Building flexible and robust analysis frameworks for molecular subtyping of cancers

Pedersen, Christina Bligaard; Campos, Benito; Rene, Lasse; Wegener, Helene Scheel; Krishnan, Neeraja M.; Panda, Binay; Vitting-Seerup, Kristoffer; Rossing, Maria; Bagger, Frederik Otzen; Olsen, Lars Rønn

Published in:
Molecular Oncology

Link to article, DOI:
10.1002/1878-0261.13580

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Molecular Oncology

FEBS PRESS science publishing by scientists

Building flexible and robust analysis frameworks for molecular subtyping of cancers

Christina Bligaard Pedersen[1,2], Benito Campos[1], Lasse Rene[1], Helene Scheel Wegener[1], Neeraja M. Krishnan[3], Binay Panda[1,3,4], Kristoffer Vitting-Seerup[1], Maria Rossing[2], Frederik Otzen Bagger[2] and Lars Rønn Olsen[1]

1 Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark
2 Center for Genomic Medicine, Rigshospitalet – Copenhagen University Hospital, Denmark
3 School of Biotechnology, Jawaharlal Nehru University, New Delhi, India
4 Special Centre for Systems Medicine, Jawaharlal Nehru University, New Delhi, India

Keywords
bioinformatics workflows; clinical bioinformatics; molecular subtyping; sample classification

Correspondence
L. R. Olsen, Department of Health Technology, Technical University of Denmark, Anker Engelunds Vej 1, 2800 Kongens Lyngby, Denmark
E-mail: lronn@dtu.dk

(Received 7 June 2023, revised 19 October 2023, accepted 28 December 2023, available online 7 January 2024)

doi:10.1002/1878-0261.13580

Molecular subtyping is essential to infer tumor aggressiveness and predict prognosis. In practice, tumor profiling requires in-depth knowledge of bioinformatics tools involved in the processing and analysis of the generated data. Additionally, data incompatibility (e.g., microarray versus RNA sequencing data) and technical and uncharacterized biological variance between training and test data can pose challenges in classifying individual samples. In this article, we provide a roadmap for implementing bioinformatics frameworks for molecular profiling of human cancers in a clinical diagnostic setting. We describe a framework for integrating several methods for quality control, normalization, batch correction, classification and reporting, and develop a use case of the framework in breast cancer.

1. Introduction

Large-scale multi-omics profiling experiments during the last decade have defined tumor subtypes in glioblastoma, breast cancer, squamous cell lung cancer, and colorectal cancer [1–4]. Morphologically similar tumor samples can differ substantially in underlying genetic aberrations and changes in gene expression. For several cancers, specific tumor subtypes have been linked to different treatment responses and prognoses [5–7], providing an important approach for patient stratification and its application in precision medicine.

In the case of breast cancer, the definition of subtypes has a history extending two decades [8–10] and tumor profiling is already integrated into many modern clinical workflows [11]. More than 100 expression profiles have been proposed [12], however, in practice, the most widely used subtyping methods are the PAM50 [9] and CIT [10] gene signatures. PAM50 is used to classify samples into one of four subtypes:

Abbreviations
BasL, basal-like; BRCA, breast invasive carcinoma; CIT, Cartes d'Identité des Tumeurs program; CITBCMST, CIT breast cancer molecular subtypes prediction; EGA, European genome-phenome archive; ER, estrogen receptor; ERBB2, receptor tyrosine-protein kinase erbB-2; ESR1, estrogen receptor 1; FDA, Food and Drug Administration; FISH, fluorescence *in situ* hybridization; GTEx, genotype-tissue expression; HER2, human epidermal growth factor receptor 2; IHC, immunohistochemistry; kNN, *k* nearest neighbors; LumA, luminal A; LumB, luminal B; LumC, luminal C; mApo, molecular-apocrine; NormL, normal-like; PAM50, Prediction Analysis of Microarray 50; PCA, principal component analysis; PCs, principal components; PR, progesterone receptor; qRT-PCR, quantitative reverse transcription polymerase chain reaction; RMA, robust multi-array average; RSEM, RNA-Seq by expectation–maximization; ssGSEA, single-sample gene set enrichment analysis; SVA, surrogate variable analysis; TCGA, The cancer genome atlas; TPM, transcripts per million.

Luminal A, Luminal B, HER2-enriched, and Basal-like. A fifth subtype derived from normal breast tissue was also included in the original work: Normal-like [9]. Classification into these four cancer subtypes was originally based on research microarray and qRT-PCR data, but later an FDA-approved platform was developed [13]. Using a similar approach relying on unsupervised clustering, Guedj et al. [10] grouped breast cancer samples into six subtypes; lumA, lumB, lumC, mApo, basL, and normL. Additional subtypes, for example, Claudin-low group [14], were discussed as a subtype in the past but later inferred as a phenotype extending across the spectrum of the other groups [15].

Despite their wide use, these subtyping methods are associated with technical challenges in a diagnostic laboratory. For example, most sample sizes in clinical settings are small and may include as few as a single sample from an individual patient. However, the subtyping methods depend on the distribution of data from a group of samples. Additionally, classification results may differ depending on whether a single sample or a batch of samples are used [16]. Additionally, subtype assignment depends on the amount of non-tumor cells in a given sample, and thus should consider tumor purity. Lastly, many pivotal subtyping methods were derived from microarray data, while most laboratories have moved or are moving towards RNA sequencing for gene expression profiling. The incompatibility of these two data types makes it challenging to leverage established methods while transitioning to more prevalent and unbiased gene expression profiling technologies.

Here, we describe practical bioinformatics solutions to address these challenges with a framework and apply the resulting framework to the PAM50 [9] and CIT [10] classifiers. We provide a generally applicable roadmap for implementing data and bioinformatics tools for robust inference of cancer subtypes in a clinical setting, both from raw and processed RNA sequencing data and on a per-sample basis. Finally, we discuss several other molecular profiling features which, together with subtyping, can be combined into a comprehensive report to support diagnostic and clinical insights.

2. Materials and methods

2.1. Data

2.1.1. TCGA

We used data from the cancer genome atlas (TCGA) for assessing samples in our use case for breast cancer. We downloaded gene expression data from the Xena Browser [17] as transcripts per million (TPM) quantified using RSEM and TOIL, mapped to Gencode GRCh38.p3. For visualization purposes, we selected 20 random samples from each cancer type.

2.1.2. GTEx

We downloaded data from the Genotype-Tissue Expression (GTEx) project for assessing whether our use case samples resembled healthy breast tissue. Gene expression data were downloaded from the UCSC Xena project [17] as TPM quantified using RSEM and TOIL, mapped to Gencode GRCh38.p3. For the visualizations presented here, we selected 20 random samples from each tissue type.

2.1.3. CIT reference data

The CIT microarray (Affymetrix HG-U133 Plus 2.0) training data were downloaded from ArrayExpress (accession: E-MTAB-365) and a subset of data for the 355 core samples and 375 probes defined by the original publication [10]. The .CEL files were read and RMA-normalized using the AFFY R package [18]. The subtype for each of the samples was available in the data of the CITBCMST R package [10] (package no longer maintained).

2.1.4. PAM50 reference data

We used the breast invasive carcinoma samples from the TCGA for which a PAM50 subtype has been assigned, to train a classifier for the PAM50 subtyping scheme. The expression of the 50 genes as defined by Parker et al. [9] were used for classification. The format of the data is TPM transformed RNA-seq data.

