AN R PACKAGE FOR DIVERGENCE ANALYSIS OF OMICS DATA

A PREPRINT

Wikum Dinalankara
Department of Cancer Biology
Johns Hopkins School of Medicine
wdinala1@jhmi.edu

Qian Ke
Department of Applied Mathematics and Statistics
Johns Hopkins University
qke1@jhu.edu

Donald Geman
Department of Applied Mathematics and Statistics
Johns Hopkins University
geman@jhu.edu

Luigi Marchionni[*]
Department of Cancer Biology
Johns Hopkins School of Medicine
marchion@jhu.edu

September 23, 2019

ABSTRACT

Given the ever-increasing amount of high-dimensional and complex omics data becoming available, it is increasingly important to discover simple but effective methods of analysis. Divergence analysis transforms each entry of a high-dimensional omics profile into a digitized (binary or ternary) code based on the deviation of the entry from a given baseline population. The divergence package, available on the R platform through the bioconductor repository collection, provides easy-to-use functions for carrying out this transformation.

Keywords divergence · omics · cancer

1 Introduction

The technologies that provide us with high-dimensional omics data continue to advance at a rapid rate. Particularly in the last decade, the available modalities of omics data have considerably expanded and now include, among others, coding and noncoding RNA expression, micro RNA expression, protein expression, epigenetic profiling related to histones and CpG methylation, copy-number and mutation profiling. As a result, the analysis of multi-modal omics data has become indispensable in many domains of biological and medical research.

Wheras the quantity and quality of available data has appreciably increased, the results of research based on such data are often not reliable, robust and replicable. A key challenge has been to quantify the level of variability and diversity in omics profiles in a given population, and to separate normal and technical variability from abnormal variability indicative of a biological property such as disease.

Recently we have introduced divergence analysis as a method for simplyfing high-dimensional omics data for bioinformatic analyses [1]. This method conceptually parallels the widespread use of deviation from normality as a disease marker in clinical testing, such as a blood based prostate specific antigen (PSA) test [2, 3]. Given a high-dimensional omics data profile, it can be converted to a binary or ternary string of the same length, where each value now indicates how the original value diverges from a baseline or reference population. The features of interest may be univariate such as a gene, a CpG site, or a protein, in which case the original profile is converted to one of three labels 0, -1 and 1 indicating whether the level of a feature is, respectively, within the reference range for that feature, or below or above that range. Or, the features of interest may be multivariate, such as sets of genes representing pathways of interest. In this case, divergence coding is binary, where a 0 or 1 for a multivariate feature indicates, respectively, that the set of feature is inside or outside the support of the multivariate, baseline distribution.

[*] corresponding author

We have prepared the 'divergence' R package[5] offered through the bioconductor package repository[4] which provides functionality for divergence based computations. Here we present how to use this package and provide a few short examples as to what types of analyses can be performed following the transformation of data to digitized divergence coding. In the following we use a set of breast normal and tumor samples from the TCGA project [6] for which RNA-seq expression profiles are provided to illustrate the workflow. A subet of this data is available with the R package.

2 Results

We may summarize the divergence method as follows. Before computing the baselin range (univariate) or support (multivariate) and the resulting divergence coding, a rank-transformation is applied to all data. Consider an omics sample respresented by a vector $X = (X_j), j = 1..m$, m being the dimensionality of the omics profiles. The following transformation is applied which converts the data to a normalized rank, with the minimum being zero.

$$Q_j(X) = \frac{|i \in 1..m : 0 < X_i \leq X_j|}{|i \in 1..m : 0 < X_i|}$$

Then $Q_j(X) \in [0, 1]$. The zero minimum is particularly useful in preserving the zero valued mass usually observed in many omics data modalities, such as RNA-Seq.

Now consider a multivariate feature indexed by S - i.e. a subset of the given m features (the univariate scenario is a special case when $|S| = 1$). We will denote the corresponding subset of a given omics profile X following the quantile transformation as $Q(X)^S$. Suppose we have n such profiles that constitute the baseline group. We esimtate the baseline support as follows: given a parameter $\gamma \in [0, 1]$, we compute l which is the floor of $n\gamma$. Then from each baseline sample k, if r_k is the distance from $Q(X)^S$ to it's l^{th} nearest neighbor in the multivariate feature space. If we denote the sphere around $Q(X)^S$ of radius r_k as U_k^S then the support of the baseline is the union of the regions covered by these speheres around each baseline sample, which we denote as $\hat{U}^S$:

$$\hat{U}^S = \bigcup_{k=1}^{n} U_k^S$$

Following the baseline estimation in this manner, for any given omics profile X, the divergence coding $Z(X)^S$ can be computed as:

$$Z(X)^S = \begin{cases} 1 & \text{if } Q(X)^S \notin \hat{U}^S \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

For the univariate scenario, the support is the union of a series of intervals. However, we apply a further simplification by replacing these with a single interval spanning the lowest end to the highest end of these intervals. Accordingly the divergence coding becomes ternary: $Z(X)^j \in \{-1, 0, 1\}$ with -1 indicating a value below the baseline interval for feature j, 1 indicating a value above the baseline interval, and 0 indicating no divergence.

In practice we use two more parameters, α and β. The β parameter is used for allowing a certain number of outliers to be exempted from the baseline. Given the radii of the spheres $r_1..r_n$, let $\bar{r}$ be the $(1-\beta)^{th}$ percentile of these values. Then we select only the spheres with $r_k \leq \bar{r}$ to compose the baseline. Once the baseline is computed, we can compute the divergence coding for the baseline samples; then α is the average proportion of divergent features (multivariate or univariate) among the baseline cohort. As discussed in the following section, we specify α and β and then select the γ value that fits these specified parameters. However the functionality for estimating a baseline for a specified γ parameter is available in the package as well.

For a more detailed description of the method, see [1].

2.1 Univariate Workflow

To carry out an analysis based on divergence, the first step is to determine the case and control samples. The control cohort here will be used for computing the baseline interval for each feature, and we will refer to it as the baseline cohort or baseline group. Once the baseline is computed, the divergence values for the case cohort will be computed with reference to the baseline.

What data should be used as the baseline cohort depends on the problem at hand and needs careful investigation based on the biologocial or clinical questions of interest that are being investigated. In many scenarios of disease based data, and in partciular cancer, normal samples would be a good choice. The more normal samples available, the more robust the baseline will be.

