Evaluating computational approaches for comparison of protein expression across cancer indications

Jixin Wang,[1*] Xiaowen Tian,[2*] Wen Yu,[3] Ben Pullman[4], John Bullen Jr,[5] Elaine Hurt,[5] Wenyan Zhong[6]

[1]Oncology Data Science and [2]Statistical Innovation, Oncology R&D, and [3]Data Science and AI, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, MD; [4]Centre for Genomics Research, Discovery Sciences, R&D, AstraZeneca, Gaithersburg, MD; [5]Oncology Targeted Delivery, Research and Early Development, Oncology R&D, AstraZeneca, Gaithersburg, MD; [6]Oncology Data Science, Oncology R&D, AstraZeneca, New York, NY

Running title: Computational approaches for pan-cancer protein expression comparison

Keywords: CPTAC, Proteomics, iBAQ, Normalization, Differential Analysis, TCGA

*Jixin Wang and Xiaowen Tian are co-first authors.

Corresponding author:

Wenyan Zhong

AstraZeneca, One MedImmune Way, Gaithersburg, MD 20878

Phone: (301) 398-4209

Fax: (301) 398-9000

E-mail: wenyan.zhong@astrazeneca.com

ABSTRACT

Background: The National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) recently generated harmonized genomic, transcriptomic, proteomic, and clinical data for over 1,000 tumors across 10 cohorts to facilitate pan-cancer discovery research. However, protein expression comparison across CPTAC cohorts remains challenging due to non-uniform missing data and varying protein expression distribution patterns across tumor types. Here, we present our efforts to evaluate various missing data handling and normalization strategies to create a normalized pan-cancer protein expression dataset.

Results: First, we developed a novel algorithm to select robustly expressed proteins in tumors within any CPTAC cohort. Second, we applied a cohort hybrid imputation approach to protein abundance values from FragPipe within each cohort based on protein expression distribution patterns. Third, we calculated intensity-based absolute quantification using protein abundance values and applied both global and smooth quantile normalization methods. Our results indicate that global quantile normalization ensured identical distribution across cohorts for both tumor and normal tissues, while smooth quantile normalization preserved distribution differences between biological conditions. We assessed our method by comparing differential protein expression analysis results with and without normalization. Additionally, we examined the ranks of protein expression in the normalized CPTAC dataset for selected proteins with high protein-to-RNA expression correlation across CPTAC cohorts. We then compared these protein expression ranks with their RNA expression ranks across corresponding cohorts in The Cancer Genome Atlas (TCGA). Differential protein expression analysis revealed a high level of agreement in the fold change of tumor versus normal tissue within cohorts before and after

normalization. Furthermore, our results indicate that global quantile normalization resulted in the highest cohort rank correlation between CPTAC and TCGA for selected proteins.

Conclusions: In summary, our thorough analysis demonstrates that global quantile normalization surpasses both smooth quantile normalization and no normalization, as evidenced by its higher rank correlation across cancer cohorts between CPTAC and TCGA for selected proteins. The findings suggest that combining cohort hybrid imputation with global quantile normalization is an effective method for creating a normalized CPTAC pan-cancer protein dataset, which can facilitate the study of protein expression across different cancer types.

BACKGROUND

Recent advancements in proteogenomic characterization of large collections of cancer patient samples in the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) [1] have provided unique opportunities for the discovery of cancer targets by using protein expression. CPTAC serves as a repository of global proteomics data of tumor and normal adjacent tissue (NAT) samples from more than 10 tumor types, including colon adenocarcinoma (COAD) [2], breast cancer (BRCA) [3], lung adenocarcinoma (LUAD) [4], ovarian cancer (OV) [5], glioblastoma multiforme (GBM) [6], clear-cell renal cell carcinoma (ccRCC) [7], pancreatic ductal adenocarcinoma (PDAC) [8], uterine corpus endometrial carcinoma (UCEC) [9], head and neck squamous-cell carcinoma (HNSCC) [10], and lung squamous-cell carcinoma (LSCC) [11]. However, proteomics data from these tumor types have been primarily generated by the multiplex tandem mass tag (TMT) approach, which is designed to provide relative protein quantification to reference samples. The relative protein expression data from CPTAC hinders the direct utilization of proteomics data for comparing protein expression across cohorts and

requires additional reprocessing of the original proteomics data to derive absolute quantitation estimations.

Mass spectrometry–based proteomics methods have been widely used for relative and absolute quantification of proteomes and post-translational modifications [12]. FragPipe [13] with the MSFragger search engine can provide protein abundance based on the intensities of the precursor and TMT reporter ions, which can be used as input to derive estimated quantitation. This approach has been recently applied to CPTAC pan-cancer data reprocessing [7, 14, 15]. Intensity-based absolute quantification (iBAQ) [16] values have been further calculated and validated with receptor density values in about 30 selected breast and lung cancer cell lines from the Cancer Cell Line Encyclopedia measured by flow cytometry [15].

Although reprocessed CPTAC pan-cancer proteomics data can be used to estimate absolute protein abundance, the comparison of protein expression levels across CPTAC cohorts remains challenging due to a heterogeneous set of samples, each with distinct missing data and expression patterns. Here, we present our efforts to evaluate and validate various strategies for handling missing data and normalization to generate a normalized pan-cancer protein expression dataset, which can serve as a valuable resource for cancer therapeutics development.

METHODS

Evaluation of CPTAC pan-cancer reprocessed proteomics data

CPTAC pan-cancer samples were reprocessed using the FragPipe computational platform (version 15) with MSFragger (version 3.2) and Philosopher (version 3.4.13) [17] for protein identification and quantitation as well as iBAQ derivation, as described in Wang et al. [15].

FragPipe was used to assess protein abundance in tumor and normal tissues from 10 cancer indications in CPTAC. Log2-transformed protein abundance was used for assessing missing value patterns and developing an imputation strategy. “Cohort” was defined as the unique combination of tissue type and indication, and “missing rate” was defined as the percentage of samples with missing values in each cohort.

Algorithm for selection of robustly expressed proteins

An algorithm was developed to select robustly expressed proteins in the samples as follows. For protein *j* in tumor samples from indication *i*, the percentage of samples with missing abundance values was denoted as M_{*ij*} and the median abundance from non-missing observations was denoted as A_{*ij*}. The corresponding percentile rank of A_{*ij*} within each cohort *i* was obtained and denoted as F_*i*{(A_{*ij*})}. The protein *j* was kept in all cohorts if there was an indication *i* such that F_*i*{(A_{*ij*})} > 50% and M_{*ij*} < 25%. In other words, the protein was kept if the median abundance percentile rank was >50% and the corresponding missing rate was <25% in at least one indication among tumor samples. The determination of these cutoffs was empirical, aimed at achieving a balance between discarding an excessive number of proteins and retaining those with a high rate of missing protein expression measurements.

iBAQ conversion and normalization after cohort hybrid imputation

The R package imputeLCMD [18] was used for imputation after filtering proteins that did not pass the above criteria. This package works via a method selection approach that first determines whether the protein was affected by a missing-not-at-random (MNAR) missingness mechanism, based on the assumption that the distributions of the mean values of protein abundance follow a normal distribution within the cohort [18]. If the determined missing mechanism was MNAR,

quantile regression imputation of left-censored data was applied. Otherwise, k–nearest neighbor (KNN) imputation was applied. Imputation was performed on robustly expressed proteins as determined by the method described above, in log2 scale. iBAQ values were then calculated by dividing the raw imputed protein abundance by the number of theoretically observable peptides of the protein from FragPipe, as described by Wang et al. [15]. The iBAQ values were then normalized using one of two quantile normalization methods: (1) global quantile normalization with quantile normalization [19] applied to all samples across indications and tissue types together; or (2) smooth quantile normalization [20] with the assumption that the statistical distribution of each sample was the same within indications or tissue types, but allowing for differences between groups. Additionally, riBAQ (relative iBAQ) [21] and riBAQ-derived copy number [22] were calculated according to Wang et al. [15]. The iBAQ values obtained from the imputed protein abundance was used to derive riBAQ and copy number values without applying additional normalization methods.