2.1.5. Use case data

The use case data set comprises RNA-seq data from the tumors of 57 breast cancer patients sequenced at the Breast Cancer Translational Research Laboratory, Institut Jules Bordet, Brussels, Belgium. It was originally presented by Fumagalli et al. [19] and derived from EGA (accession number EGAD00001000627). For visualization purposes, 10 of these samples (named HER2-03, HER2-21, LUMA-18, LUMA-24, LUMA-27, LUMA-29, LUMB-01, LUMB-17, TN-18, and TN-22) were selected to be the use case data set in this study, such that the included samples spanned the four IHC- and grading-based subtypes from the original work (Table S1). The full analysis of all 57 samples

was also performed. For CIT classification, RNA sequencing reads were mapped to the Affymetrix HG-U133 Plus 2.0 probe set sequences [20], and subset to the 375 CIT probes. For comparison to GTEx, TCGA, and PAM50 classification using TCGA-BRCA as a reference, reads were mapped to Gencode GRCh38.p3. Furthermore, the reads were also subset to the 375 CIT probes for PAM classification.

2.2. Methods

2.2.1. Quality assessment

We used ARRAYQUALITYMETRICS v3.54.0 to assess data quality for microarray data [21] and FASTQC v0.12.0 [22] for RNA sequencing data.

2.2.2. Tumor purity estimates

We estimated tumor purity using a functional class scoring based method, ESTIMATE v1.0.13, which provides enrichment scores for a stromal content gene set and an immune infiltrate gene set in a given sample [23]. These scores are then aggregated into a score that serves as an indicator of tumor purity. As this method works by calculating the enrichment of two gene sets, it is important to ensure overlap between the gene symbols in the gene sets and the samples.

2.2.3. Data harmonization

Data sets were harmonized using either ComBat [24] (implemented in the SVA package v3.46.0) or simple rank transformation of expression values.

2.2.4. Dimensionality reduction and projection

All dimensionality reduction was done using principal component analysis (PCA) with the STATS package in R. Initial PCA spaces were constructed on reference genes (centered and scaled) and additional samples were projected into the reference space by multiplying the rotations from the reference space with the additional sample vectors. For construction of the TCGA and GTEx reference spaces, features were first reduced by subsetting significantly differentially expressed genes between samples from each subtype versus the rest of subtypes collectively, using a Mann–Whitney U test. For the TCGA BRCA reference PCA, the 1092 samples with assigned PAM50 subtypes were subsetted to genes with a $\log_2$-transformed fold change greater than 1 or smaller than -1. The choice of this filtering setting was based on cross-validation performance. For construction of the CIT PCA reference space, we used the 375 probe sets originally defined by the authors.

2.2.5. Definition of subtype-specific gene sets

Subtype-specific gene sets were defined using a Mann–Whitney U test on samples in each subtype versus the rest of the samples. Features were subsetted first based on Bonferroni multiple testing corrected P values lower than 0.05, and then by $\log_2$-fold changes greater than 1.

2.2.6. Classification

Classification of samples was carried out using three different approaches: k-Nearest Neighbor (kNN) using the e1071 v1.7-13 package in R, distance-to-centroid, and subtype gene set single sample gene set enrichment using the SINGSCORE package v1.18.0 in R [25]. For the kNN classifier, the best k was empirically determined by leave-one-out cross-validation. Then, a winner-takes-all approach was used to assign subtypes. For the distance-to-centroid classifier, centroids were calculated as the mean expression of genes in each class, and the minimum centroid distance subtype was assigned. It has been proposed to shrink the centroid means to offset the effects of outliers on the mean. We observed that defining centroids on the median expression worked just as well. For centroid distance using gene expression, Euclidean distance was used, and for the rank transformed data, we used the Kendall tau distance. For subtype gene set single sample gene set enrichment, we used Singscore. For performance evaluation, we used a leave-one-out approach and calculated precision and recall for each subtype as well as a subtype frequency weighted accuracy.

3. Results

3.1. Pipeline for molecular subtyping

A robust workflow for subtyping involves several steps (Fig. 1). Briefly, the process starts with data preparation where one or more reference data sets (e.g., subtype training samples) and a test data set (new samples to be profiled) are processed. The latter step may also involve some additional steps, to make test data compatible with additional reference sets like the TCGA or GTEx sets. After processing each data set separately, it is necessary to harmonize them to ensure that they can be directly compared, after which the actual subtyping can take place. In this paper, we describe each step of the pipeline and exemplify the analysis

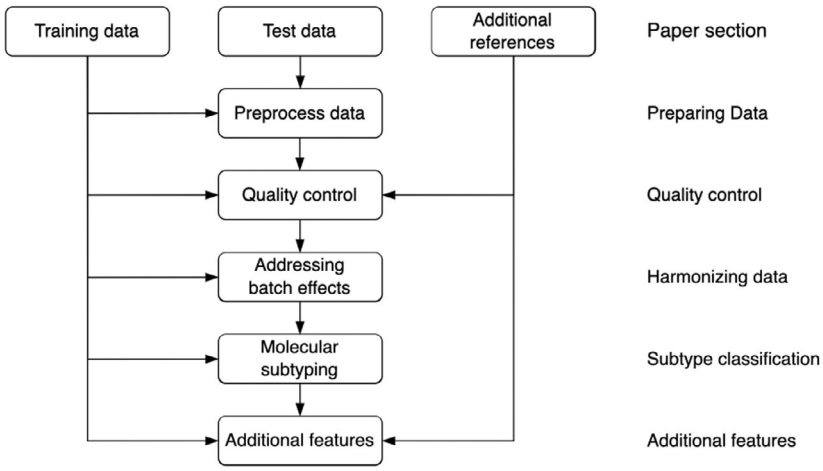

Fig. 1. Proposed pipeline for molecular characterization of cancer. The section describing each step of the workflow is indicated on the right.

using a data set of 10 breast cancer samples analyzed with RNA sequencing. However, the workflow is designed to be applicable, in principle, to any other samples, cancers, or data types.

3.2. Preparing data

The first step in building a subtyping framework is to prepare the data set(s) that should serve as a reference for the subtypes. This set will typically be published along with the method describing the subtyping scheme, for example, raw data from the CIT subtyping method published by Guedj et al. [10]. As a general rule of thumb, the training data should be pre-processed in accordance with the methods used by the authors of the original publication, although this may not affect classification performance in practice.

For samples analyzed with the same Affymetrix platform, extracting the matching probe set intensities is trivial. However, gene expression profiling has largely shifted towards RNA sequencing, as this technique is now both cheaper and more comprehensive. To address this technology incompatibility, we previously devised a method for integrating RNA sequencing data with DNA microarray data [20]. In short, rather than mapping RNA sequencing reads to a reference transcriptome, we propose mapping the reads to the probe set target sequences of the microarray. This also ensures full overlap between the features of the reference set and the test set, which is necessary for harmonization, visualization, and classification. An additional note regarding the use of sequencing data, is that length-normalization is necessary for inter-gene comparisons, on which subtyping is typically highly dependent. One approach is to use transcript per million-normalization.

3.3. Quality control

Quality control has two major components: technical and non-technical quality control. Technical quality control strictly deals with the quality of the data from an instrumental point of view. Depending on the platform used, different software packages can be applied to assess the raw data. Data files with poor quality data should be discarded as this can severely affect the results of the downstream analyses. Quality assessment of microarray data can be done using tools such as the BIOCONDUCTOR package ARRAYQUALITYMETRICS [21], while RNA sequencing data can be assessed using FASTQC [22].