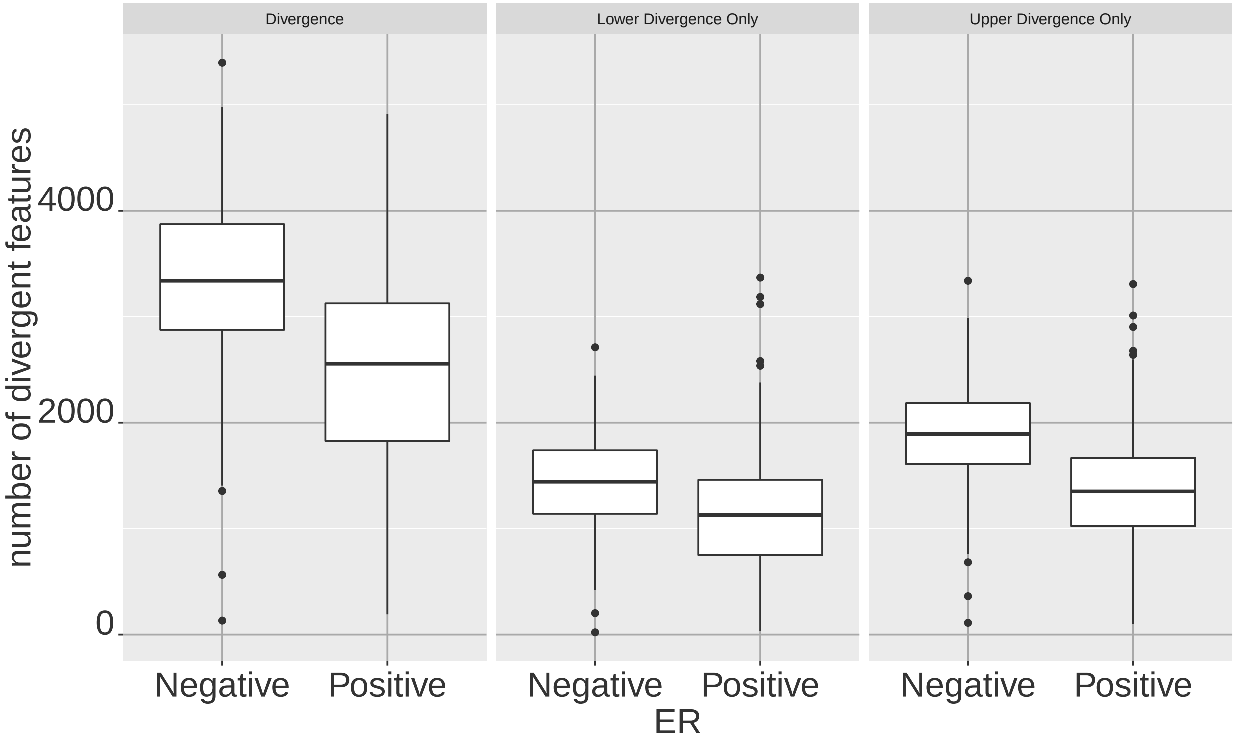

Figure 1: **Number of divergent features as a measure of sample divergence.** The size of the divergent set in each sample - i.e. the set of features that are divergent in a given sample - can be uses as a measurement of the divergence of a sample. This allows the comparison of sample level divergences between sample groups. When digitized with respect to a normal breast basenline, the boxplots compare the number of divergent features per sample between ER+ and ER- breast tumors. Given the higher risk associated with ER- breast tumors, we expect to observe higher divergence from normality compared to ER+. The same trend is observed if the consideration is limited to the number of upper or lower divergent features only.

Table 1: 10 Most differentially divergent genes between ER+ and ER- breast tumor samples.

gene	divergent probability, ER+	divergent probability, ER-	χ^2 test statistic	χ^2 test p-value
ESR1	0.0184	0.7416	472.4381	$P \ll 10^{-6}$
ACADSB	0.0686	0.7472	359.2727	$P \ll 10^{-6}$
TBC1D9	0.0251	0.6236	356.8933	$P \ll 10^{-6}$
PSAT1	0.6538	0.6685	344.7521	$P \ll 10^{-6}$
RHOB	0.0635	0.7079	337.8027	$P \ll 10^{-6}$
RABEP1	0.0435	0.6461	330.9501	$P \ll 10^{-6}$
SCUBE2	0.0886	0.7416	318.5029	$P \ll 10^{-6}$
C6orf97	0.1355	0.5393	310.8407	$P \ll 10^{-6}$
A2ML1	0.0702	0.6685	294.9983	$P \ll 10^{-6}$
SKP1	0.0552	0.6292	293.7114	$P \ll 10^{-6}$

All samples, both in the case cohort and the baseline cohort should have the same featues - for example genes or microarray probes, and be from the same platform (e.g. RNASeq or a specific microarray platform). Ideally the baseline cohort should be from the same experiment to avoid study-specific batch effects as much as possible. In general we suggest at least 20 samples be available to compute the baseline.

The divergence algorithm applies a scaled rank transformation to all samples before the digitization of the data. The package provides functionality for the user to compute this rank transformation manually, or it can be set to compute internally when the divergence computation is requested by the user. No other normalization procedures are applied, and the user may apply any normalization procedures beforehand as necessary.

