-reference) sample. The resulting z-statistic vectors are then combined to get the final z-statistic matrix. After that, the z-statistics associated with each protein are squared and summed, and the result is compared to the corresponding chi-square distribution to get a p-value. The results of applying reverse-zMAP to the TMT data sets about serous ovarian carcinoma cell lines, HCC patients, PBC patients, and the HCC label-free data set are given in Tables S2-S5 in Additional file 3.

Pathway enrichment analysis for the PC1-correlated proteins

Gene sets of biological pathways were downloaded from the Molecular Signatures Database (https://www.gsea-msigdb.org/gsea/msigdb) [66]. Fisher's exact test was used to evaluate the enrichment of PC1-positively/negatively correlated proteins within each biological pathway, with the background protein set limited to those proteins that were detected in at least half of the TMT samples.

Classification for the HCC TMT data set

The HCC patients were hierarchically clustered based on the PC1 scores of their tumor samples. We first invoked the `cluster.hierarchy.linkage` function of the scipy package to generate a linkage matrix, with the parameter setting `method="complete"`. Then, the linkage matrix was passed to the `cluster.hierarchy.fcluster` function, with `criterion="maxclust"` to form flat clusters. The number of clusters was empirically set to 3.

Functional assays of NPC1, UBE2C, and TSC2 in HepG2 and Huh7 cell lines

The human liver cancer cell lines HepG2 and Huh7 were obtained from American Type Culture Collection (ATCC). Cells were grown in an incubator at 37 °C under 5% CO_2, and supplemented in DMEM medium (Gibco) contained with 10% fetal bovine serum (Biological Industries) and 100 units/ml penicillin and streptomycin (Gibco).

For the overexpression of *UBE2C*, the template for PCR of *UBE2C* was purchased from Bio-Research Innovation Center Suzhou (plasmid number, SP-100953). It was then inserted into pCDH-CMV vector to construct pCDH-CMV-UBE2C plasmid, which was used to overexpress *UBE2C* in the cells. For the knock-down of *NPC1* or *TSC2*, the corresponding shRNA was constructed into PLKO.1 vector. All the plasmids including packaging ones were transfected into human embryonic kidney (HEK) 293 T cells using lipofectamine 3000 regent (L3000008, Thermo Fisher Scientific) to package and release the virus according to the instructions. Then, HepG2 and Huh7 cells were cultured and infected with virus to achieve stable overexpression or knock-down.

For the MTT assays, HepG2 and Huh7 cells were seeded in 96-well microplate at a density of 2×10^3 cells per well in 100 μL culture medium and were detected at 12, 36, 60, 84, and 108 h respectively. Next, 10 μL of MTT solution (5 mg/ml, Sigma) was added to each well, which was then incubated for 4 h. After that, the supernatants containing MTT were discarded, and we added 100 μL/well DMSO to dissolve formazan crystals. Finally, the plate was recorded at the absorbance of 490 nm. Data were repeated three times.

For the transwell assays, cells were seeded in the upper layer of 12-well Transwell plates (Corning) by 10,000 cells per well, which diluted in serum-free DMEM medium contained with 2% BSA. DMEM medium contained with FBS was added to the lower layer. After culturing for 20 or 24 h, the cells were fixed with 4% PFA and were stained by crystal violet.

For protein extraction, cells were scraped and lysed in 200 μL RIPA buffer (20 mM Tris–HCl pH 8.0, 150 mM NaCl, 1% NP40, 1% SDS, 10 mM NaF) containing protease inhibitors (aprotinin, pepstatin, leupeptin, vanadate, PMSF). Cell lysates were incubated on ice for 30 min and were centrifuged at 12,000 *g* for 10 min at 4 °C, and the supernatant was then harvested. Proteins were quantified by using the Pierce™ BCA (Thermo Scientific™), and were boiling with SDS loading.