Differential protein expression analysis with normalized iBAQ

To determine whether our missing data imputation and normalization strategy affected downstream analyses, we compared the fold change in differentially expressed proteins (DEP) between tumor and matched normal adjacent tissue (NAT) samples for each indication, using non-normalized, global quantile normalized, and smooth quantile normalized protein iBAQ values.

FragPipe protein abundance–derived iBAQ data from CPTAC tumor and NAT samples were used for DEP analysis using R (version 4.1.1), with empirical Bayes statistics on protein-wise linear models with limma [23] embedded in the DEP package (version 1.16.0) [24]. The

correlations between the log2 fold change from DEP analysis were examined between each comparison using Pearson correlation.

Comparison of the indication ranks of selected proteins between CPTAC and TCGA

We identified proteins whose protein and RNA expression were highly correlated across CPTAC cohorts. Within each indication, we calculated the Pearson correlation between protein and RNA expression using tumor tissues. Then, proteins with a correlation greater than 0.5 across all 10 indications were kept for subsequent analysis. We then compared their protein expression rank across CPTAC cohorts with their RNA expression rank across corresponding cohorts from The Cancer Genome Atlas (TCGA) [25]. Specifically, the median log2(iBAQ) before and after normalization, the median log2(riBAQ) of CPTAC and the median log2(TPM) of TCGA were calculated for those proteins within each indication. Next, indications were ranked by the median log2(iBAQ) before and after normalization, the median log2(riBAQ) and the median log2(TPM) of CPTAC and TCGA, respectively. A weighted rank correlation approach [26] was used to measure rank agreement between each comparison, where the correlation between the protein and RNA expression in CPTAC was used as the weight.

RESULTS

Evaluation of CPTAC pan-cancer reprocessed proteomics data

We used the FragPipe computational platform to assess protein abundance in tumor tissue and normal tissue samples from 10 cancer indications in CPTAC. Log2-transformed protein abundance was used to identify missing value patterns and develop an imputation strategy.

The workflow used in this study is outlined in Fig. 1. We developed an algorithm to select robustly expressed proteins across cohorts as described in the methods section. A total of 15,762 proteins were identified in the union of all cohorts, with 8,419 commonly detected proteins across all indications. The protein identifications per cohort at 1% protein-level false discovery rate (FDR) are shown in Supplementary Table 1. In addition to the 8,419 proteins identified in all cohorts, we included 1,718 proteins that, while not present in all cohorts, are considered robustly expressed in certain indications by our robustly expressed proteins selection algorithm defined in the methods section. Finally, we selected 10,137 proteins for further analysis and the selected proteins per cohort are shown in Supplementary Table 2. Moreover, when we assess the total iBAQ values for each sample, excluding the filtered proteins versus including them, our algorithm shows a mere 1% reduction (see Supplementary Fig. S1).

Evaluation of missing values pattern

Our evaluation of the missing values showed both missing-at-random (MAR) and missing-not-at-random (MNAR) patterns within and across cohorts (Fig. 2a and 2b). As shown in Fig. 2a, the COAD cohort had the highest missing values among all cohorts. The most prominent categories of proteins with missing values are those with a missing rate of 0% to 25% in the MAR pattern and those with a missing rate of 75% to 100% in the MNAR pattern. Interestingly, BRCA exhibits a markedly higher number of proteins with the MNAR pattern in normal tissue compared to tumor tissue (Fig. 2b). Furthermore, proteins with higher missing rates tended to have lower expression levels (Supplementary Fig. S2), suggesting an MNAR pattern. Because it was critical to apply the appropriate missing value imputation method before performing any further downstream analysis, we chose a cohort hybrid algorithm to impute missing values [18].

Assessment of normalization methods

Global quantile normalization ensured identical distribution across cohorts for both tumor and normal tissues, whereas smooth quantile normalization preserved variability across cohorts (Fig. 3a and 3b). The PCA analysis of protein expression across pan-cancer indications revealed that cohort separation based on protein expression is more distinct than tissue types, both before and after normalization. The most noticeable tissue separation occurred following global quantile normalization (Supplementary Fig. S3).

Differential protein expression analysis with normalized and imputed iBAQ values

To determine whether our missing data imputation and normalization strategy affected downstream analyses, we compared the fold change in DEP between tumor tissue and matched normal tissue of selected cohorts (ccRCC, COAD, LSCC, and LUAD), using non-normalized, riBAQ, global quantile normalized, and smooth quantile normalized protein iBAQ values. Our results (Pearson r values) demonstrate a strong correlation in fold change between riBAQ normalized data and non-normalized data (ccRCC, $r = 0.99823$; COAD, $r = 0.99791$; LUAD, $r = 0.99749$; LSCC, $r = 0.99813$) (Fig. 4a), global quantile normalized data and non-normalized data (ccRCC, $r = 0.97262$; COAD, $r = 0.97882$; LUAD, $r = 0.99105$; LSCC, $r = 0.99142$) (Fig. 4b). Similar results were observed when we compared smooth quantile normalized data with non-normalized data (ccRCC, $r = 0.99863$; COAD, $r = 0.99993$; LUAD, $r = 0.99819$; LSCC, $r = 0.99996$) (Fig. 4c). The Pearson r values tend to decrease as more variability is removed from the dataset; for example, the smooth quantile normalization resulted in $r = 0.99863$ while the global quantile normalized resulted in $r = 0.97262$ in ccRCC. However, with global quantile normalization, the correlations are still greater than 0.97 in the four selected indications,

indicating that the global quantile normalization method retained biological differences between tumor tissues and matched normal tissues within cohorts.

Comparison of the indication ranks of selected proteins between CPTAC and TCGA

We identified proteins whose protein and RNA expression were highly correlated ($r > 0.5$) across CPTAC cohorts: ERAP2, CA9, GSTM3, MX1, and STAT1. We then compared their median protein expression rank across CPTAC cohorts with their median RNA expression rank across corresponding TCGA cohorts. A weighted rank correlation approach [26] was used to measure rank agreement between each comparison (e.g., denoted as "v" in the Supplementary Fig. S4). Ovarian cancer (OV) and glioblastoma multiforme (GBM) are excluded from the weighted rank correlation calculation because they have a small number of proteins with $r > 0.5$, and those proteins do not overlap with those selected from other indications. Global quantile normalization had a higher rank correlation (weighted rank correlation of 0.597–0.931) than smooth quantile normalization (weighted rank correlation of 0.168–0.76) and no normalization (weighted rank correlation of 0.168–0.76) (Supplementary Fig. S4).

DISCUSSION

CPTAC recently generated harmonized genomic, transcriptomic, proteomic, and clinical data for over 1,000 tumors in 10 cohorts to facilitate pan-cancer discovery research [14]. To use these data for prioritizing cancer surface antigens and selecting indications in the discovery and development of drug targets, protein expression levels often need to be compared and ranked across cohorts [27, 28]. Efforts to do so are hindered, however, by non-uniform missing data and varying protein expression distribution patterns across tumor types. To evaluate various strategies for missing data handling and normalization for the generation of a normalized pan-

cancer protein expression dataset, we built a computational workflow that incorporates the selection of robustly expressed proteins, cohort hybrid imputation, and quantile normalization [29]. We then evaluated our strategy by comparing indication ranking between CPTAC and TCGA using a set of proteins with high correlation between protein and RNA expression (Fig. 1). To our knowledge, this work represents the first cross-cohort normalized CPTAC pan-cancer data containing estimated absolute protein expression quantification derived from mass spectrometry–based proteomic data.

There are considerable challenges in harmonizing proteomics data across the 10 CPTAC cohorts since they have been generated over multiple years and analyzed by various labs utilizing different processing pipelines. To avoid potential confounding factors, it was necessary to reanalyze TMT global proteomics data with the same pipeline. We first estimated the absolute protein abundance in reprocessed CPTAC pan-cancer data as described previously [30]. Additionally, to create a unified dataset, it is necessary to define a set of proteins that maximizes the identification of proteins across all 10 cohorts with high quality while minimizing the occurrence of missing values. We designed a protein selection algorithm to effectively balance these requirements. Our robustly expressed proteins selection algorithm defined 10,137 proteins, which indicated that more than half of the human proteome (assuming 20,000 protein products) [31, 32] is robustly expressed.