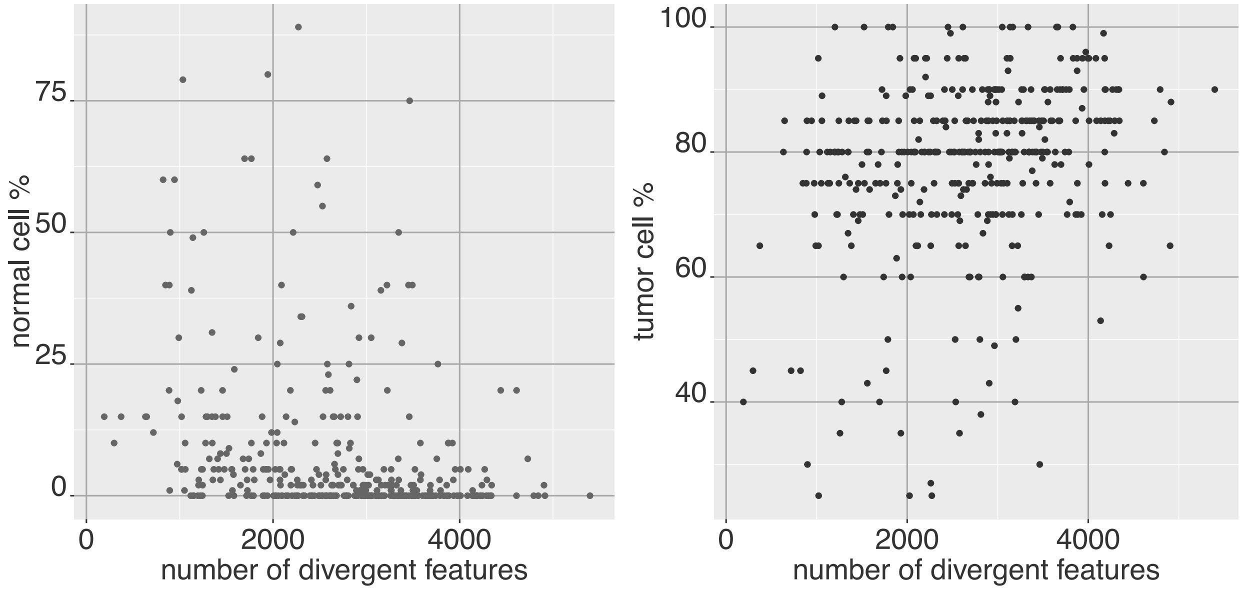

Figure 2: **Comparing sample divergence with clinical covariates.** Sample divergence of breast tumor samples (as measured by the size of the divergent set in each sample), with respect to a normal breast baseline, shows negative correlation (spearman correlation = -0.398) with the normal cell percentage estimate of the sample, and positive correlation (spearman correlation = 0.278) with the tumor cell percentage estimate of the sample, illustrating that sample divergence is indicative of the sample level deviation from normality.

There are three parameters involved in the divergence computation: α, β and γ. All paramters are in the $(0, 1)$ range. The β parameter adjusts the support to account for a certain percentage of outliers included in the baseline data, and the γ parameter provides a way to widen or tighten the support around each baseline sample. The closer γ is to 1, the further the support around each normal sample will extend. For a given support, the α value is simply the average number of divergent features per sample for the baseline cohort.

In usage, we usually provide the α and β values and a range of possible γ values and let the package find the most appropriate γ value out of these for the given α and β values. By default, the package uses $\alpha = 0.01$, $\beta = 0.95$, and $\gamma \in \{0.01, 0.02, ..., 0.09, 0.1, 0.2, ..., 0.9\}$ as a list of candidate γ values. Thus if you use these default values, the package will consider 95% of the baseline cohort to be included in the support and find the smallest possible γ value from the given list that will provide the average number of divergent features per sample in the baseline cohort to be 1% or less. For more details, see [1].

We use RNA-Seq data spanning 20531 genes from the TCGA Breast Cancer dataset [6] for the following analysis. This data cosists of 776 tumor samples and 111 normal samples, the latter which we will use as our baseline cohort.

Using the default parameters available, we compute the digitized ternary divergence format for the tumor samples with respect to the normal baseline of 111 samples. The algorithm selects a γ value of 0.2, which yields an α value of 0.0091 which is below the α threshold of 1% which was specified as the default.

The tumor data contains 598 ER+ and 178 ER- samples. We can see whether the number of divergent genes per sample are similar between the two ER groups or not (Figure 1).

As another example of using the size of the divergent feature set as an indicator of the divergence of the sample, we can compare it against cilinically obtained covariates such as the percentage of normal and tumor cells found in the sample. We see that these percentages track with sample level divergence (Figure 2).

To perform a differential expression analysis at the feature level for digitized divergence data, the package provides a χ^2 test functionality, which we can use to perform a χ^2 test for each gene between ER+ and ER- samples. Even simpler, we can merely compute the divergent probability of each gene over the ER+ and ER- samples respectively. Note that the divergent probability of a feature over a group of samples is simply the number of samples for which the feature is non-zero in the divergence space divided by the number of samples.

Figure 3 shows the divergent probability of each gene for ER+ and ER- samples. The samples in blue are genes that are significant under a bonferroni-adjusted $P \leq 0.05$ threshold from the χ^2 test. The ESR1 gene, which can be considered