For western blot, samples were separated by 10% SDS-PAGE and were transferred to PVDF membrane (Millipore) for 200 mA, 120 min. Then, the membranes were sealed in 5% defatted milk and were blotted with antibody for 4 °C overnight. On the second day, the membranes were rinsed with PBST for 3 × 10 min and were then incubated with secondary antibody for 1 h at RT. Finally, the membranes were rinsed with PBST for 3 × 10 min to develop by ECL chemiluminescence (Millipore). The first antibodies were UBE2C (SantaCruz), NPC1 (SantaCruz), TSC2 (Abclonal), HSP90 (CST), Actin (Proteintech). The secondary antibodies were purchased from Jackson immune research (115–035-003 and 111–035-003).

All the data were repeated two or three times, and all the related comparison analyses were performed by applying *t*-test. The marks *, **, and *** suggested a *p*-value less than 0.05, 0.01, and 0.001, respectively.

Applying the GSVA method to the z-statistic matrix of the HCC TMT data set

Gene sets of KEGG pathways were downloaded from the MSigDB. Then, the gsva function of the `gsva` package (v1.44.3) was invoked with the parameter setting `kcdf = "none"`, since the *z*-transformation performed by reverse-zMAP had made the expression measurements of different proteins comparable with each other.

QTL analysis for the HCC TMT data set

We first picked out the genes that were associated with non-silent somatic mutations in at least 10 of the HCC patients, resulting in 112 genes in total. Then, we picked out the proteins whose expression was detected in at least half of the tumor samples, resulting in 8935 proteins in total. Finally, for each candidate gene-protein pair, we performed a regression of the protein expression against the mutation status of the gene using MatrixEQTL [67] (v2.3) package in R (v4.2.1), by fitting the following linear model:

$$Y = \alpha + \beta X + \gamma S + \delta A + \varepsilon,$$

$$\varepsilon \sim N\left(0, \sigma^2 I\right).$$

Here, Y was a vector of the z-statistics (derived by reverse-zMAP) of the protein in tumor samples; X was the non-silent somatic mutation indicators of the gene; S and A referred to the sex and age variables, respectively; ε was a vector of independent and identically distributed noise variables; α, β, γ, and δ were unknown parameters. This model was fitted by applying the least squares method, and the null hypothesis $\beta = 0$ was then tested by applying the t-test.

Co-expression network construction and module detection

A protein co-expression network was constructed for the HCC TMT data set, based on the z-statistics (derived by reverse-zMAP) associated with the 159 tumor samples. Only the 8935 proteins that were detected in at least half of the tumor samples were used for this analysis. First, the KNN method was used to impute missing values in the z-statistic matrix. Then, the PTCCs for all protein pairs were calculated by a previously developed method for covariance matrix estimation [68]. Finally, the statistical significance of each PTCC was assessed based on a mixture model [69], which computed a local FDR estimate to describe the probability of observing it under null hypothesis. Technically, this step was done by applying the `network.test.edges` function in the GeneNet package (v1.2.16) in R to the estimated PTCCs, with the default parameter settings. For constructing the co-expression network, an edge was added to link a protein pair if and only if the corresponding PTCC was positive and significant (by default, the local FDR < 0.2).

After constructing this network, we specifically took out the sub-network consisting of the identified hypervariable proteins and the associated edges. Module detection for this sub-network was performed by the WGCNA [70] package (v1.70–3) in R. Functional annotation for each detected module was performed by identifying enriched biological processes. In detail, we downloaded GO, Reactome, KEGG, and BioCart gene sets from the MSigDB, and Fisher's exact test was used to search for significantly enriched ones (BH-adjusted p-value < 0.05). Associated network visualization was realized by using Cytoscape (v3.8.0) [71].