Proteomics data often contain missing values due to a variety of technical factors, such as protein abundance below the instrument limit of detection, poor ionization efficiency, and low signal-to-noise ratio [33-35]. When merging the 10 cohorts, we faced similar issues. Although many algorithms have been developed for missing value imputation[36], it is critical to select the ones that are best suited to the specific characteristics of the data and downstream analysis needs. An

analysis of the missing values in the CPTAC cohort uncovered patterns indicative of both Missing at Random (MAR) and Missing Not at Random (MNAR). Based on these observations, a cohort-hybrid imputation approach was applied. Our study introduces a general framework to manage missing data when comparing protein expression across cohorts. For other specific analyses, one can adopt this model, integrating multiple imputation methods to measure the uncertainty linked to imputation, thereby obtaining more accurate estimations.

Due to the lack of a definitive list of proteins with known expression levels across different tumor types, we chose to use proteins that exhibit a strong correlation between their RNA and protein expression levels. We then assessed our approach by comparing the rank order of these proteins' expression in our normalized dataset to their RNA expression rank in The Cancer Genome Atlas (TCGA) as an indirect method of evaluation. Furthermore, we have implemented a weighted rank correlation method that accounts for the correlation between protein and RNA expression levels. Through our validation method, we've shown a significantly enhanced rank correlation with the application of global cohort normalization as opposed to using a dataset without normalization (see Supplementary Fig. S4). Further, when comparing the fold change of tumor versus matched normal tissues using the normalized dataset versus the original dataset, there was significant agreement (Fig. 4), indicating that our normalization method did not inadvertently eliminate biological variation.

In conclusion, our assessment indicates that a combination of cohort hybrid imputation and global quantile normalization is a reasonable approach to generate a normalized CPTAC pan-cancer protein dataset that could be leveraged to compare protein expression across different cancer types. Nonetheless, the limited number of proteins and indications used for validation in our study represents an important limitation. While our approach was tested using selected

highly correlated proteins and rank alignment with TCGA RNA-Seq data, we recommend further validation using a larger number of proteins with expression levels measured across cohorts using other orthogonal experimental methods such as targeted proteomics to enhance the robustness of our approach.

CONCLUSIONS

We have established a computational workflow for cross-indication comparison from proteomics data, thereby providing a unique data resource to interrogate protein expression across different cancer types. Our results demonstrate that missing data imputation and normalization strategies do not affect downstream analyses. The methodology and the data sources used in this study can serve as valuable resources for cancer research.

Supplementary Information

The online version contains supplementary material available at[https://doi.org/10.5281/zenodo.13376456].

Additional file 1. Supplemental Table 1. The number of proteins and subjects in the pan-cancer indication.

Additional file 2. Supplemental Table 2. The number of selected proteins in each cohort.

Additional file 3. Supplemental Figure 1. Percentage change in the sum of iBAQ values for each sample.

Additional file 4. Supplemental Figure 2. Protein expression missing pattern.

Additional file 5. Supplemental Figure 3. Principal component analysis of protein expression across the pan-cancer indication.

Additional file 6. Supplemental Figure 4. Comparison of expression ranks of highly correlated proteins in CPTAC and TCGA.

Acknowledgments

We gratefully acknowledge the CPTAC for providing open-source proteomics data and Deborah Shuman of AstraZeneca for editing the manuscript and formatting the figures.

Availability of data and materials

The imputed and normalized iBAQ values of FragPipe output of protein abundance estimation for the 10 CPTAC indications used in this work can be downloaded from https://doi.org/10.5281/zenodo.13376456.

Authors' contributions

JW and WZ conceived and designed the study. JW, XT, YW, BP, JB, EH, and WZ analyzed and interpreted the data. All authors contributed to the writing, review, and/or revision of the

manuscript, have approved the final version of the manuscript, and agree to be accountable for all aspects of the work.

Funding

This study was funded by AstraZeneca.

Competing interests

All authors are employees of AstraZeneca and may have stock ownership, options, or interests in the company.

DECLARATIONS

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Abbreviations: BRCA, breast cancer; ccRCC, clear-cell renal cell carcinoma; COAD, colon adenocarcinoma; CPTAC, Clinical Proteomic Tumor Analysis Consortium; DEP, differential protein analysis; GBM, glioblastoma multiforme; HNSCC, head and neck squamous-cell carcinoma; iBAQ, intensity-based absolute quantification; riBAQ, relative intensity-based absolute quantification; KNN, k–nearest neighbor; LSCC, lung squamous-cell carcinoma; LUAD, lung adenocarcinoma; MNAR, missing not at random; NAT, normal adjacent tissue; OV, ovarian cancer; PDAC, pancreatic ductal adenocarcinoma; QRILC, quantile regression imputation of left-censored data; TMT, tandem mass tag; UCEC, uterine corpus endometrial carcinoma.

REFERENCES

1. Lindgren CM, Adams DW, Kimball B, Boekweg H, Tayler S, Pugh SL, Payne SH: **Simplified and unified access to cancer proteogenomic data**. *J Proteome Res* 2021, **20**(4):1902-1910.
2. 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**(4):1035-1049.
3. Krug K, Jaehnig EJ, Satpathy S, Blumenberg L, Karpova A, Anurag M, Miles G, Mertins P, Geffen Y, Tang LC *et al*: **Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy**. *Cell* 2020, **183**(5):1436-1456.
4. Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, Krug K, Petralia F, Li Y, Liang WW, Reva B *et al*: **Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma**. *Cell* 2020, **182**(1):200-225.
5. McDermott JE, Arshad OA, Petyuk VA, Fu Y, Gritsenko MA, Clauss TR, Moore RJ, Schepmoes AA, Zhao R, Monroe ME *et al*: **Proteogenomic characterization of ovarian HGSC implicates mitotic kinases, replication stress in observed chromosomal instability**. *Cell Rep Med* 2020, **1**(1):100004.
6. Wang L-B, Karpova A, Gritsenko MA, Kyle JE, Cao S, Li Y, Rykunov D, Colaprico A, Rothstein JH, Hong R *et al*: **Proteogenomic and metabolomic characterization of human glioblastoma**. *Cancer Cell* 2021, **39**(4):509-528.
7. Clark DJ, Dhanasekaran SM, Petralia F, Pan J, Song X, Hu Y, da Veiga Leprevost F, Reva B, Lih TM, Chang HY *et al*: **Integrated proteogenomic characterization of clear cell renal cell carcinoma**. *Cell* 2019, **179**(4):964-983.

8. 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**(19):5031-5052.

9. 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**(4):729-748.

10. Huang C, Chen L, Savage SR, Eguez RV, Dou Y, Li Y, da Veiga Leprevost F, Jaehnig EJ, Lei JT, Wen B *et al*: **Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma**. *Cancer Cell* 2021, **39**(3):361-379.

11. Satpathy S, Krug K, Jean Beltran PM, Savage SR, Petralia F, Kumar-Sinha C, Dou Y, Reva B, Kane MH, Avanessian SC *et al*: **A proteogenomic portrait of lung squamous cell carcinoma**. *Cell* 2021, **184**(16):4348-4371.

12. Ankney JA, Muneer A, Chen X: **Relative and absolute quantitation in mass spectrometry-based proteomics**. *Annu Rev Anal Chem* 2018, **11**(1):49-77.

13. Yu F, Teo GC, Kong AT, Fröhlich K, Li GX, Demichev V, Nesvizhskii AI: **Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform**. *Nat Commun* 2023, **14**(1):4154.

14. Li Y, Dou Y, Da Veiga Leprevost F, Geffen Y, Calinawan AP, Aguet F, Akiyama Y, Anand S, Birger C, Cao S *et al*: **Proteogenomic data and resources for pan-cancer analysis**. *Cancer Cell* 2023, **41**(8):1397-1406.

15. Wang J, Yu W, D'Anna R, Przybyla A, Wilson M, Sung M, Bullen J, Hurt E, D'Angelo G, Sidders B *et al*: **Pan-Cancer Proteomics Analysis to Identify Tumor-Enriched and**

Highly Expressed Cell Surface Antigens as Potential Targets for Cancer Therapeutics. *Mol Cell Proteomics* 2023, **22**(9):100626.

16. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: **Global quantification of mammalian gene expression control**. *Nature* 2011, **473**(7347):337-342.

17. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI: **MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics**. *Nat Methods* 2017, **14**(5):513-520.

18. Lazar C, Burger T, Wieczorek S: **imputeLCMD: a collection of methods for left-censored missing data imputation**. In.; 2022.

19. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.

20. Hicks SC, Okrah K, Paulson JN, Quackenbush J, Irizarry RA, Bravo HC: **Smooth quantile normalization**. *Biostatistics* 2018, **19**(2):185-198.

21. Shin JB, Krey JF, Hassan A, Metlagel Z, Tauscher AN, Pagana JM, Sherman NE, Jeffery ED, Spinelli KJ, Zhao H *et al*: **Molecular architecture of the chick vestibular hair bundle**. *Nat Neurosci* 2013, **16**(3):365-374.

22. Milo R: **What is the total number of protein molecules per cell volume? A call to rethink some published values**. *Bioessays* 2013, **35**(12):1050-1055.

23. Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3**(1).

24. Zhang X, Smits AH, van Tilburg GBA, Ovaa H, Huber W, Vermeulen M: **Proteome-wide identification of ubiquitin interactions using UbIA-MS**. *Nat Protoc* 2018, **13**(3):530-550.

25. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, Ally A, Balasundaram M, Birol I, Butterfield YSN *et al*: **The Cancer Genome Atlas Pan-Cancer analysis project**. *Nature Genetics* 2013, **45**(10):1113-1120.

26. Plaia A, Buscemi S, Sciandra M: **Consensus among preference rankings: a new weighted correlation coefficient for linear and weak orderings**. *Adv Data Anal Classif* 2021, **15**(4):1015-1037.

27. Jarnuczak AF, Najgebauer H, Barzine M, Kundu DJ, Ghavidel F, Perez-Riverol Y, Papatheodorou I, Brazma A, Vizcaíno JA: **An integrated landscape of protein expression in human cancer**. *Sci Data* 2021, **8**(1):115.

28. Kosti I, Jain N, Aran D, Butte AJ, Sirota M: **Cross-tissue analysis of gene and protein expression in normal and cancer tissues**. *Sci Rep* 2016, **6**(1):24799.

29. Gong TQ, Jiang YZ, Shao C, Peng WT, Liu MW, Li DQ, Zhang BY, Du P, Huang Y, Li FF *et al*: **Proteome-centric cross-omics characterization and integrated network analyses of triple-negative breast cancer**. *Cell Rep* 2022, **38**(9):110460.

30. !!! INVALID CITATION !!! [15].

31. Aebersold R, Agar JN, Amster IJ, Baker MS, Bertozzi CR, Boja ES, Costello CE, Cravatt BF, Fenselau C, Garcia BA *et al*: **How many human proteoforms are there?** *Nat Chem Biol* 2018, **14**(3):206-214.

32. Ponomarenko EA, Poverennaya EV, Ilgisonis EV, Pyatnitskiy MA, Kopylov AT, Zgoda VG, Lisitsa AV, Archakov AI: **The Size of the Human Proteome: The Width and Depth**. *Int J Anal Chem* 2016, **2016**:7436849.

33. McGurk KA, Dagliati A, Chiasserini D, Lee D, Plant D, Baricevic-Jones I, Kelsall J, Eineman R, Reed R, Geary B *et al*: **The use of missing values in proteomic data-independent acquisition mass spectrometry to enable disease activity discrimination**. *Bioinformatics* 2020, **36**(7):2217-2223.

34. Jin L, Bi Y, Hu C, Qu J, Shen S, Wang X, Tian Y: **A comparative study of evaluating missing value imputation methods in label-free proteomics**. *Sci Rep* 2021, **11**(1):1760.

35. Gardner ML, Freitas MA: **Multiple imputation approaches applied to the missing value problem in bottom-up proteomics**. *Int J Mol Sci* 2021, **22**(17):9650.

36. Lazar C, Gatto L, Ferro M, Bruley C, Burger T: **Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies**. *J Proteome Res* 2016, **15**(4):1116-1125.