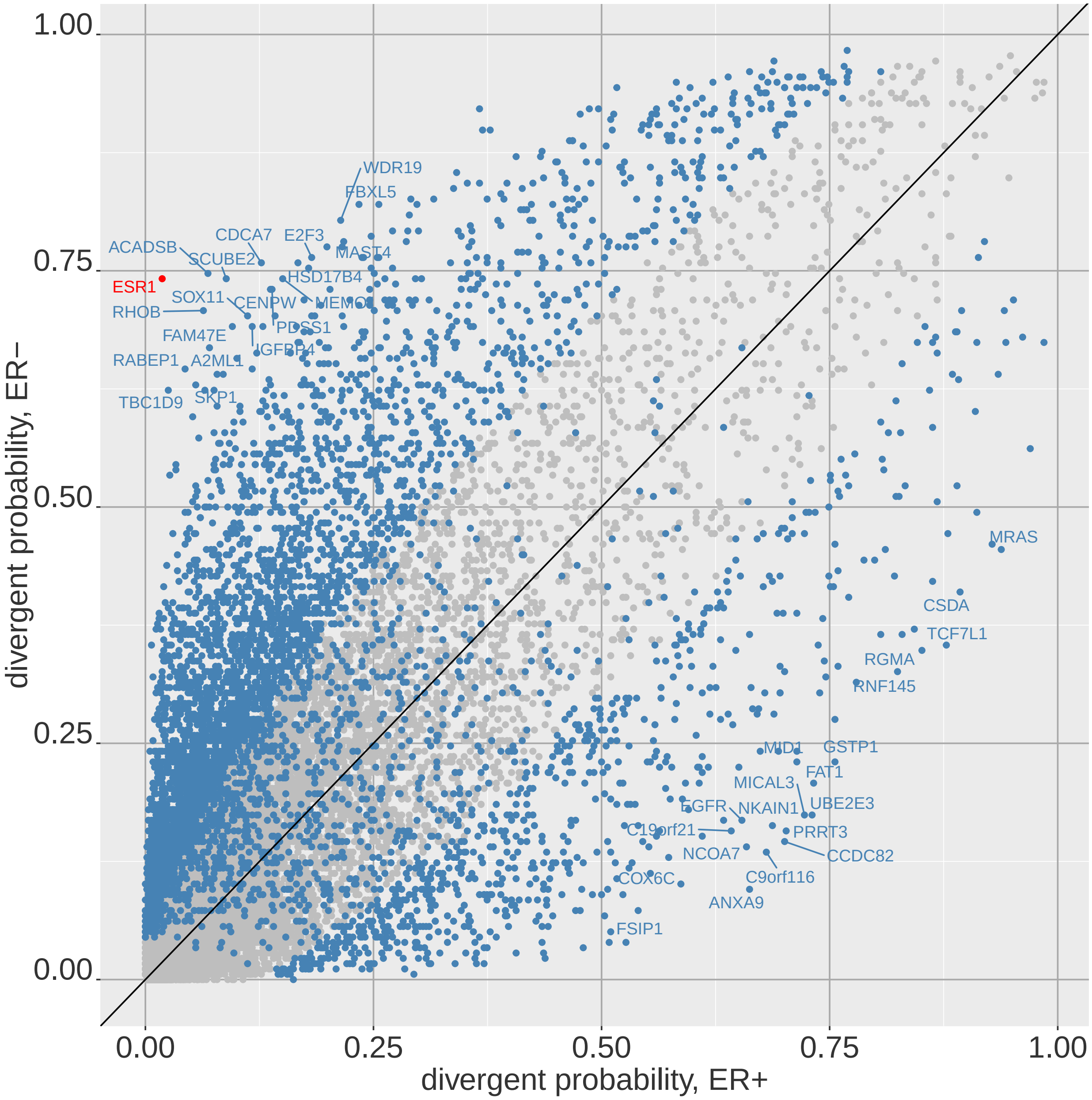

Figure 3: **Differentially divergent features between two phenotypes.** The probability of divergence for each feature is plotted for two sub-groups of breast tumors, ER+ and ER-. Genes far from the diagonal are likely to have highly different divergence patterns between the two groups (e.g. ESR1). Alternatively, a χ^2 test can be performed for each feature between the two groups, and points in blue indicate genes that are χ^2 test p-value significiant at a bonferroni adjusted $P \leq 0.05$ threshold.

a marker for ER status, is highly significant and is colored in red. Table 1 shows the top ten genes by rank of χ^2 test p-value.

We can observe how these genes track with the ER level by comparing the expression level to that of ESR1 expression, both in the regular expression value space and in the divergence space. Figures 4 and 5 show the values of ESR1 expression among ER+ and ER- breast tumor samples against two of the genes highly differentially expressed between the two sample groups, ACADSB and PSAT1, in the regular expression space ($\log_2$ transcripts per million) and in the divergence space.

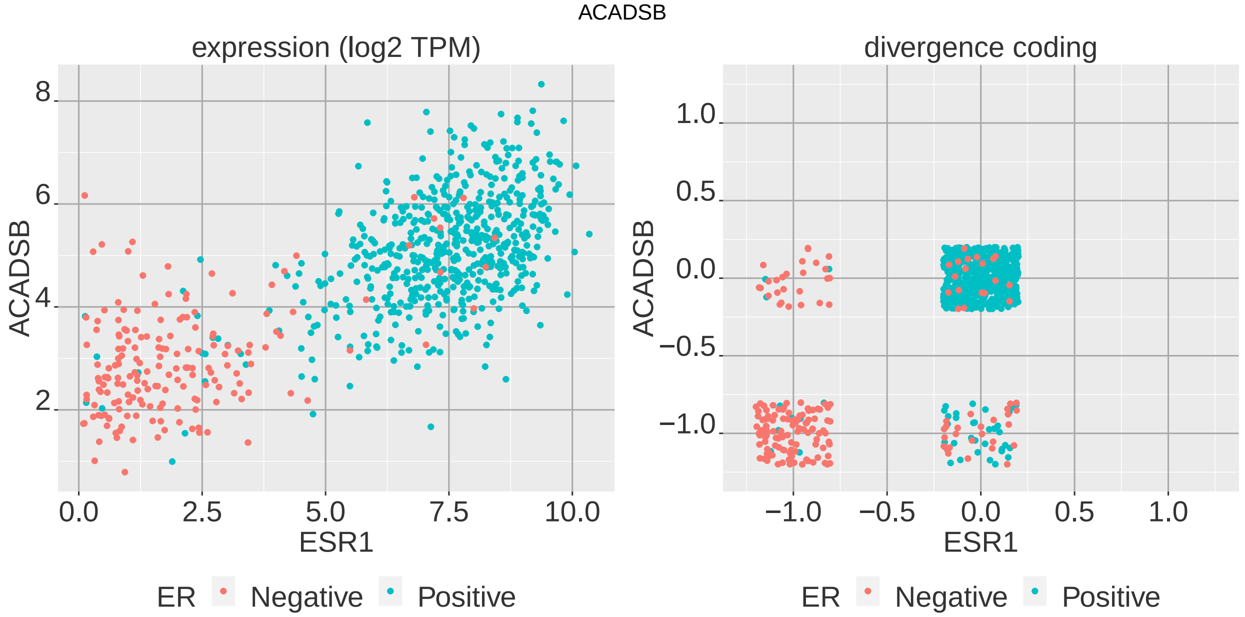

Figure 4: **Comparing two genes in expression space and divergence space.** Comparison of ACADSB and ESR1 gene expression for ER+ and ER- breast tumor samples in regular expression space ($\log_2$ TPM) and the digitized divergence space. ESR1 is an indicator of ER status and ACADSB shows to be highly differentiated between the two groups as well.