Non-technical quality issues, such as sample impurity or sample swaps can be harder to detect. However, the majority of these non-technical quality parameters can be estimated by comparing sample characteristics to appropriate reference data sets. Tumor purity naturally varies, and may not be an issue for subtyping classification, unless the purity of a sample is outside of the range of the reference set. Tumor purity can be estimated using the software package ESTIMATE [23], which utilizes ssGSEA [26] to calculate enrichment of an immune signature and a stromal signature, which are then aggregated into a tumor purity score. As the method is rank-based, it is less sensitive to batch effects. A subtyping analysis of a breast cancer sample with potential low tumor cell content can be done by comparing the purity scores of the samples of interest to the purity scores of all breast cancer samples in a reference set. The 10 samples used for exemplifying the framework throughout this paper provide a very broad range of tumor purity (Fig. 2).

Purity reflects the degree to which an expression profile reflects patterns considered to represent normal cells which are also present in tumors including stromal and

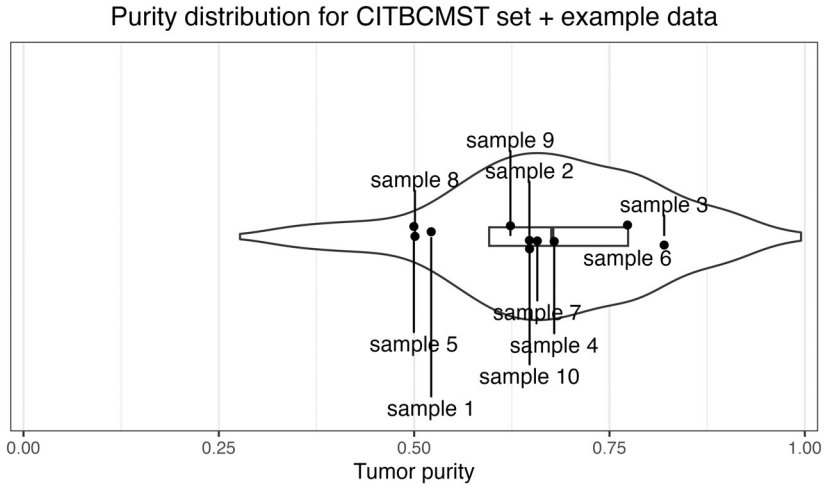

Fig. 2. Distribution of purity scores of the CIT reference samples (shown as violin with boxplot) and the 10 samples from our test set (shown as labeled points). The *x* axis represents purity in the range [0; 1]. A lower score indicates lower purity.

immune cells, and may need to be explicitly included in tumor expression profiling. For example, two tumor samples which are both characterized by high expression of a specific receptor, but which have different purity, may have very different measured expression values for the given receptor gene based on the tumor sample. Accordingly, if a sample is at the extremes of the purity range, one may consider normalizing expression by purity before reporting.

Generally, for subtyping purposes, it is important to remember that subtypes are defined on bulk samples which include the tumor microenvironment in addition to malignant cells. Accordingly, defined subtypes can actually be correlated with normal cell content, as is the case for both the CIT and PAM50 classifiers (Figs S1 and S2). In some subtyping schemes, the content of non-malignant cells is even the defining feature of specific categories [27].

Sample swaps or major contamination can be difficult to detect, but clues can be provided by comparing sample gene expression to expression characteristics of comprehensive tissue-specific expression profiles such as GTEx [28] or the TCGA [29]. As can be seen in Fig. 3, the 10 breast cancer test samples appear highly likely to be derived from breast tissue, which correlates with our expectations. Naturally, projection of swapped samples originating from the same tissue into the space of reference sets would not reveal the error. It does, however, indicate whether the analyzed sample actually has an expression profile resembling its true tissue of origin. Furthermore, it may be used to reveal if a sample has been severely impacted by any experimental processing step, for example, late freezing [30]. Validation of this procedure is shown for both GTEx and TCGA, where samples from each type of tissue can be shown corresponding to its tissue of origin (Fig. S3).

3.4. Harmonizing data

Before classification, harmonization of expression values from the reference samples is necessary. This is critical when considering data from the same or very similar platforms, for example, two microarray experiments. The need for harmonization is even greater when comparing samples across technology platforms. Even after mapping RNA sequencing reads to a reference consisting of probe sets, data from different experiments have completely different distributions, as illustrated in Fig. 4A. Harmonization can be carried out using batch correction methods, for example, ComBat [24] (Fig. 4B). While ComBat is designed to work with small sample sizes, at least eight samples must be submitted to adequately model technical variance [24]. If subtyping of fewer samples is needed, expression values can also be rank normalized (Fig. 4C).

When evaluating batch correction, it is important to consider that the method should provide complete and unbiased integration of the different data sets since even a small shift of the test set can change the distribution of assigned subtypes. ComBat and rank transformation solve the problem of this variation. ComBat explicitly models both technical and biological variance, the latter relying on biological cofactors – that is, sample condition, or in this case, the hitherto unknown subtype. If batch correcting two large test sets with equal distribution of biological conditions, this may not pose a problem, but if the test data are "biased" towards certain conditions, for example, includes only