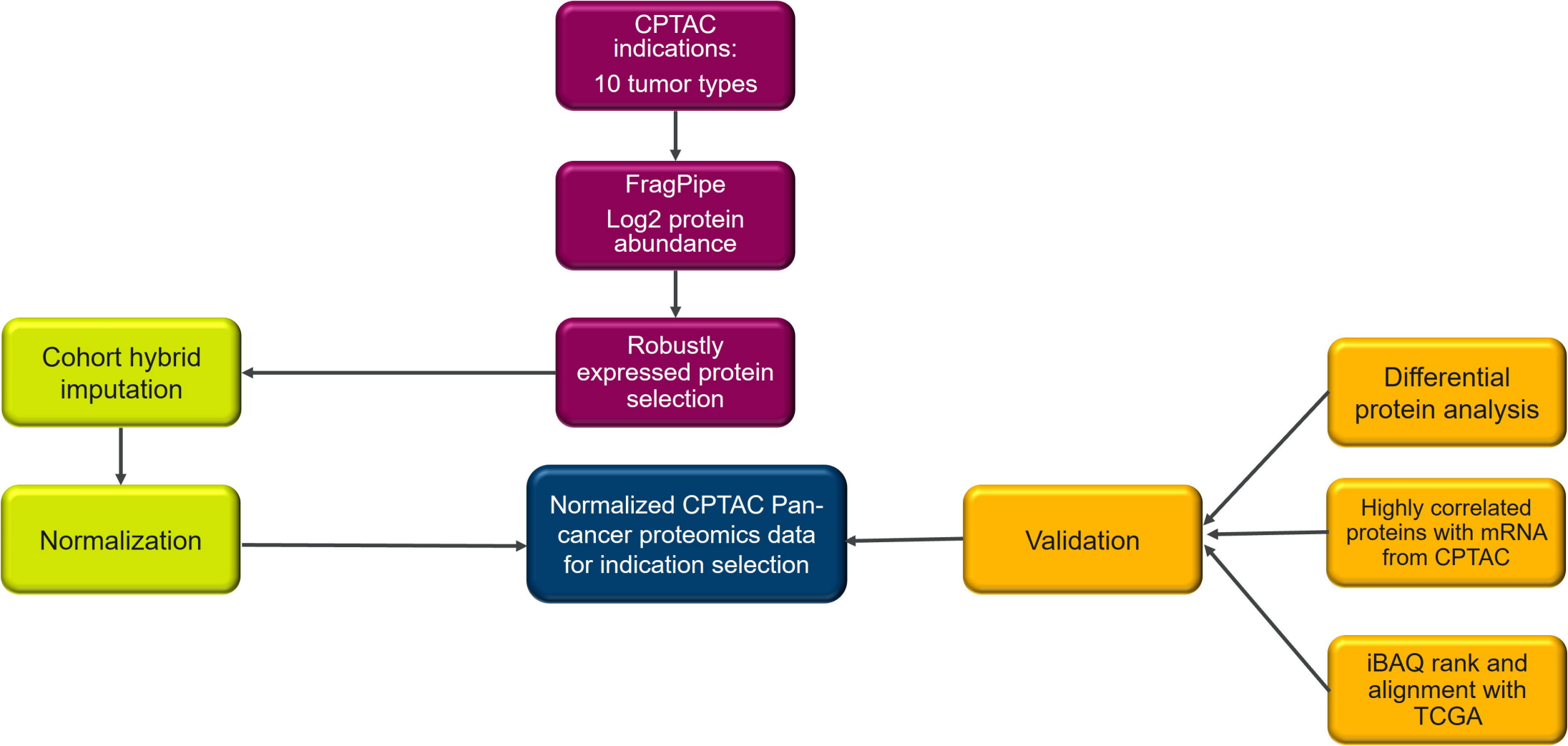
FIGURE LEGENDS

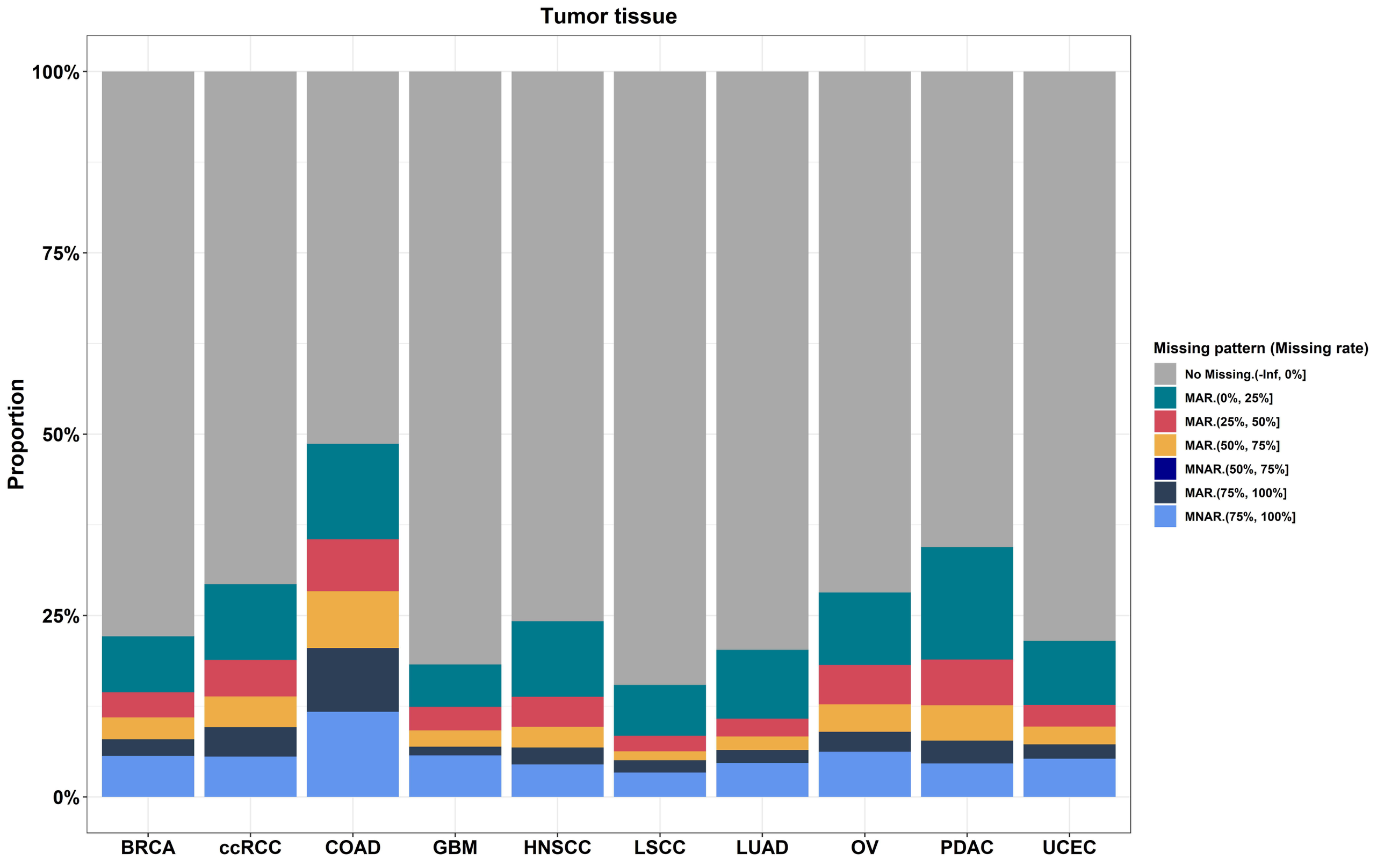
Fig. 1 Workflow for a computational approach that enables the comparison of protein expression across CPTAC cohorts. Tissue samples from 10 cancer indications in CPTAC were processed through FragPipe. Protein abundance data were analyzed by iBAQ to derive estimated absolute protein abundance. A cohort hybrid imputation strategy was developed to address missing values. Quantile normalization was applied to iBAQ values. Imputed and normalized iBAQ values were used for validation with differential protein analysis, ranking indications by protein expression and comparing these ranks with those derived from TCGA RNA-Seq data.

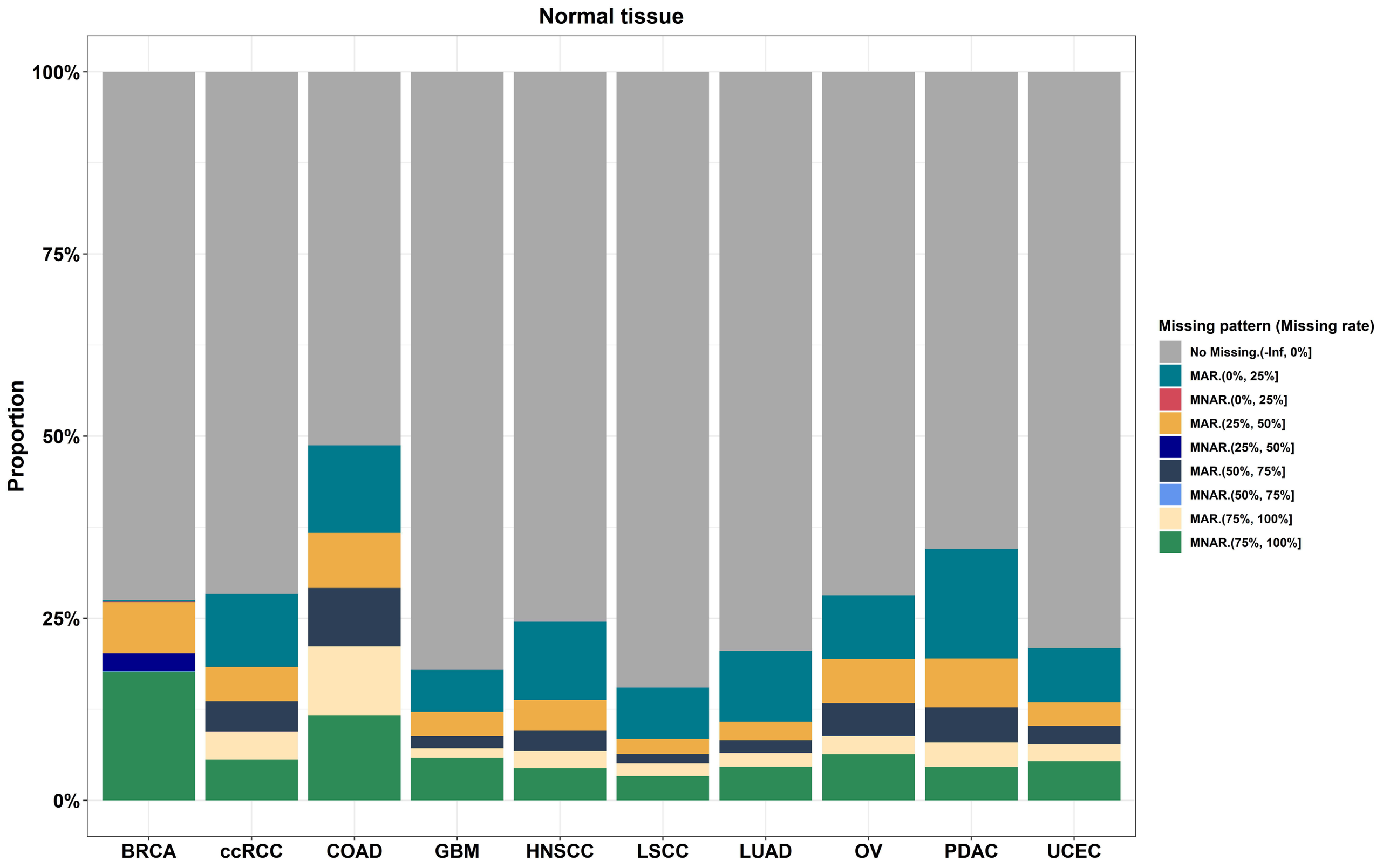
Fig. 2 CPTAC missing values pattern assessment in **a** tumor and **b** normal tissue. “Cohort” was defined as the unique combination of tissue type and indication. “Missing rate” was defined as the percentage of samples with missing values in each cohort.