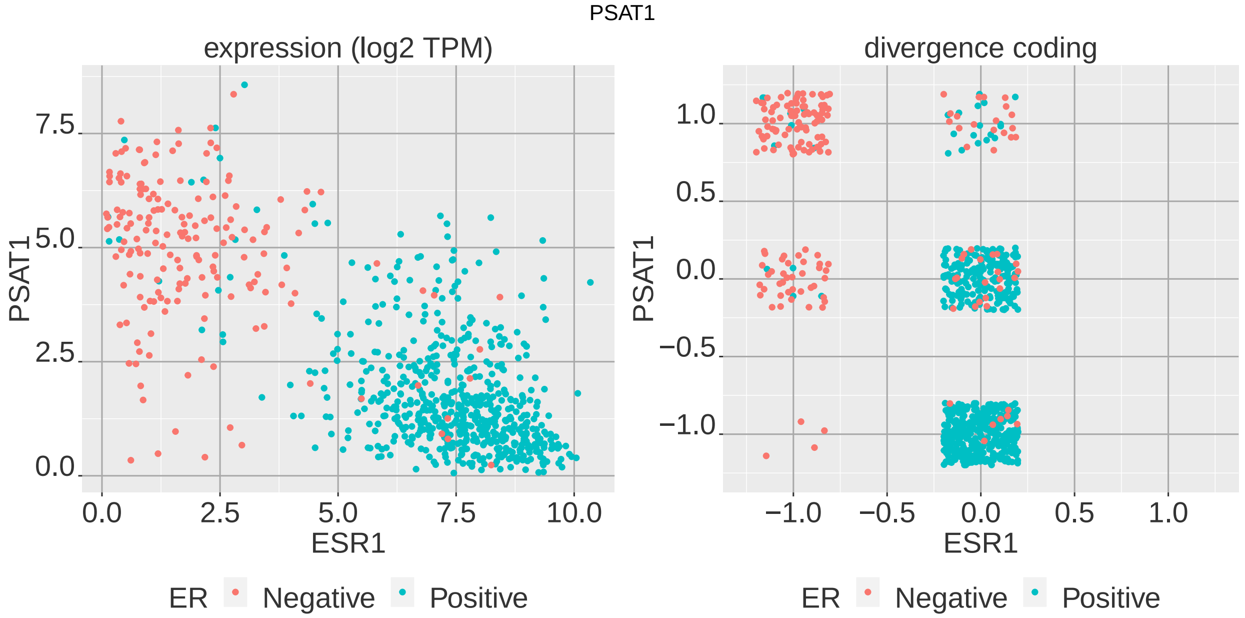

Figure 5: **Comparing two genes in expression space and divergence space.** Comparison of PSAT1 and ESR1 gene expression for ER+ and ER- breast tumor samples in regular expression space ($\log_2$ TPM) and the digitized divergence space.

We can also examine which genes are highly divergent for at least one out of multiple phenotypes. For example, figure 6 shows 50 genes that are highly significantly differentially divergent from a χ^2 test performed among the PAM50 subtypes of breast cancer. For each of these genes, we compute the average divergence (which will be in the interval $[0, 1]$) for each subype. The genes in the plot are color-coded by which subtype contains the highest or lowest average.

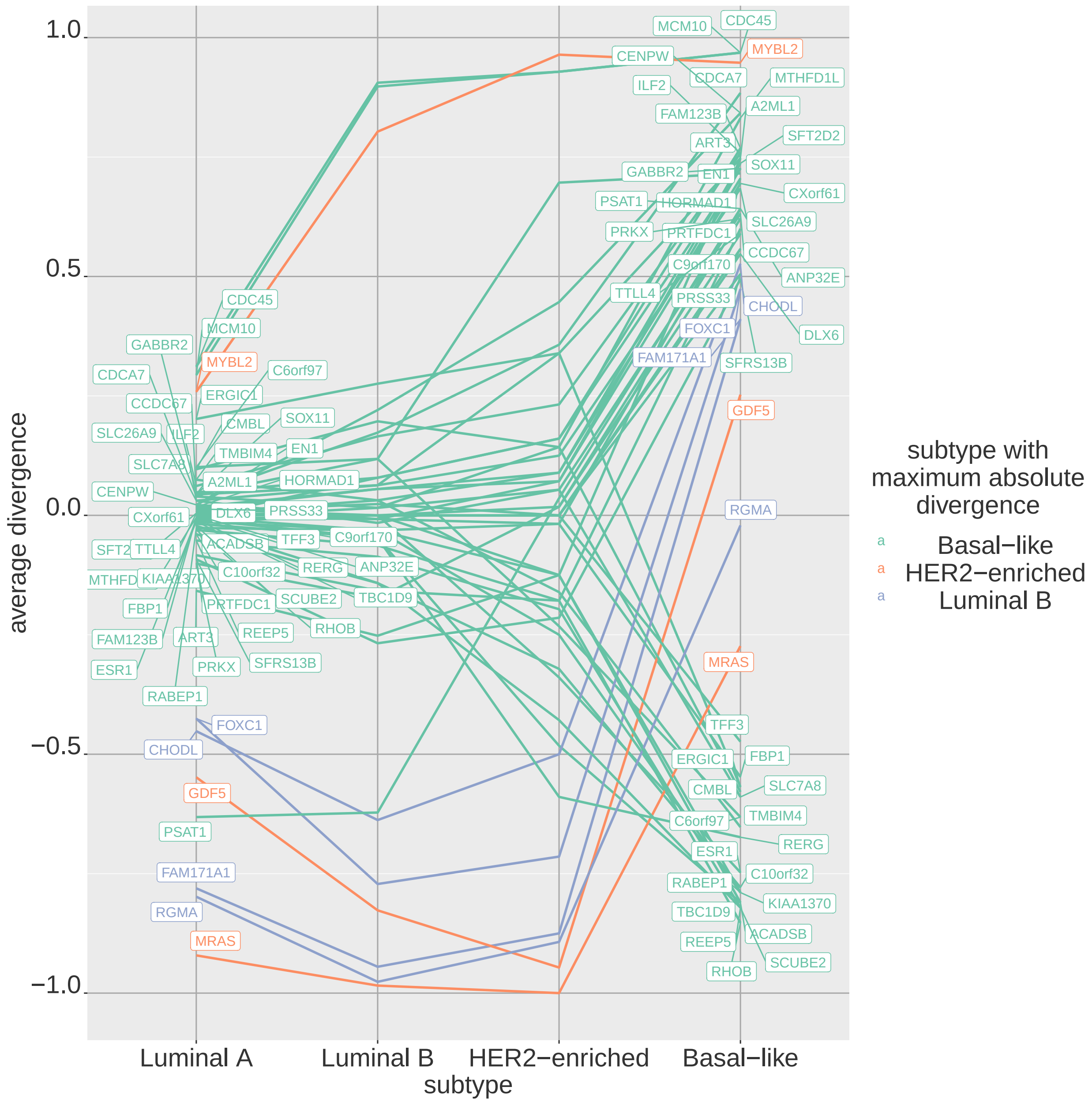

Figure 6: **Features with differring divergent patterns among PAM50 subtypes.** Genes with the largest variances between the average divergences among the PAM50 subtypes of breast tumor samples. For many of these genes, the Basal-like subtype, which is one of the most malignant of the PAM50 subtypes, corresponds to an extreme upper or lower divergence with respect to a normal breast baseline.