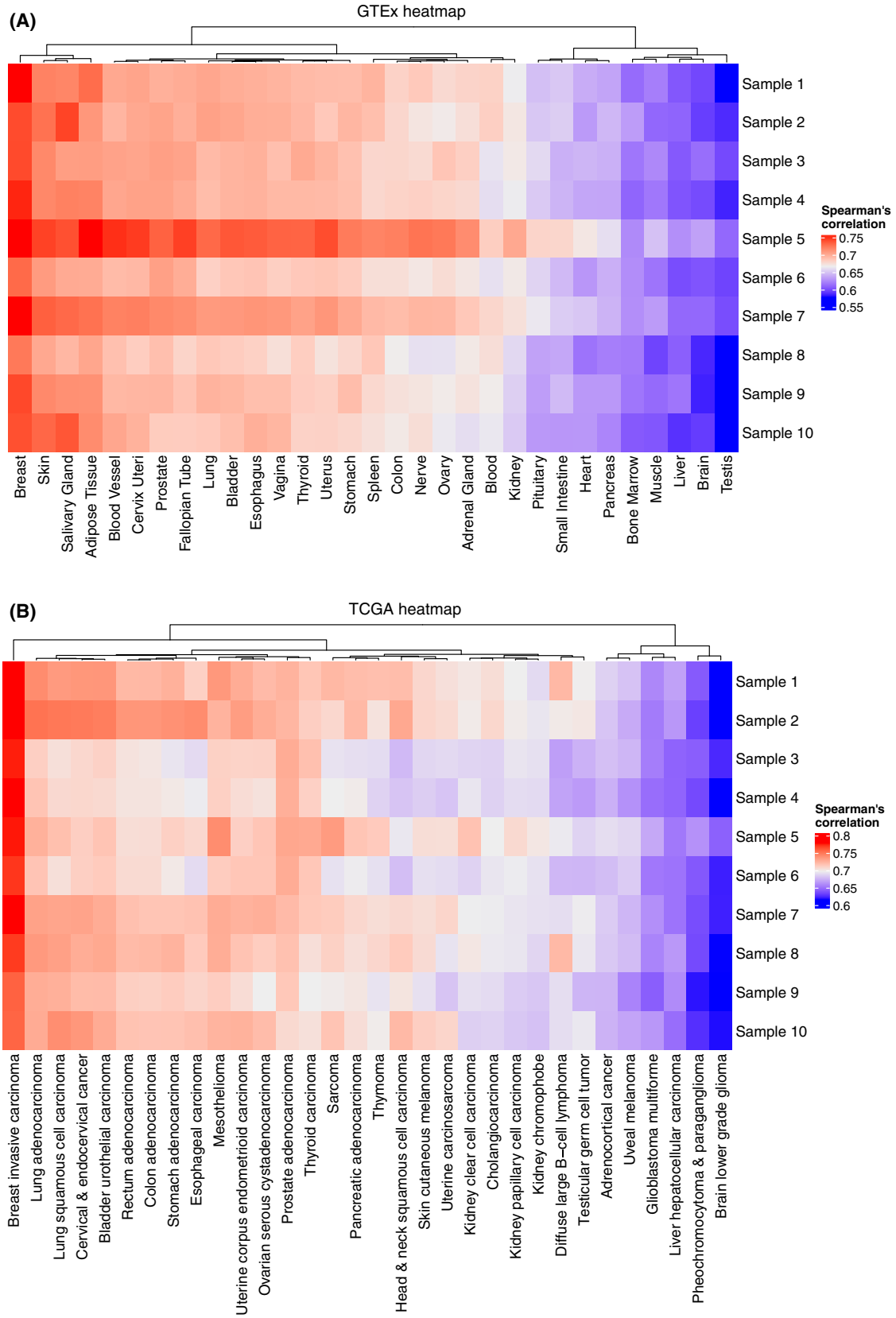

Fig. 3. Heatmaps of Spearman's correlations from samples to tissue mean-based centroids in data sets of comprehensive tissue-specific expression profiles. Columns are clustered and dendrograms are shown. (A) Heatmap of example samples versus centroids defined from a collection of healthy tissues from the GTEx project. (B) Heatmap of example samples versus centroids from a collection of cancer tissues from the TCGA project.

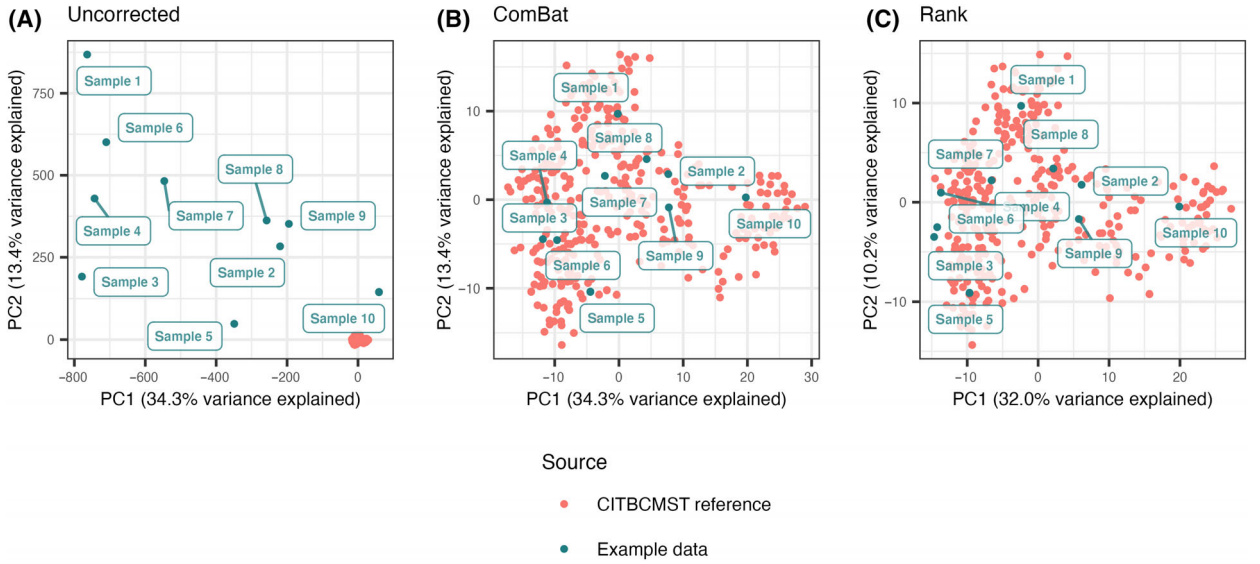

Fig. 4. Principal component analysis (PCA) plots of the CIT reference data (microarray) with the use case data (probe-mapped RNA-seq, transcripts per million (TPM)) projected in. Percent variance explained by the principal components (PCs) refer to the reference data set. (A) PCA of samples without batch correction. (B) PCA of samples batch corrected with ComBat. (C) PCA of samples integrated using rank.

one subtype, ComBat is not the optimal solution as illustrated in Fig. 5A and Fig. S4. Rather, correction methods for which the sample-wise transformation is independent of the other samples in the test set is desirable [16,31], and as can be seen in Fig. 5B, rank transformed data are readily comparable to the reference. In situations where the reference data are not provided as raw data, or only include a subset of features, it may be useful to use a signature enrichment approach such as biological process transformation.

3.5. Subtype classification

Once the reference and test data are harmonized, subtypes can be assigned to the test samples. Some subtyping studies provide a software package for classifying new samples, for example, the CIT classifier for breast cancer [10], while others do not. In the latter cases, an appropriate classifier must be trained and applied. There are no one-size-fits-all approaches for this, and hence, a classifier may be selected based on its cross-validation performance. To demonstrate how this is carried out, we have applied the CIT reference data and leave-one-out cross-validation to test two classification algorithms: a *k*-nearest neighbor classifier and a distance-to-centroid classifier. Advantages of these algorithms are (a) they are quite simple and thus the results are easily interpretable, and (b) subtype assignment probability can be inferred (fuzzy classification), rather than the methods providing a "winner-takes-all" classification. Distance-to-centroid additionally allows for easy outlier detection. Another strategy for classification is to define subtype specific gene expression signatures using ssGSEA based on the training set, and calculate the enrichment score of each signature for each additional sample. An advantage of this approach is that it is less sensitive to missing features resulting from data sparsity (as observed in single cell transcriptomics). As seen in Table 1, all three classifiers perform reasonably well, with a slight performance advantage for the distance-to-centroid classifier.

The performance of the leave-one-out classification of the reference data set reveals that if reference data and test data is appropriately harmonized, even simple classifiers can perform quite well. Of course, some methods will be preferable to others as illustrated above, but this can easily be established using cross-

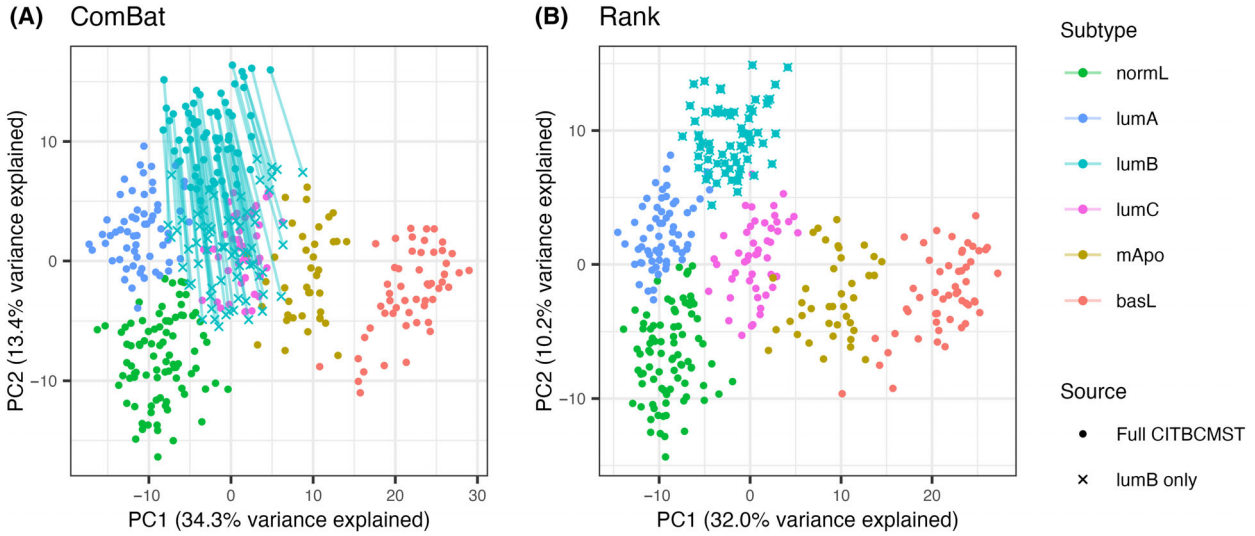

Fig. 5. Examples of harmonization of data sets with unbalanced biological conditions. Percent variance explained by the PCs refer to the full CIT reference data. (A) Batch correction of the entire CIT reference data against the lumB samples only using ComBat with the full set as reference. Without biological conditions as cofactors, ComBat under-estimates the biological variance causing the lumB samples to become centered in the PC space. Similar plots for the remaining subtypes are shown in Fig. S4. (B) A similar example using rank transformation.