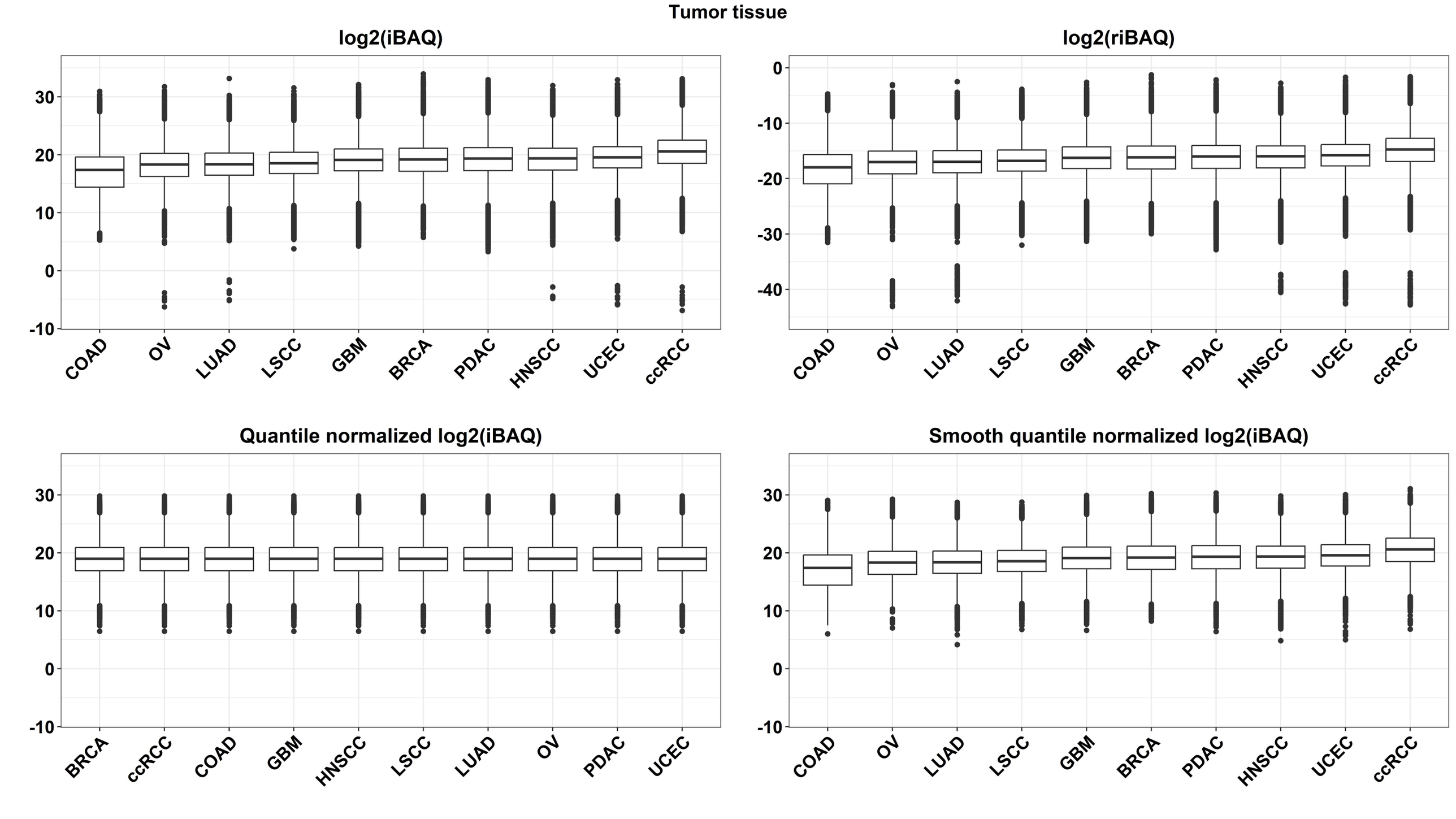
Fig. 3 Quantile normalization. **a** Quantile normalization applied to derived iBAQ values after imputation in **a** tumor and **b** normal tissue. Two quantile normalization methods, global and smooth quantile normalization, were used.

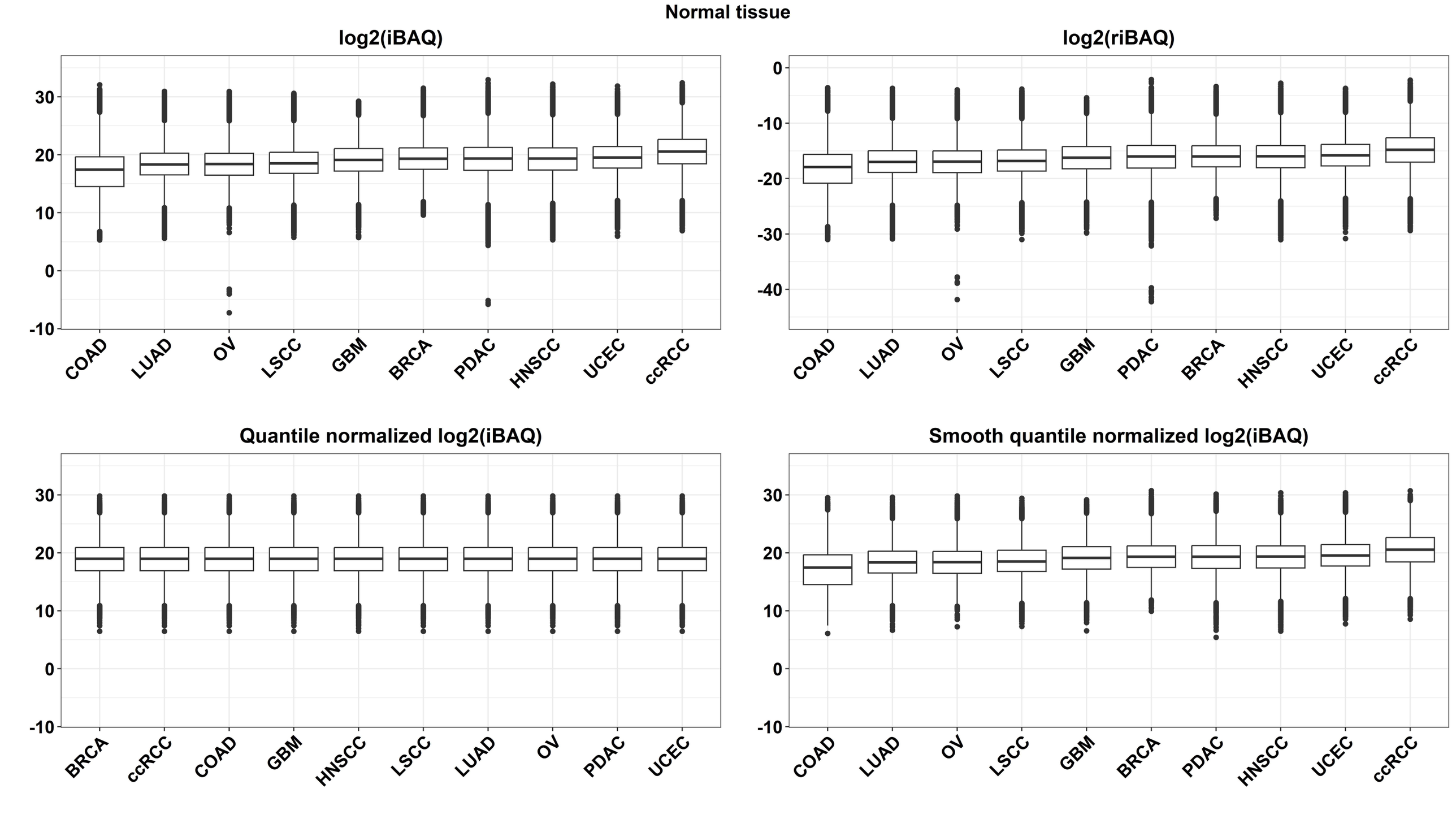
Fig. 4 Correlation between fold change from differential expression analysis. **a** riBAQ normalization and no normalization. **b** Global quantile normalization and no normalization. **c** Smooth quantile normalization and no normalization.