Given that the univariate divergence values are in the $\{-1, 0, 1\}$ set, another way to visualize divergence coded values is by proportion of samples in each one of the three states among different phenotypes. Figure 7 show this for 5 of the PAM50 genes. The size of each point is the percentage of samples in the -1, or 0, or 1 state, with the average divergence shown in the dotted lines. Alternatively, they can simply be viewed as barplots showing the proportion of values in each divergence state, as we show in the following section.

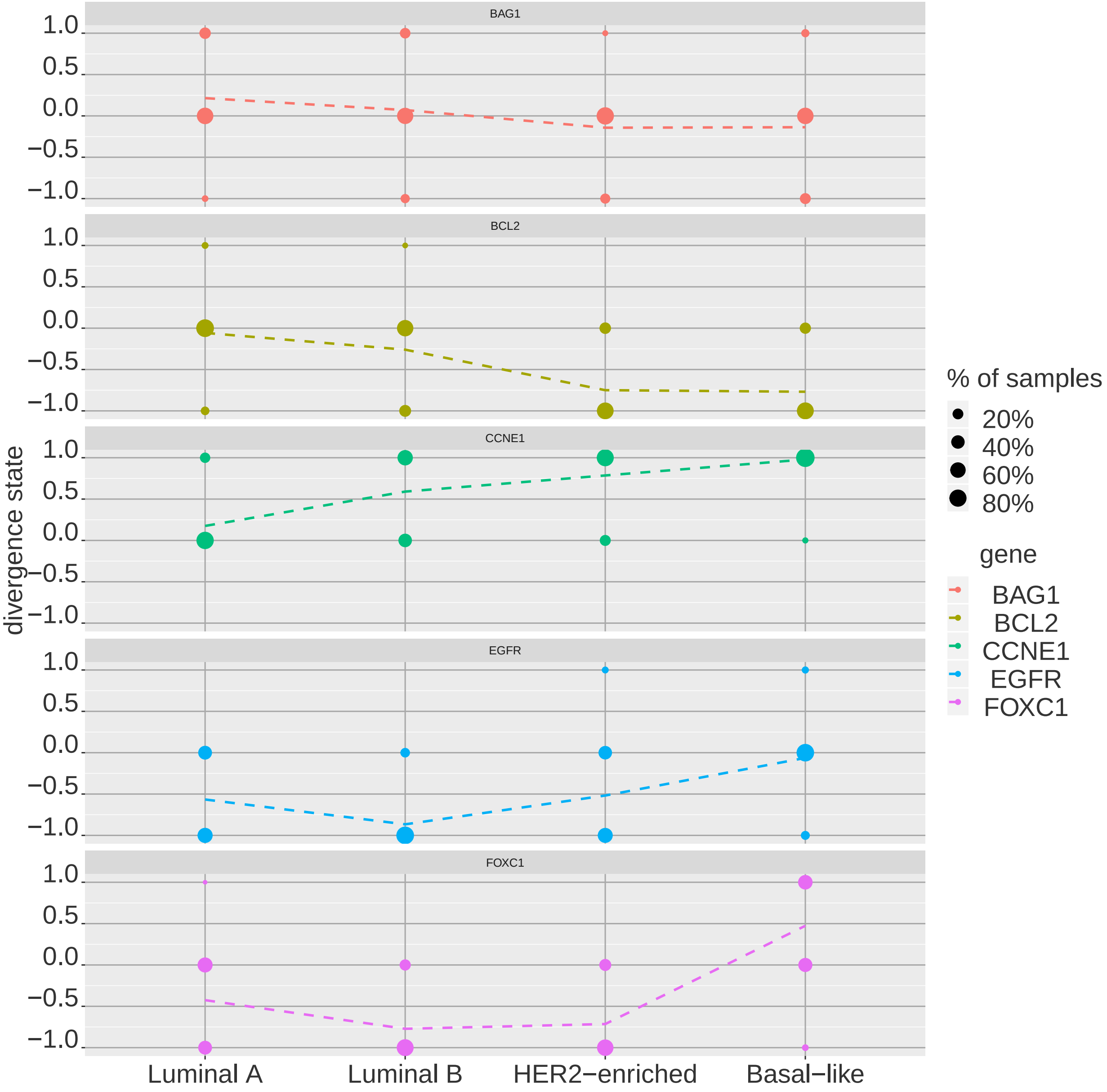

Figure 7: **An alternative representation of average divergences for a given feature.** Alternatively, we can plot the percentage of samples in a given group in each possible divergence state: -1, 0, or 1. We show this by PAM50 subtypes for a few PAM50 genes with breast tumor samples. The size of each point indicates the percentage of samples of each subtype in a given state.

2.2 Multivariate Workflow

In the multivariate scenario, the features of interest are composite - that is, they are sets of univariate features, for example a gene set indicating a disease specific signature or a pathway. The workflow is similar to that of the univariate case, with the exception that the divergence coding will be either 0 or 1 - that is, it indicates whether the feature is divergent or not, whereas for the univariate case it provided a direction of divergence as well.

With the same breast gene expression data from TCGA which we have used so far, in the following we examine the divergence coding obtained for the KEGG pathways from MSIGDB gene set collection[7]. Computing the divergence coding is similar to the univariate case, except that the multivariate features need to be specified. As with the univariate

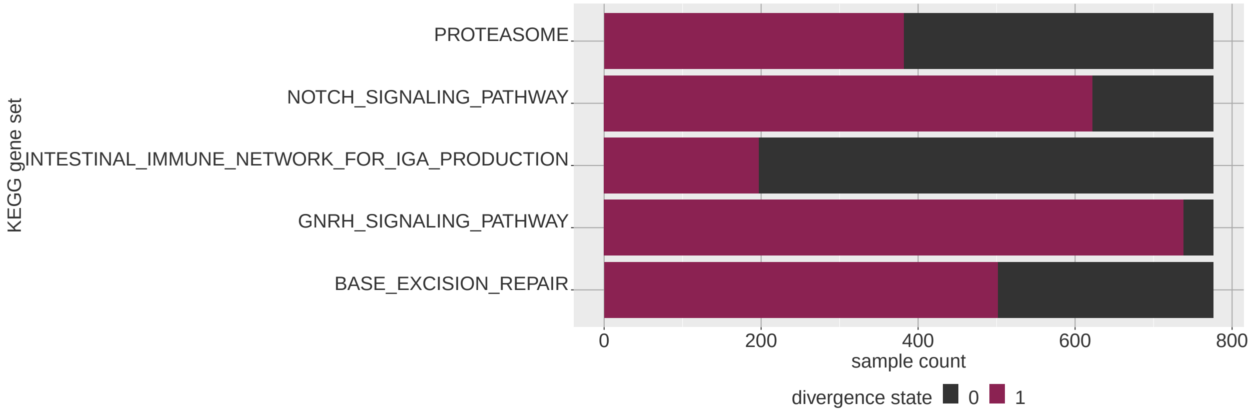

Figure 8: **Visualizing multivariate divergence by feature.** Given that multivariate divergence results in a binary digitization, for each feature we can observe how many samples are in a given state. Multivariate features (which are sets of genes, in this case) NOTCH signaling pathway and GNRH signaling pathway, for example, are much more likely to be divergent compared to the other features shown.