Table 1. Precision, recall, and weighted accuracy from leave-one-out cross-validation of *k*-nearest neighbor, nearest centroid, and subtype signature enrichment for the CIT reference data. Highest precision per subtype, highest recall, and highest weighted accuracy are highlighted in bold text.

	k-nearest neighbor		Distance-to-centroid (Euclidean distance)		Subtype signature ssGSEA	
Subtype	Precision	Recall	Precision	Recall	Precision	Recall
normL	0.966	0.989	**1**	**1**	0.827	0.920
lumA	0.909	**0.984**	**1**	**0.984**	0.875	0.918
lumB	0.955	0.955	**0.957**	**1**	0.844	0.985
lumC	1	0.854	**1**	**0.958**	0.722	0.542
mApo	**1**	**1**	**1**	**1**	**1**	0.795
basL	**1**	**1**	**1**	**1**	**1**	0.925
Weighted accuracy	0.963		**0.990**		0.847	

validation. This conclusion is not only applicable for the specific case of the CIT tool for breast cancer but the approach may be used in much wider contexts across data sets, subtyping schemes, and cancer types.

As mentioned, many classification methods, such as those used here, have the added advantage of fuzzy classification. This means that samples with almost equal distances to two or more subtypes may be considered as mixed cases [10]. In clinical diagnostic settings, such information may be of interest, but if more simple classification is preferred, the single closest centroid can also be reported as winner-takes-all. An example of fuzzy classification is shown in Fig. 6, where the 10 test samples are classified according to the CIT scheme, using rank transformation followed by a distance-to-centroid classifier. While some samples clearly resemble a single subtype (e.g., sample 10), the label of others is more unclear (e.g., sample 6).

A similar approach to classification can be taken for PAM50. The primary difference between CIT and PAM50, is that the former study provided a training set, a feature set, and a classifier, while the latter provided a 50-gene signature. Using the TCGA BRCA data with the ranks of the 50 PAM50 genes, a nearest centroid classifier yields a weighted accuracy of 0.894 (See Table S2 for full performance metrics). As shown in Fig. 7, a majority of our included samples are labeled as LumA. While this fits nicely with the IHC-

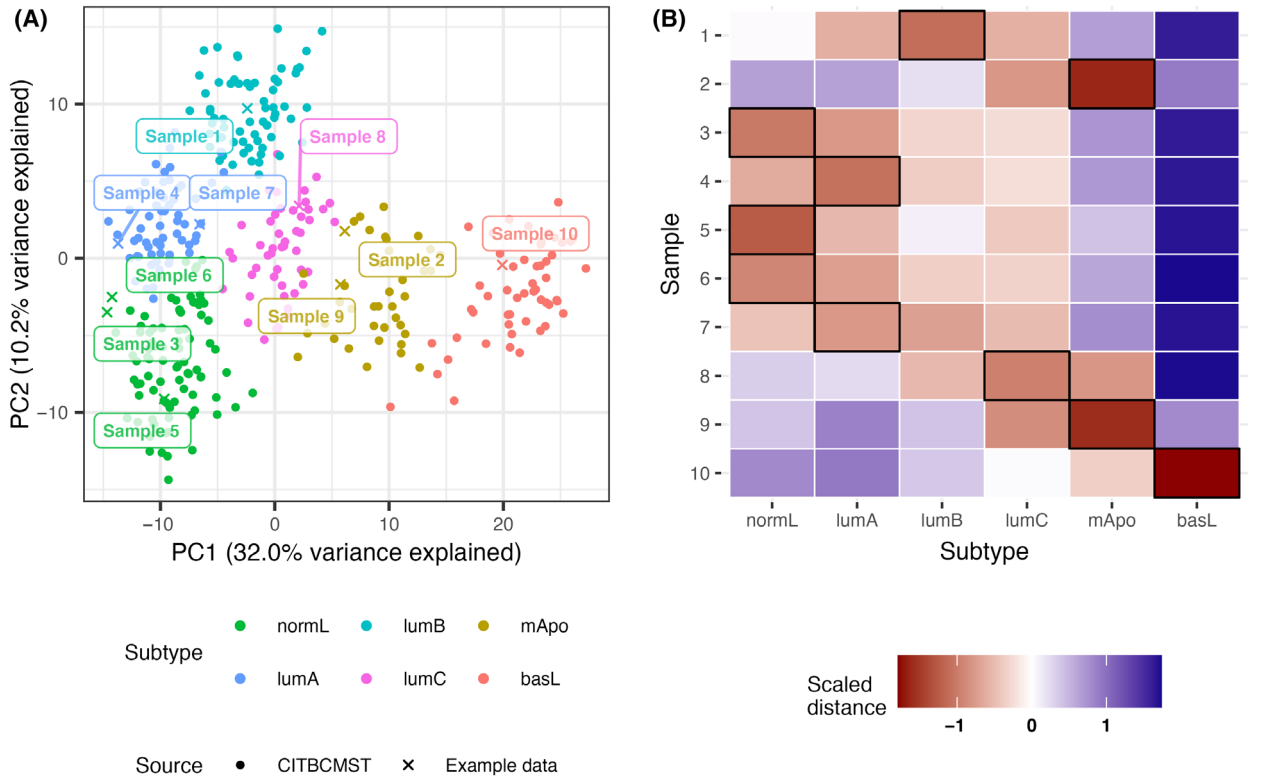

Fig. 6. Subtyping of 10 test samples using rank transformation and a distance-to-centroid classifier (Kendall Tau distance). (A) Projection of the samples from the use case data into CIT principal component space. Percent variance explained by the PCs refer to the CIT reference data. (B) Distance-to-centroid matrix, scaled per sample. Black squares indicate the closest centroid for each sample.

based ER-positive status of all of these six samples, it is somewhat conflicting that sample 1, which is also HER2-positive, is not classified as Her2, but this sample was similarly classified as luminal with CIT. In the case of sample 7, it is also quite close to the LumB subtype (Fig. 7B), so this could be considered as a mixed assignment. Sample 2, sample 8, and sample 10 behave as expected from their receptor status, but the assignment of sample 9 to the Normal-like subtype may seem unexpected. Of note, the PAM50 definition of Normal-like is actually based on non-cancerous breast tissue samples but the assignment of receptor-negative samples to the Normal-like PAM50 subtype is not unusual [32].

To further validate the results, classification using the CIT scheme was also done on the full use case data set. With this, 4/57 samples were classified differently between rank transformed microarray data and rank transformed RNA-seq data. The misclassifications were between lumA, lumB, lumC, and normL. These four misclassifications lie closely together in their PCA space and seem like edge cases, when looking at the centroids (Figs S5 and S6, Table S3).