Downstream analyses for the PBC TMT data set

Missing values were imputed by using the mean value of five nearest neighbors with the Python function `sklearn.impute.KNNImputer` (v1.2.1). Consensus clustering of the PBC samples was performed by using the ConsensusClusterPlus R package (v1.66). The associated parameter settings were as follows: number of repetitions is 1000 bootstraps; `pItem = 0.8` (resampling 80% of any sample); `pFeature = 1; distance = "euclidean"; clusterAlg = "km"` (K-means). Hierarchical ward-linkage clustering of the proteins used for sample clustering was performed based on the correlation distance with the Python function `scipy.cluster.hierarchy.fcluster` (v1.10.0), with `criterion = "maxclust"`. For a better visualization of the protein expression heat map, smoothing was applied to the corresponding z-statistic matrix by using the Python function scipy.ndimage.gaussian_filter (v1.10.0), with `sigma = 1`. Pathway enrichment analysis for each protein cluster was performed by using the Python function `gseapy.enrichr` (v0.9.5), with `gene_sets = "KEGG_2021_Human"`. Pathways with BH-adjusted p-value < 0.05 were considered to be significant. A log-rank test was performed to compare the survival outcomes among the subgroups of PBC samples. The associated Kaplan–Meier survival curves were plotted by using the Python function `lifelines.KaplanMeierFitter` (v0.27.8). For this survival analysis, only the patients with a follow-up time shorter than 3600 days and longer than 5 days, excluding the ones who died because of unknown reasons or reasons other than the disease, were involved (resulting in 190 patients in total).

Downstream analyses for the HCC label-free data set

The HCC samples were classified into three subgroups by using *K*-means with the Python function `sklearn.cluster.KMeans` (v1.2.1). The significant hypervariable proteins with detected expression in at least half of the HCC samples were hierarchically clustered by using the same method as applied to the PBC data set. Pathway enrichment analysis for each protein cluster and the survival analysis (the log-rank test and the plotting of Kaplan–Meier curves) were conducted with the same methods as well.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03382-9.

Addtional file 1: Supplementary Figures S1-S25. All supplementary figures in this study.

Additional file 2: Supplementary Notes S1-S3. All supplementary notes in this study.

Additional file 3: Supplementary Tables S1-S5. Outputs of zMAP toolset when applied to the data sets used in this study.

Additional file 4: Uncropped western blots.

Additional file 5: Review history.

Acknowledgements
We thank Dr. Jian Xu from St. Jude Children's Research Hospital and Dr. Chris Turck from Max Planck Institute of Psychiatry for their help and suggestions.

Review history
The review history is available as Additional file 4.

Peer review information
Andrew Cosgrove was the primary editor of this article at Genome Biology and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
Z.S. conceived the study; X.G., J.H., S.T., and Z.S. developed the algorithms; X.G. and J.H. analyzed the data under the supervision of S.T. and Z.S.; L.R. and Y.W. performed the experiments under the supervision of X.L. and L.Z.; X.G., S.T., and Z.S. wrote the manuscript with contributions from all the other authors. All authors read and approved the final manuscript.

Funding
This work was supported by the National Basic Research Program of China (2018YFA0800203), the Strategic Priority Research Program of Chinese Academy of Sciences (XDB38040100), and the National Natural Science Foundation of China (32370705 and 32200519).

Availability of data and materials
All the proteomic profiling data used in this study have been previously published. The iTRAQ proteomic data set regarding human erythropoiesis was available at ProteomeXchange under accession number PXD006170 [72]. The TMT proteomic data set of serous ovarian carcinoma cell lines was obtained from ProteomeXchange under accession number PXD019544 [73]. The HCC TMT proteomic data set was obtained from https://www.biosino.org/node/project/detail/OEP000321 [74]. The PBC TMT proteomic data set was obtained from Proteomics Data Commons under accession number PDC000180 [75]. The label-free proteomic data set about early-stage HCC was obtained from ProteomeXchange under accession number PXD006512 [76]. A Web-based application of zMAP is provided at http://bioinfo.cemps.ac.cn/shaolab/zMAP. The source code of zMAP toolset is available at GitHub under the GPL-3 license [77] and is also deposited to Zenodo with a DOI of https://doi.org/10.5281/zenodo.12206918 [78]. Data and code used to generate the analyses and figures in this study are available at Zenodo as well with a DOI of https://doi.org/10.5281/zenodo.13337951 [79].

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China. [2]Key Laboratory of Epigenetic Regulation and Intervention, Shanghai Institute of Biochemistry and Cell Biology, CAS Center for Excellence in Molecular Cell Science, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China. [3]CAS Key Laboratory of Nutrition, Metabolism and Food Safety, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China. [4]Analytical Research Center for Organic and Biological Molecules, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China. [5]Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA.