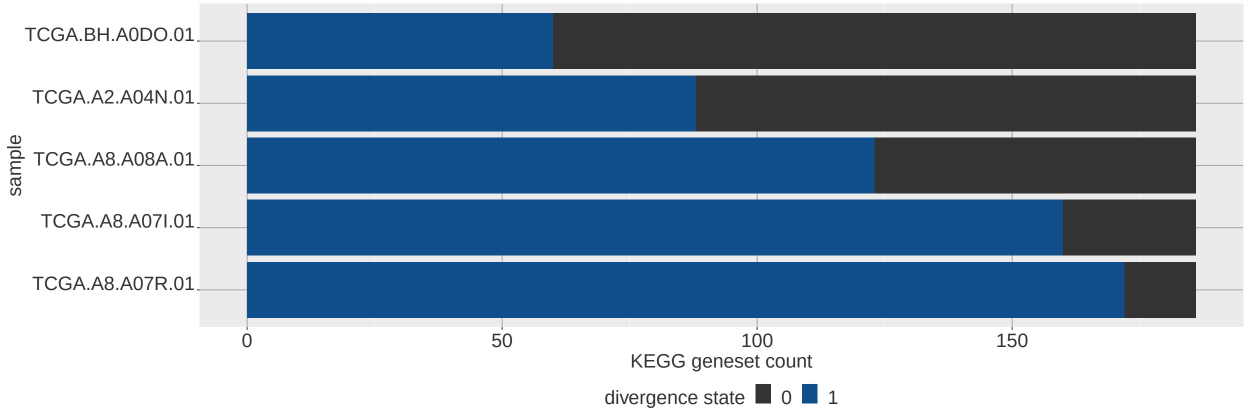

Figure 9: **Visualizing multivariate divergence by sample.** For a given set of multivariate features, we can see how likely they are to be divergent in a given sample. Here five selected breast tumor samples are ordered by how many gene sets in the KEGG gene set collection are divergent in each sample, with respect to a normal breast baseline.

case, the reference sample data and the samples for which the divergence coding need to be computed are provided in matrix form, along with a range of γ values to choose from, a β value, and the required α threshold. The multivariate features are provided as a list, where each element of the list is a vector of univariate features which are available in the data matrices provided. In the example we present, the same TCGA sourced RNA-Seq expression data used previously are used. The KEGG pathway list has 186 gene sets, where each set ranges from 10 to 389 gene symbols. These gene symbols are available in the RNA-Seq expression data and will be used to estimate the baseline support from the matched normal samples and the placement of the tumor samples either within or outside the support in the high dimensional space representing each KEGG gene set. The output will be a matrix comprising of the binary divergence coding for each KEGG pathway and each tumor sample.

Given that the divergence coding is a highly simplified representation, we can simply observe it in terms of either the features or the samples and observe the proportion of values in each divergence state. Figure 8 shows the sample proportions for some of the KEGG pathways, and similarly at the sample level in figure 9.

As before, we can apply a χ^2 test to each pathway between the ER+ and ER- sample groups to identify which pathways are highly differentially divergent between the two groups (Figure 10). The pathways that meet a $P \leq 0.05$ cutoff after bonferroni adjustment for multiple testing are shown in blue.

Figure 11 shows principal component analysis (PCA) applied to the KEGG pathway divergence coding. We simply take the binary matrix of breast tumor samples with the divergence coding for each KEGG pathway and apply PCA analysis