3.6. Additional features

Besides the assignment of subtypes based on gene expression analysis, additional tumor features may be of interest when optimizing treatment regimens. In the case of breast cancer, this includes expression levels of specific receptor genes, such as the estrogen receptor (ER), progesterone receptor (PR), and the HER2-receptor [33]. While receptor status is mainly based on immunohistochemistry or ERBB2 FISH probes, multiple studies have shown that measurements at the mRNA-level have reasonably strong correlations with those from IHC [10,33–35]. Of note, a sample's expression level of any tumor-associated gene may be impacted by the sample purity, as also described in brief in Section 3.3.

Generally speaking, any feature may be visualized in relation to any reference (e.g., subtyping reference, TCGA, and in-house samples), as seen in Fig. 8. In addition to genes, enrichment of relevant gene signatures or pathways related to prognostics can also be visualized. More widely across cancer types, this can include detection of specific mutations [36,37], or

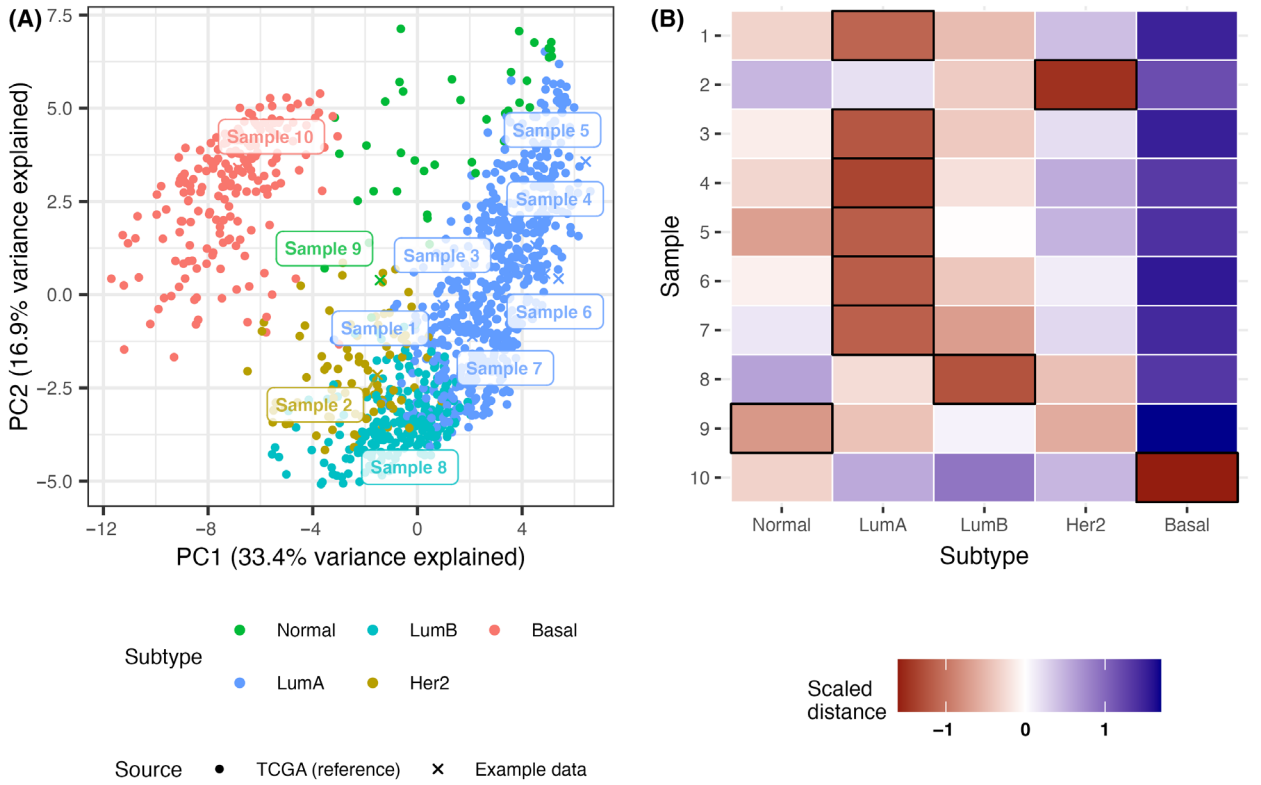

Fig. 7. Subtyping of 10 test samples using rank transformation and a distance-to-centroid classifier (Kendall Tau distance). (A) Projection of the samples from the use case data into PAM50 principal component space. Percent variance explained by the PCs refer to the TCGA reference set. (B) Distance-to-centroid matrix, scaled per sample. Black squares indicate the closest centroid for each sample.

methylation analyses [38]. Such analyses add to the complexity and costs of molecular profiling, but can also provide valuable information.

4. Discussion

Molecular characterization of tumors is an important tool for precision medicine. In this paper, we describe a bioinformatics framework based on years of practical experience. Most characteristics rely on comparisons to reference samples. For tumor purity, the sample can be compared to the previously collected and analyzed samples to establish whether the sample is within a reasonable range. For quality control, comparisons to different cancer samples in the TCGA data or the healthy tissues in GTEx may reveal more serious sample issues, if samples do not resemble the expected tissues.

For subtyping, samples are directly compared to the reference data sets. For these comparisons to yield sensible results, mapping, normalization, and harmonization must be considered, as transcriptomic measurements are sensitive to batch effects and technology incompatibility (e.g., comparing microarray data and RNA sequencing data). In the case of CIT (and many other subtyping methods), gene expression profiles were measured using the Affymetrix HG-U133 Plus 2.0 DNA microarray. In the original study, the subtyping itself was not based on summarized gene expression values, but a selection of probe set intensities. This means that while gene-level classification is possible using the classifier offered by the authors of CIT, additional samples should optimally be classified based on intensities of the same probe sets. Consequently, for comparing RNA sequencing data to microarray, mapping to probe set sequences greatly increases compatibility.

The batch correction tool of choice in the vast amount of transcriptomics applications is ComBat. However, ComBat is dependent on biological cofactors to accurately model biological variance. Therefore, using ComBat carries the risk of removing biological variance without knowing the subtypes in advance. Instead, it is recommended [31,39,40] to use the expression ranks of genes. One caveat is that gene expression ranks may be sensitive to noise – particularly in the low end of the expression spectrum, where