Received: 5 February 2024 Accepted: 29 August 2024
Published online: 14 October 2024

References

1. Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem. 2003;75:1895–904.
2. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics. 2004;3:1154–69.
3. Zieske LR. A perspective on the use of iTRAQ reagent technology for protein complex and profiling studies. J Exp Bot. 2006;57:1501–8.
4. Rauniyar N, Yates JR 3rd. Isobaric labeling-based relative quantification in shotgun proteomics. J Proteome Res. 2014;13:5293–309.
5. Mitchell DC, Kuljanin M, Li J, Van Vranken JG, Bulloch N, Schweppe DK, Huttlin EL, Gygi SP. A proteome-wide atlas of drug mechanism of action. Nat Biotechnol. 2023;41:845–57.
6. Dou Y, Katsnelson L, Gritsenko MA, Hu Y, Reva B, Hong R, Wang YT, Kolodziejczak I, Lu RJ, Tsai CF, et al. Proteogenomic insights suggest druggable pathways in endometrial carcinoma. Cancer Cell. 2023;41(9):1586-1605.e15.
7. Mun DG, Bhin J, Kim S, Kim H, Jung JH, Jung Y, Jang YE, Park JM, Kim H, Jung Y, et al. Proteogenomic Characterization of Human Early-Onset Gastric Cancer. Cancer Cell. 2019;35(111–124): e110.
8. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016;534:55–62.
9. Moulder R, Bhosale SD, Goodlett DR, Lahesmaa RJMSR. Analysis of the plasma proteome using iTRAQ and TMT-based Isobaric labeling. Mass Spectrom Rev. 2018;37:583–606.
10. Chen X, Sun Y, Zhang T, Shu L, Roepstorff P, Yang F. Quantitative Proteomics Using Isobaric Labeling: A Practical Guide. Genomics Proteomics Bioinformatics. 2021;19:689–706.
11. Sivanich MK, Gu TJ, Tabang DN, Li L. Recent advances in isobaric labeling and applications in quantitative proteomics. Proteomics. 2022;22: e2100256.
12. Dou Y, Kawaler EA, Cui Zhou D, Gritsenko MA, Huang C, Blumenberg L, Karpova A, Petyuk VA, Savage SR, Satpathy S, et al. Proteogenomic Characterization of Endometrial Carcinoma. Cell. 2020;180:729-748.e726.
13. Jiang Y, Sun AH, Zhao Y, Ying WT, Sun HC, Yang XR, Xing BC, Sun W, Ren LL, Hu B, et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature. 2019;567:257-+.
14. Cao L, Huang C, Cui Zhou D, Hu Y, Lih TM, Savage SR, Krug K, Clark DJ, Schnaubelt M, Chen L, et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell. 2021;184:5031-5052.e5026.
15. Trakarnsanga K, Wilson MC, Griffiths RE, Toye AM, Carpenter L, Heesom KJ, Parsons SF, Anstee DJ, Frayne J. Qualitative and quantitative comparison of the proteome of erythroid cells differentiated from human iPSCs and adult erythroid cells by multiplex TMT labelling and nanoLC-MS/MS. PLoS ONE. 2014;9: e100874.
16. Brenes A, Hukelmann J, Bensaddek D, Lamond AI. Multibatch TMT Reveals False Positives, Batch Effects and Missing Values. Mol Cell Proteomics. 2019;18:1967–80.
17. Savitski MM, Mathieson T, Zinn N, Sweetman G, Doce C, Becher I, Pachl F, Kuster B, Bantscheff M. Measuring and Managing Ratio Compression for Accurate iTRAQ/TMT Quantification. J Proteome Res. 2013;12:3586–98.
18. Mahoney DW, Therneau TM, Heppelmann CJ, Higgins L, Benson LM, Zenka RM, Jagtap P, Nelsestuen GL. Bergen III HR. Oberg ALJJopr: Relative quantification: characterization of bias, variability and fold changes in mass spectrometry data from iTRAQ-labeled peptides. 2011;10:4325–33.
19. Huang T, Choi M, Tzouros M, Golling S, Pandya NJ, Banfai B, Dunkley T, Vitek O. MSstatsTMT: Statistical Detection of Differentially Abundant Proteins in Experiments with Isobaric Labeling and Multiple Mixtures. Mol Cell Proteomics. 2020;19:1706–23.
20. Zhu Y, Orre LM, Zhou Tran Y, Mermelekas G, Johansson HJ, Malyutina A, Anders S, Lehtiö J. DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis. Mol Cell Proteomics. 2020;19:1047–57.