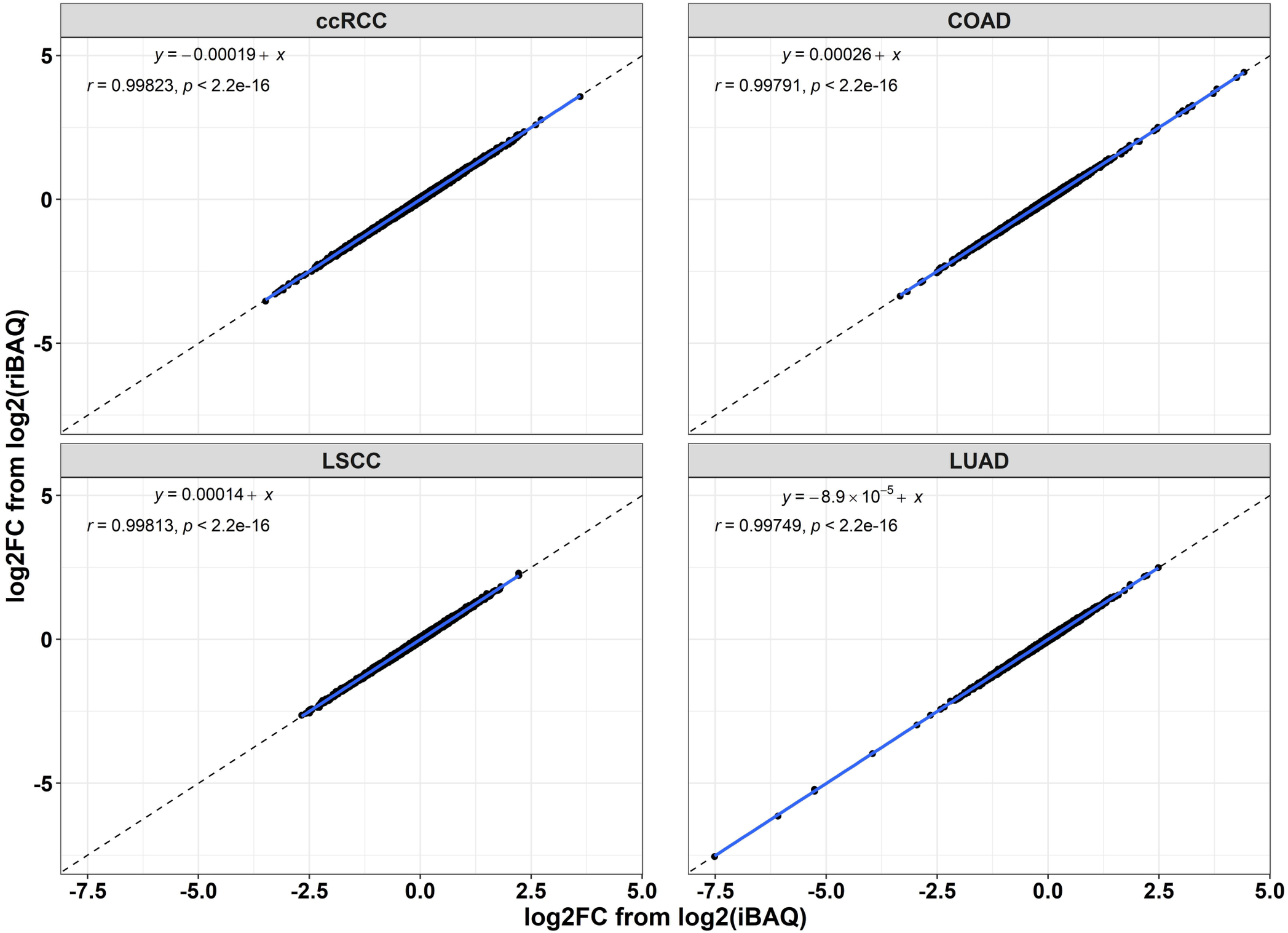

CPTAC indications:
10 tumor types
FragPipe
Log2 protein abundance
Robustly expressed protein selection
Cohort hybrid imputation
Normalization
Normalized CPTAC Pan-cancer proteomics data for indication selection
Validation
Differential protein analysis
Highly correlated proteins with mRNA from CPTAC
iBAQ rank and alignment with TCGA

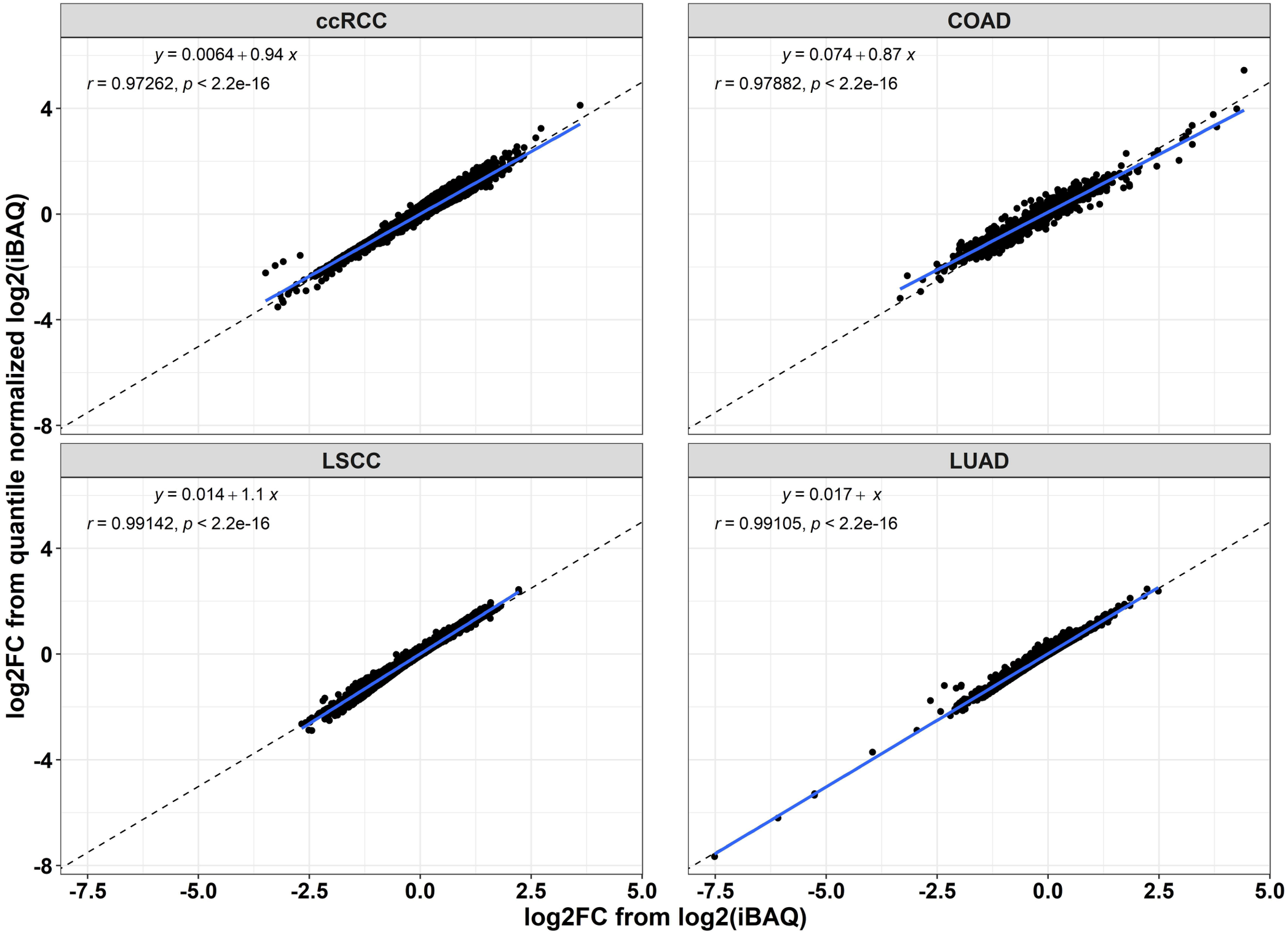

Tumor tissue
Proportion
100%
75%
50%
25%
0%
BRCA
ccRCC
COAD
GBM
HNSCC
LSCC
LUAD
OV
PDAC
UCEC
Missing pattern (Missing rate)
No Missing.(-Inf, 0%]
MAR.(0%, 25%]
MAR.(25%, 50%]
MAR.(50%, 75%]
MNAR.(50%, 75%]
MAR.(75%, 100%]
MNAR.(75%, 100%]