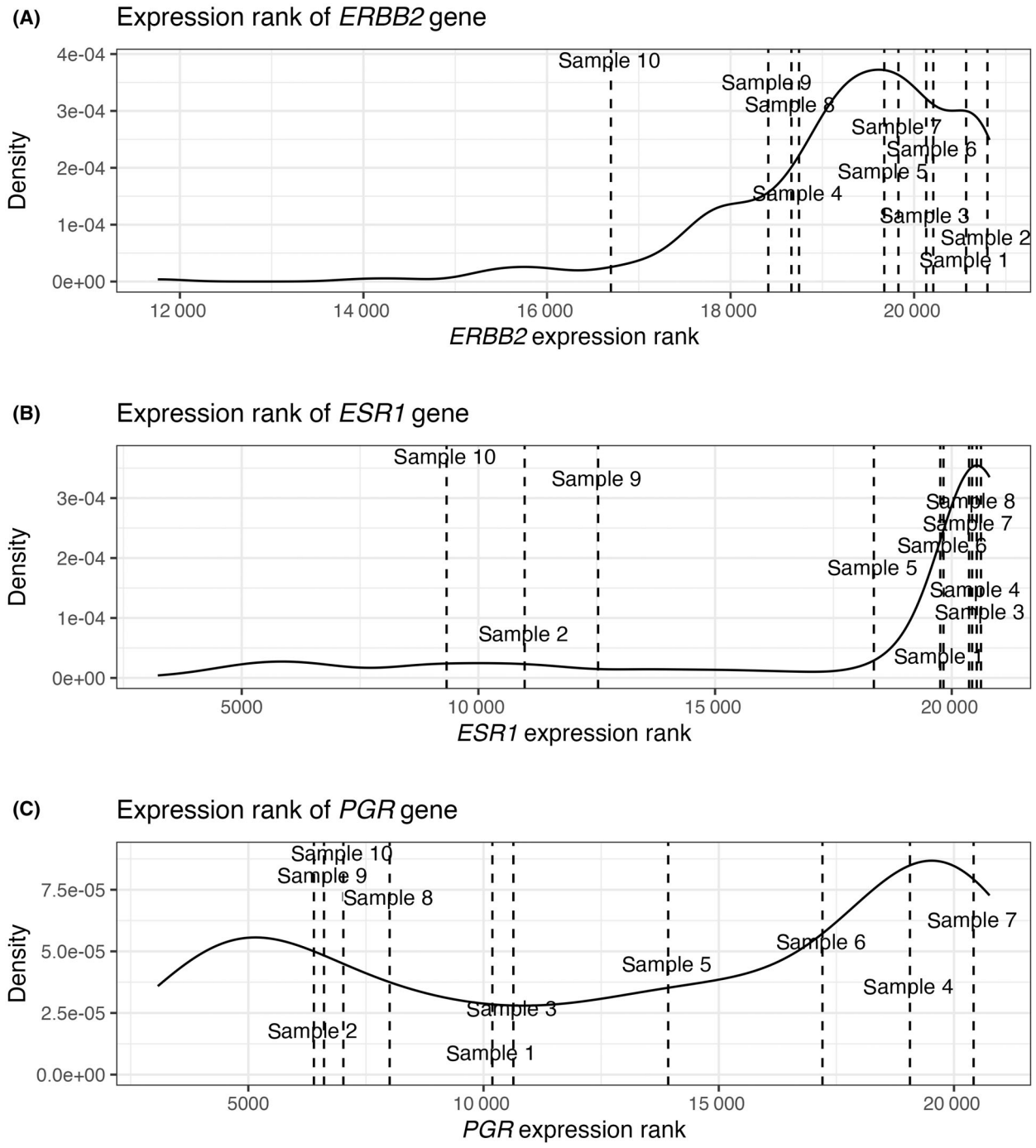

Fig. 8. Visualizing the expression ranks of specific genes of interest in relation to the expression rank distributions in the reference data (CIT). (A) Expression of *ERBB2* (HER2). (B) Expression of *ESR1* (Estrogen receptor). (C) Expression of *PGR* (Progesterone receptor).

even a small change can induce swaps in the ranks. One option is to apply the robust rank aggregation [41], although when working with small, curated gene sets such as PAM50, the effect is negligible.

Once data are harmonized, dimensionality reduction and classification can be performed on the ranks. For distance-based classifiers, a distance metric suitable for ranks should be selected. Some studies have presented advanced machine learning classifiers for breast cancer samples [42–44]. In the present paper, we intentionally focused on three very simple methods, and show that with proper data harmonization, even simple models

will work. Simple models come with the added bonus of interpretability and are widely useful, especially in resource constrained settings. Our classifiers were based on the core gene sets defined by the original studies. Others define their own gene sets to optimize cross-validation performance [16,43]. While neither approach is methodologically wrong, it does open the debate of whether the subtype calls or the genetic features on which they are based are of greater importance.

One of the classification methods we chose to apply, stands out from the rest: the subtype-specific single sample gene set enrichment analysis. Though it was the poorest performing of the three tested methods, we chose to highlight this method as it is more robust to data sparsity [45], which may prove important in single cell diagnostics [46]. One important thing to consider in this context is that signatures derived from bulk transcriptomics include signals from the entire micro-environment, and as we show here, tumor purity not only impacted the definitions of the subtypes, but also the subsequent molecular characterizations. This means that single tumor cells may not readily fit into current subtyping schemes.

For diagnostic applications, stability is essential. This means that the stability of the software packages used, R versioning, etc., must be taken into consideration. This is essential both from an operational point of view, as well as from an analytical point. This may be addressed by using a docker environment.

5. Conclusion

In this paper, we have presented a framework which enables robust molecular characterization of clinical cancer samples. Particularly, the work emphasizes that harmonization of the new data to be classified relative to the reference is essential for deriving a molecular subtype. The method for harmonization is important. Methods relying on intra-sample relative expression values are particularly suited for classification of single samples. We showed that once data are properly harmonized, even simple classification models can yield high accuracy.

Acknowledgements

This research was funded by the Independent Research Fund, Denmark, grant number 8048-00078A to LRO. BP was funded by a grant from the Otto Mønsted Foundation to the Technical University of Denmark.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

LRO, BC, and CBP contributed to the conceptualization. LRO, CBP, KVS, and FOB contributed to the methodology. LRO and CBP contributed to the software. LRO, CBP, LR, and HSW contributed to the formal analysis. CBP, LRO, MCR, and FOB contributed to the investigation. LRO contributed to the resources, project administration, and funding acquisition. LRO, CBP, BP, NMK, BC, MCR, and FOB contributed to the data curation. CBP, LRO, and MCR contributed to the writing—original draft preparation. CBP, BC, NMK, BP, KVS, MCR, FOB, LR, and HSW contributed to the writing review and editing. CBP and LRO contributed to the visualization. LRO and MCR contributed to the supervision.

Peer review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/1878-0261.13580.

Data accessibility

All data are available via public repositories. TCGA BRCA and GTEx data was downloaded from Xenabrowser.net (dataset ID: TcgaTargetGtex_rsem_gene_tpm). The CIT microarray (Affymetrix HG-U133 Plus 2.0) training data was downloaded from ArrayExpress (accession number: E-MTAB-365). The use case data set is available in EGA (accession number: EGAD00001000627).

References

1 Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, et al. TCGA research network the somatic genomic landscape of glioblastoma. *Cell*. 2013;**155**:462–77. https://doi.org/10.1016/j.cell.2013.09.034

2 Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;**490**:61–70. https://doi.org/10.1038/nature11412

3 Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;**487**:330–7. https://doi.org/10.1038/nature11252

4 Campbell JD, Yau C, Bowlby R, Liu Y, Brennan K, Fan H, et al. Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. *Cell Rep*. 2018;**23**:194–212.e6. https://doi.org/10.1016/j.celrep.2018.03.063

5 Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, et al. Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast*. 2015;**24** (Suppl 2):S26–35. https://doi.org/10.1016/j.breast.2015.07.008

6 Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res*. 2008;**14**:5198–208. https://doi.org/10.1158/1078-0432.CCR-08-0196

7 Wilkerson MD, Yin X, Hoadley KA, Liu Y, Hayward MC, Cabanski CR, et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin Cancer Res*. 2010;**16**:4864–75. https://doi.org/10.1158/1078-0432.CCR-10-0199

8 Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. *Nature*. 2000;**406**:747–52. https://doi.org/10.1038/35021093

9 Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;**27**:1160–7. https://doi.org/10.1200/JCO.2008.18.1370

10 Guedj M, Marisa L, de Reynies A, Orsetti B, Schiappa R, Bibeau F, et al. A refined molecular taxonomy of breast cancer. *Oncogene*. 2012;**31**:1196–206. https://doi.org/10.1038/onc.2011.301