21. O'Brien JJ, Raj A, Gaun A, Waite A, Li W, Hendrickson DG, Olsson N, McAllister FE. A data analysis framework for combining multiple batches increases the power of isobaric proteomics experiments. Nat Methods. 2024;21:290–300.
22. Chen YJ, Roumeliotis TI, Chang YH, Chen CT, Han CL, Lin MH, Chen HW, Chang GC, Chang YL, Wu CT, et al. Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. Cell. 2020;182:226-244.e217.
23. Gao Q, Zhu H, Dong L, Shi W, Chen R, Song Z, Huang C, Li J, Dong X, Zhou Y, et al. Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. Cell. 2019;179:561-577.e522.
24. Karp NA, Huber W, Sadowski PG, Charles PD, Hester SV, Lilley KS. Addressing accuracy and precision issues in iTRAQ quantitation. Mol Cell Proteomics. 2010;9:1885–97.
25. Bantscheff M, Boesche M, Eberhard D, Matthieson T, Sweetman G, Kuster BJM, Proteomics C. Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. 2008;7:1702–13.
26. Griffin TJ, Xie H, Bandhakavi S, Popko J, Mohan A, Carlis JV. Higgins LJJopr: iTRAQ reagent-based quantitative proteomic analysis on a linear ion trap mass spectromete. 2007;6:4200–9.
27. Jiang Y, Sun A, Zhao Y, Ying W, Sun H, Yang X, Xing B, Sun W, Ren L, Hu B. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature. 2019;567:257–61.
28. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26:1367–72.
29. Li M, Tu S, Li Z, Tan F, Liu J, Wang Q, Zhang Y, Xu J, Zhang Y, Zhou F, Shao Z. MAP: model-based analysis of proteomic data to detect proteins with significant abundance changes. Cell Discov. 2019;5:40.
30. Liu X, Zhang YY, Ni M, Cao H, Signer RAJ, Li D, Li MS, Gu ZM, Hu ZP, Dickerson KE, et al. Regulation of mitochondrial biogenesis in erythropoiesis by mTORC1-mediated protein translation. Nature Cell Biology. 2017;19:626-+.
31. Zhou C, Walker MJ, Williamson AJ, Pierce A, Berzuini C, Dive C, Whetton AD. A hierarchical statistical modeling approach to analyze proteomic isobaric tag for relative and absolute quantitation data. Bioinformatics. 2014;30:549–58.
32. Zhang Y, Askenazi M, Jiang J, Luckey CJ, Griffin JD, Marto JA. A robust error model for iTRAQ quantification reveals divergent signaling between oncogenic FLT3 mutants in acute myeloid leukemia. Mol Cell Proteomics. 2010;9:780–90.
33. Wood WG. Haemoglobin synthesis during human fetal development. Br Med Bull. 1976;32:282–7.
34. Wilber A, Nienhuis AW, Persons DA. Transcriptional regulation of fetal to adult hemoglobin switching: new therapeutic opportunities. Blood. 2011;117:3945–53.
35. Ow SY, Salim M, Noirel J, Evans C, Wright PC. Minimising iTRAQ ratio compression through understanding LC-MS elution dependence and high-resolution HILIC fractionation. Proteomics. 2011;11:2341–6.
36. Kondoh H, Lleonart ME, Gil J, Wang J, Degan P, Peters G, Martinez D, Carnero A, Beach D. Glycolytic enzymes can modulate cellular life span. Can Res. 2005;65:177–85.
37. Heiden MGV, Cantley LC, Thompson CB. Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. Science. 2009;324:1029–33.