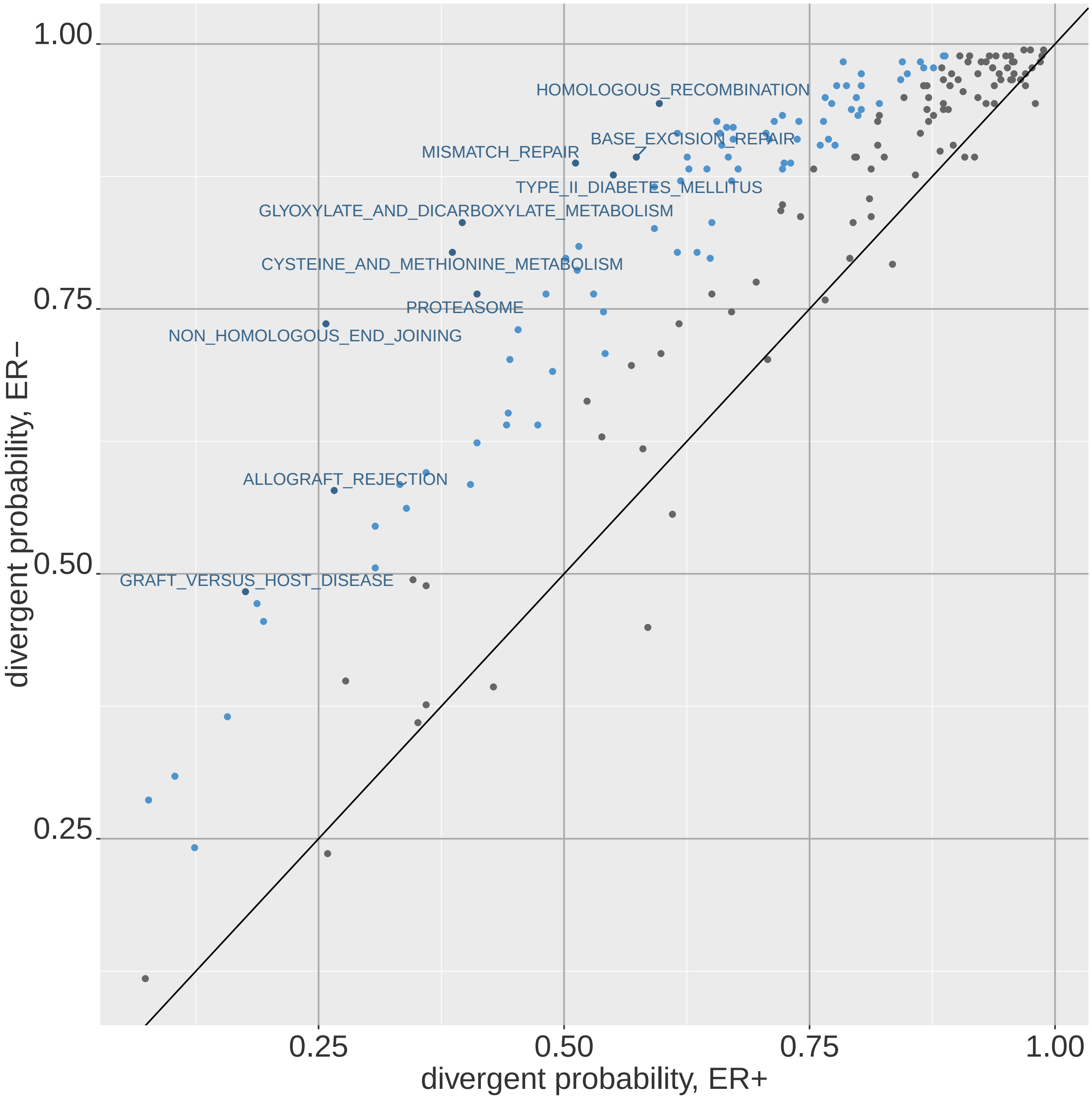

Figure 10: **Differentially divergent gene sets between phenotypes.** Similar to the univariate scenario, we can compare the divergence probability of multivariate features between groups of samples. The divergent probabilites among the ER+ and ER- breast tumor samples are shown here for KEGG gene sets, which the points in blue indicating meeting a bonferroni adjusted $P \leq 0.05$ threshold from a χ^2 test.

as we would with any other data matrix. The second princiapl component separates a large number of the Luminal A samples from the rest.

3 Conclusion

In the above results we have showcased a variety of analyses that can be performed with divergence at the univariate and multivariate level. While we have used RNA-seq data here, the software is applicable to many other modalities of high dimensional omics data, such as microarray data, CpG level methylation data, protein data, and microRNA expression data, some of which we have showed in [1].

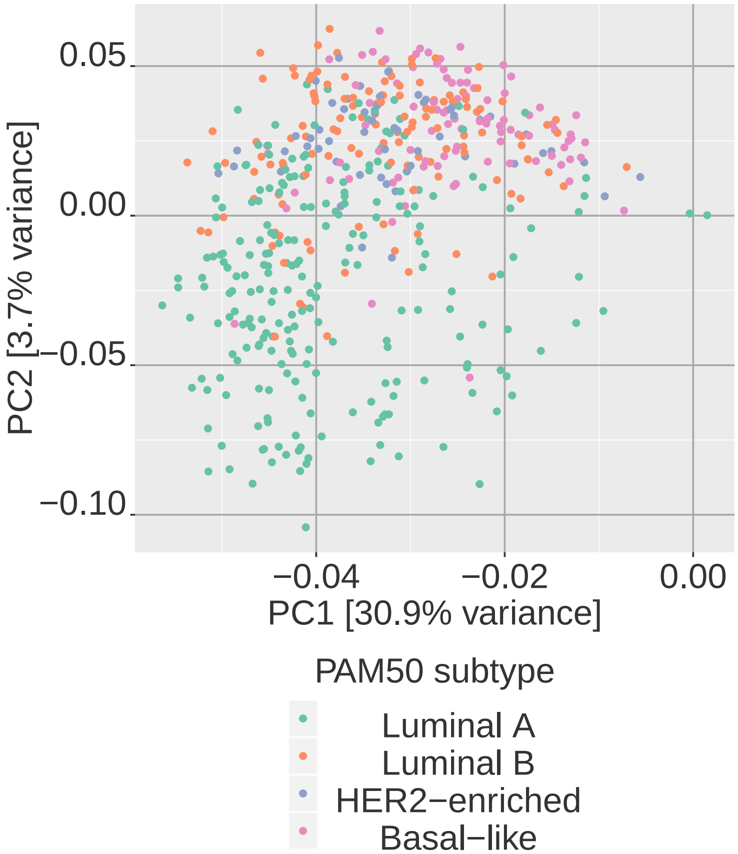

Figure 11: **Principal component analysis of multivariate divergence.** Digitized data can be analyzed with regular statistical tools such a principal component analysis (PCA). Here the first two PCs are shown for multivariate divergence values of the KEGG geneset collection for breast tumor samples with respect to a normal baseline.

Once the data has been processed as necessary and the baseline cohort identified, the R package can be used to compute the divergence coding quite easily. A full outline of the functions that can be used and the workflows possible are provided in the package vignette[5]. Here we have presented some of the many different ways that the digitized divergence coding can be visualized and analyzed by the user.

4 Acknowledgements

This publication was made possible though support from the NIH-NCI grants P30CA006973 and R01CA200859, and the Department of Defense (DoD) office of the Congressionally Directed Medical Research Programs (CDMRP) award W81XWH-16-1-0739.

5 References

References

[1] Wikum Dinalankara, Qian Ke, Yiran Xu, Lanlan Ji, Nicole Pagane, Anching Lien, Tejasvi Matam, Elana J Fertig, Nathan D Price, Laurent Younes, et al. Digitizing omics profiles by divergence from a baseline. *Proceedings of the National Academy of Sciences*, 115(18):4545–4552, 2018.

[2] Wikum Dinalankara and Héctor Corrada Bravo. Gene expression signatures based on variability can robustly predict tumor progression and prognosis. *Cancer Inform*, 14:71–81, 2015.

[3] Michael F. Ochs, Jason E. Farrar, Michael Considine, Yingying Wei, Soheil Meschinchi, and Robert J. Arceci. *Outlier Gene Set Analysis Combined with Top Scoring Pair Provides Robust Biomarkers of Pathway Activity*, pages 47–58. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[4] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[5] Wikum Dinalankara, Luigi Marchionni, and Qian Ke. *divergence: Divergence Computations*, 2019. R package version 1.1.0.

[6] The cancer genome atlas. [Online].

[7] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.