11 Rossing M, Østrup O, Majewski WW, Kinalis S, Jensen M-B, Knoop A, et al. Molecular subtyping of breast cancer improves identification of both high and low risk patients. *Acta Oncol*. 2018;**57**:58–66. https://doi.org/10.1080/0284186X.2017.1398416

12 Tofigh A, Suderman M, Paquet ER, Livingstone J, Bertos N, Saleh SM, et al. The prognostic ease and difficulty of invasive breast carcinoma. *Cell Rep*. 2014;**9**:129–42. https://doi.org/10.1016/j.celrep.2014.08.073

13 Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based prosigna breast cancer gene signature assay. *BMC Med Genomics*. 2015;**8**:54. https://doi.org/10.1186/s12920-015-0129-6

14 Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res*. 2010;**12**:R68. https://doi.org/10.1186/bcr2635

15 Fougner C, Bergholtz H, Norum JH, Sørlie T. Re-definition of claudin-low as a breast cancer phenotype. *Nat Commun*. 2020;**11**:1787. https://doi.org/10.1038/s41467-020-15574-5

16 Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. *J Natl Cancer Inst*. 2015;**107**:357. https://doi.org/10.1093/jnci/dju357

17 Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol*. 2020;**38**:675–8. https://doi.org/10.1038/s41587-020-0546-8

18 Gautier L, Cope L, Bolstad BM, Irizarry RA. affy – analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004;**20**:307–15. https://doi.org/10.1093/bioinformatics/btg405

19 Fumagalli D, Blanchet-Cohen A, Brown D, Desmedt C, Gacquer D, Michiels S, et al. Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-sequencing technology. *BMC Genomics*. 2014;**15**:1008. https://doi.org/10.1186/1471-2164-15-1008

20 Pedersen CB, Nielsen FC, Rossing M, Olsen LR. Using microarray-based subtyping methods for breast cancer in the era of high-throughput RNA sequencing. *Mol Oncol*. 2018;**12**:2136–46. https://doi.org/10.1002/1878-0261.12389

21 Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics – a bioconductor package for quality assessment of microarray data. *Bioinformatics*. 2009;**25**:415–6. https://doi.org/10.1093/bioinformatics/btn647

22 Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010 [cited 2010 April 26]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

23 Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun*. 2013;**4**:2612. https://doi.org/10.1038/ncomms3612

24 Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;**8**:118–27. https://doi.org/10.1093/biostatistics/kxj037

25 Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. Single sample scoring of molecular phenotypes. *BMC Bioinformatics*. 2018;**19**:404. https://doi.org/10.1186/s12859-018-2435-4

26 Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*. 2009;**462**:108–12. https://doi.org/10.1038/nature08460

27 Lehmann BD, Jovanović B, Chen X, Estrada MV, Johnson KN, Shyr Y, et al. Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection. *PLoS One*. 2016;**11**:e0157368. https://doi.org/10.1371/journal.pone.0157368

28 GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nat Genet*. 2013;**45**:580–5. https://doi.org/10.1038/ng.2653

29 Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;**45**:1113–20. https://doi.org/10.1038/ng.2764
30 De Cecco L, Musella V, Veneroni S, Cappelletti V, Bongarzone I, Callari M, et al. Impact of biospecimens handling on biomarker research in breast cancer. *BMC Cancer*. 2009;**9**:409. https://doi.org/10.1186/1471-2407-9-409
31 Patil P, Bachant-Winner P-O, Haibe-Kains B, Leek JT. Test set bias affects reproducibility of gene signatures. *Bioinformatics*. 2015;**31**:2318–23. https://doi.org/10.1093/bioinformatics/btv157
32 Lehmann BD, Pietenpol JA. Identification and use of biomarkers in treatment strategies for triple-negative breast cancer subtypes. *J Pathol*. 2014;**232**:142–50. https://doi.org/10.1002/path.4280
33 Roepman P, Horlings HM, Krijgsman O, Kok M, Bueno-de-Mesquita JM, Bender R, et al. Microarray-based determination of estrogen receptor, progesterone receptor, and HER2 receptor status in breast cancer. *Clin Cancer Res*. 2009;**15**:7003–11. https://doi.org/10.1158/1078-0432.CCR-09-0449
34 Brueffer C, Vallon-Christersson J, Grabau D, Ehinger A, Häkkinen J, Hegardt C, et al. Clinical value of RNA sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter Sweden Cancerome Analysis Network-Breast Initiative. *JCO Precis Oncol*. 2018;**2**:PO.17.00135. https://doi.org/10.1200/PO.17.00135
35 Sorokin M, Ignatev K, Poddubskaya E, Vladimirova U, Gaifullin N, Lantsov D, et al. RNA sequencing in comparison to immunohistochemistry for measuring cancer biomarkers in breast cancer and lung cancer specimens. *Biomedicine*. 2020;**8**:114. https://doi.org/10.3390/biomedicines8050114
36 Dang L, Yen K, Attar EC. IDH mutations in cancer and progress toward development of targeted therapeutics. *Ann Oncol*. 2016;**27**:599–608. https://doi.org/10.1093/annonc/mdw013
37 Lee JM, Ledermann JA, Kohn EC. PARP inhibitors for BRCA1/2 mutation-associated and BRCA-like malignancies. *Ann Oncol*. 2014;**25**:32–40. https://doi.org/10.1093/annonc/mdt384
38 Locke WJ, Guanzon D, Ma C, Liew YJ, Duesing KR, Fung KYC, et al. DNA methylation cancer biomarkers: translation to the clinic. *Front Genet*. 2019;**10**:1150. https://doi.org/10.3389/fgene.2019.01150
39 Tang K, Ji X, Zhou M, Deng Z, Huang Y, Zheng G, et al. Rank-in: enabling integrative analysis across microarray and RNA-seq for cancer. *Nucleic Acids Res*. 2021;**49**:e99. https://doi.org/10.1093/nar/gkab554
40 Lyu Y, Li Q. A semi-parametric statistical model for integrating gene expression profiles across different platforms. *BMC Bioinformatics*. 2016;**17**(Suppl 1):5. https://doi.org/10.1186/s12859-015-0847-y
41 Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*. 2012;**28**:573–80. https://doi.org/10.1093/bioinformatics/btr709
42 Gao F, Wang W, Tan M, Zhu L, Zhang Y, Fessler E, et al. DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis*. 2019;**8**:44. https://doi.org/10.1038/s41389-019-0157-8
43 Seo M-K, Paik S, Kim S. An improved, assay platform agnostic, absolute single sample breast cancer subtype classifier. *Cancers (Basel)*. 2020;**12**:3506. https://doi.org/10.3390/cancers12123506
44 Rhee S, Seo S, Kim S. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence main track. 2018. p. 3527–34.
45 Wang Q, Hu B, Hu X, Kim H, Squatrito M, Scarpace L, et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell*. 2017;**32**:42–56.e6. https://doi.org/10.1016/j.ccell.2017.06.003
46 Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet*. 2021;**53**:1334–47. https://doi.org/10.1038/s41588-021-00911-1

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.
Fig. S1. Purity estimates for CIT reference data.
Fig. S2. Purity estimates for TCGA data.
Fig. S3. Heatmaps of sample correlation to tissues.
Fig. S4. Combat batch correction on five subtypes.
Fig. S5. PCA of the four differently classified use case samples projected into CIT reference space.
Fig. S6. Heatmaps of distances to different centroids.
Table S1. Example patient characteristics.
Table S2. Precision, recall, and weighted accuracy from different classification methods.
Table S3. Distances to centroids for the four differently classified use case samples.