38. Petralia F, Tignor N, Reva B, Koptyra M, Chowdhury S, Rykunov D, Krek A, Ma W, Zhu Y, Ji J, et al. Integrated Proteogenomic Characterization across Major Histological Types of Pediatric Brain Cancer. Cell. 2020;183:1962-1985.e1931.
39. Shrestha R, Llaurado Fernandez M, Dawson A, Hoenisch J, Volik S, Lin YY, Anderson S, Kim H, Haegert AM, Colborne S, et al. Multiomics Characterization of Low-Grade Serous Ovarian Carcinoma Identifies Potential Biomarkers of MEK Inhibitor Sensitivity and Therapeutic Vulnerability. Cancer Res. 2021;81:1681–94.
40. Gudas LJ. Retinoid metabolism: new insights. J Mol Endocrinol. 2022;69:T37-t49.
41. D'Ambrosio DN, Clugston RD, Blaner WS. Vitamin A metabolism: an update. Nutrients. 2011;3:63–103.
42. Almazroo OA, Miah MK, Venkataramanan R. Drug Metabolism in the Liver. Clin Liver Dis. 2017;21:1–20.
43. Bradbury MW. Lipid metabolism and liver inflammation I Hepatic fatty acid uptake: possible role in steatosis. Am J Physiol Gastrointest Liver Physiol. 2006;290:G194-198.
44. Cox DR. Regression models and life-tables. J Roy Stat Soc: Ser B (Methodol). 1972;34:187–202.
45. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017;169:1327-1341.e1323.
46. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
47. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:1–15.
48. Banerjee S, Biehl A, Gadina M, Hasni S, Schwartz DM. JAK-STAT Signaling as a Target for Inflammatory and Autoimmune Diseases: Current and Future Prospects. Drugs. 2017;77:521–46.
49. Muller WA. Mechanisms of leukocyte transendothelial migration. Annu Rev Pathol. 2011;6:323–44.
50. McQuitty CE, Williams R, Chokshi S, Urbani L. Immunomodulatory Role of the Extracellular Matrix Within the Liver Disease Microenvironment. Front Immunol. 2020;11: 574276.
51. Delgado-Rizo V, Martínez-Guzmán MA, Iñiguez-Gutierrez L, García-Orozco A, Alvarado-Navarro A, Fafutis-Morris M. Neutrophil Extracellular Traps and Its Implications in Inflammation: An Overview. Front Immunol. 2017;8:81.
52. Schwartz AB, Campos OA, Criado-Hidalgo E, Chien S, Del Álamo JC, Lasheras JC, Yeh YT. Elucidating the Biomechanics of Leukocyte Transendothelial Migration by Quantitative Imaging. Front Cell Dev Biol. 2021;9: 635263.
53. Chen F, Kang R, Liu J, Tang D. The V-ATPases in cancer and cell death. Cancer Gene Ther. 2022;29:1529–41.
54. Santos-Pereira C, Rodrigues LR, Côrte-Real M. Emerging insights on the role of V-ATPase in human diseases: Therapeutic challenges and opportunities. Med Res Rev. 2021;41:1927–64.
55. Xu QR, Du XH, Huang TT, Zheng YC, Li YL, Huang DY, Dai HQ, Li EM, Fang WK. Role of Cell-Cell Junctions in Oesophageal Squamous Cell Carcinoma. Biomolecules. 2022;12:1378.
56. Knights AJ, Funnell AP, Crossley M, Pearson RC. Holding Tight: Cell Junctions and Cancer Spread. Trends Cancer Res. 2012;8:61–9.
57. Martin TA. The role of tight junctions in cancer metastasis. Semin Cell Dev Biol. 2014;36:224–31.
58. Mani SA, Guo W, Liao MJ, Eaton EN, Ayyanan A, Zhou AY, Brooks M, Reinhard F, Zhang CC, Shipitsin M, et al. The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell. 2008;133:704–15.