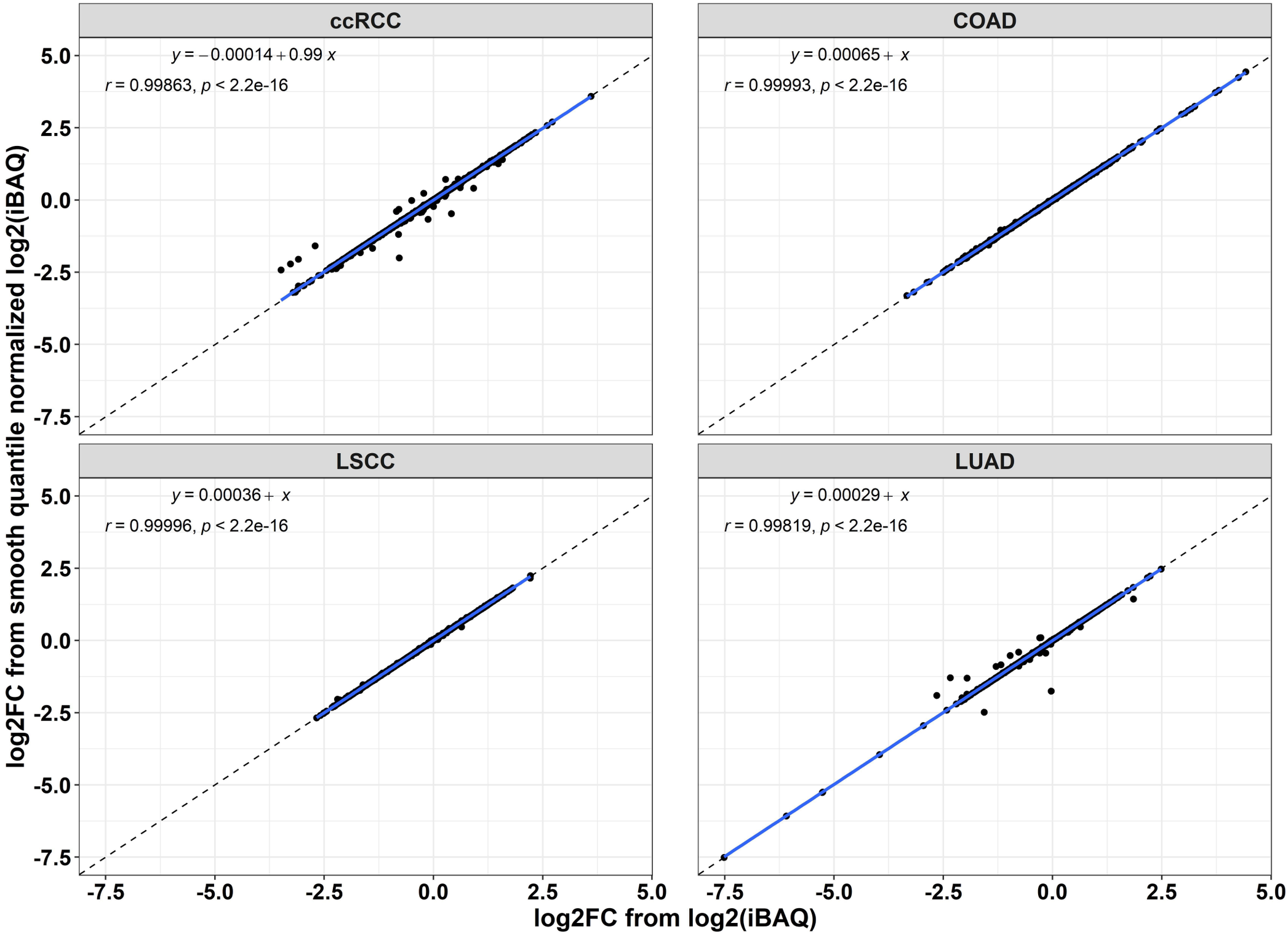

Normal tissue
Proportion
100%
75%
50%
25%
0%
BRCA
ccRCC
COAD
GBM
HNSCC
LSCC
LUAD
OV
PDAC
UCEC
Missing pattern (Missing rate)
No Missing.(-Inf, 0%]
MAR.(0%, 25%]
MNAR.(0%, 25%]
MAR.(25%, 50%]
MNAR.(25%, 50%]
MAR.(50%, 75%]
MNAR.(50%, 75%]
MAR.(75%, 100%]
MNAR.(75%, 100%]


Tumor tissue
log2(iBAQ)
30
20
10
0
-10
COAD
OV
LUAD
LSCC
GBM
BRCA
PDAC
HNSCC
UCEC
ccRCC
log2(riBAQ)
0
-10
-20
-30
-40
COAD
OV
LUAD
LSCC
GBM
BRCA
PDAC
HNSCC
UCEC
ccRCC
Quantile normalized log2(iBAQ)
30
20
10
0
-10
BRCA
ccRCC
COAD
GBM
HNSCC
LSCC
LUAD
OV
PDAC
UCEC
Smooth quantile normalized log2(iBAQ)
30
20
10
0
-10
COAD
OV
LUAD
LSCC
GBM
BRCA
PDAC
HNSCC
UCEC
ccRCC


Normal tissue
log2(iBAQ)
30
20
10
0
-10
COAD
LUAD
OV
LSCC
GBM
BRCA
PDAC
HNSCC
UCEC
ccRCC
log2(riBAQ)
0
-10
-20
-30
-40
COAD
LUAD
OV
LSCC
GBM
PDAC
BRCA
HNSCC
UCEC
ccRCC
Quantile normalized log2(iBAQ)
30
20
10
0
-10
BRCA
ccRCC
COAD
GBM
HNSCC
LSCC
LUAD
OV
PDAC
UCEC
Smooth quantile normalized log2(iBAQ)
30
20
10
0
-10
COAD
LUAD
OV
LSCC
GBM
BRCA
PDAC
HNSCC
UCEC
ccRCC


ccRCC
y = −0.00019 + x
r = 0.99823, p < 2.2e-16
COAD
y = 0.00026 + x
r = 0.99791, p < 2.2e-16
LSCC
y = 0.00014 + x
r = 0.99813, p < 2.2e-16
LUAD
y = −8.9 × 10^−5 + x
r = 0.99749, p < 2.2e-16
log2FC from log2(riBAQ)
log2FC from log2(iBAQ)
5
0
-5
-7.5
-5.0
-2.5
0.0
2.5
5.0


ccRCC
y = 0.0064 + 0.94 x
r = 0.97262, p < 2.2e-16
COAD
y = 0.074 + 0.87 x
r = 0.97882, p < 2.2e-16
LSCC
y = 0.014 + 1.1 x
r = 0.99142, p < 2.2e-16
LUAD
y = 0.017 + x
r = 0.99105, p < 2.2e-16
log2FC from quantile normalized log2(iBAQ)
log2FC from log2(iBAQ)
4
0
-4
-8
-7.5
-5.0
-2.5
0.0
2.5
5.0


ccRCC
y = −0.00014 + 0.99 x
r = 0.99863, p < 2.2e-16
COAD
y = 0.00065 + x
r = 0.99993, p < 2.2e-16
LSCC
y = 0.00036 + x
r = 0.99996, p < 2.2e-16
LUAD
y = 0.00029 + x
r = 0.99819, p < 2.2e-16
log2FC from smooth quantile normalized log2(iBAQ)
log2FC from log2(iBAQ)
5.0
2.5
0.0
-2.5
-5.0
-7.5