59. Bitting RL, Schaeffer D, Somarelli JA, Garcia-Blanco MA, Armstrong AJ. The role of epithelial plasticity in prostate cancer dissemination and treatment resistance. Cancer Metastasis Rev. 2014;33:441–68.
60. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26:1572–3.
61. Hatoum R, Chen JS, Lavergne P, Shlobin NA, Wang A, Elkaim LM, Dodin P, Couturier CP, Ibrahim GM, Fallah A, et al. Extent of Tumor Resection and Survival in Pediatric Patients With High-Grade Gliomas: A Systematic Review and Meta-analysis. JAMA Netw Open. 2022;5: e2226551.
62. Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, Dou Y, Zhang Y, Shi Z, Arshad OA, et al. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell. 2019;177:1035-1049.e1019.
63. Rozanova S, Barkovits K, Nikolov M, Schmidt C, Urlaub H, Marcus K. Quantitative Mass Spectrometry-Based Proteomics: An Overview. Methods Mol Biol. 2021;2228:85–116.
64. Anzenbacher P, Anzenbacherová E. Cytochromes P450 and metabolism of xenobiotics. Cell Mol Life Sci. 2001;58:737–47.
65. Xu J, Shao Z, Glass K, Bauer DE, Pinello L, Van Handel B, Hou S, Stamatoyannopoulos JA, Mikkola HK, Yuan GC, Orkin SH. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. Dev Cell. 2012;23:796–811.
66. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739–40.
67. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28:1353–8.
68. Schäfer J, Strimmer K: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol. 2005, 4:Article32.
69. Efron BJJotASA. Large-scale simultaneous hypothesis testing the choice of a null hypothesis. J Am Stat Assoc2004;99:96–104.
70. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
71. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
72. Liu X, Zhang Y, Ni M, Cao H, Signer RAJ, Li D, Li M, Gu Z, Hu Z, Dickerson KE, et al: Regulation of mitochondrial biogenesisin erythropoiesis by mTORC1-mediated protein translation. *ProteomeXchange*. https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD006170 (2017).
73. Shrestha R, Llaurado Fernandez M, Dawson A, Hoenisch J, Volik S, Lin YY, Anderson S, Kim H, Haegert AM, Colborne S, et al: Multiomics Characterization of Low-Grade Serous Ovarian Carcinoma Identifies Potential Biomarkers of MEK Inhibitor Sensitivity and Therapeutic Vulnerability. ProteomeXchange. https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD019544 (2021).
74. Gao Q, Zhu H, Dong L, Shi W, Chen R, Song Z, Huang C, Li J, Dong X, Zhou Y, et al: Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. *National Omics Data Encyclopedia*. https://www.biosino.org/node/project/detail/OEP000321 (2019).
75. Petralia F, Tignor N, Reva B, Koptyra M, Chowdhury S, Rykunov D, Krek A, Ma W, Zhu Y, Ji J, et al: Integrated Proteogenomic Characterization across Major Histological Types of Pediatric Brain Cancer. *Proteomics Data Commons*. https://pdc.cancer.gov/pdc/study/PDC000180 (2020).
76. Jiang Y, Sun A, Zhao Y, Ying W, Sun H, Yang X, Xing B, Sun W, Ren L, Hu B, et al: Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *ProteomeXchange*. https://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD006512 (2019).
77. Gui. X, Huang. J, Ruan. L, Wu. Y, Guo. X, Cao. R, Zhou. S, Tan. F, Zhu. H, Li. M, et al: zMAP toolset: model-based analysis of large-scale proteomic data via a variance stabilizing z-transformation. *GitHub*. https://github.com/guixiuqi/zMAP (2024).
78. Gui X, Huang J, Ruan L, Wu Y, Guo X, Cao R, Zhou S, Tan F, Zhu H, Li M, et al. zMAP toolset: model-based analysis of large-scale proteomic data via a variance stabilizing z-transformation. Zenodo; 2024. https://doi.org/10.5281/zenodo.12206918.
79. Gui X, Huang J, Ruan L, Wu Y, Guo X, Cao R, Zhou S, Tan F, Zhu H, Li M, et al. zMAP toolset: model-based analysis of large-scale proteomic data via a variance stabilizing z-transformation. 2024. Zenodo. https://doi.org/10.5281/zenodo.13337951.

Publisher's Note