Sparse classification with paired covariates - Appendix

(Advances in Data Analysis and Classification)

Paired Data: Tumour Tissue vs Normal Tissue

Armin Rauschenberger[1,2], Iuliana Ciocănea-Teodorescu[1],
Marianne A. Jonker[3], Renée X. Menezes[1], and Mark A. van de Wiel[1,4]
(mark.vdwiel@amsterdamumc.nl)

[1]Department of Epidemiology and Biostatistics, Amsterdam UMC,
VU University Amsterdam, The Netherlands
[2]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg
[3]Department for Health Evidence, Radboud University Medical Center, The Netherlands
[4]MRC Biostatistics Unit, University of Cambridge, United Kingdom

1 Introduction

In cancer studies, the main tissue of interest is that of the primary tumour. However, there are many circumstances in which additional samples from the same patient may be collected and analysed. In whole exome/genome sequencing experiments, normal blood DNA samples from patients are often collected in order to determine somatic mutations present in the tumour. In gene expression experiments, normal tissue, adjacent to the tumour, is collected and used as a "baseline" sample, against which to measure differential gene expression. However, it is not agreed upon whether histologically normal tissue from the vicinity of a tumour can be considered entirely normal in the strictest sense of the word, as it is reasonable to assume that the molecular profile of such tissue will be influenced by the malignant process unfolding in its proximity (see [1, 9] and references therein).

In the particular case of survival studies, when both tumour tissue and normal tissue data are available, a question that arises is whether normal tissue contains survival signal that is complementary to that of the tumour tissue. This is a typical example of a paired data setting, where there are two data sets available, but only one response vector, and it is unknown which of the two data sets is expected to have a stronger predictive ability. Available methods for statistical analysis are not tailored to incorporate the full power of having paired data.

Our primary interest is in evaluating the performance of the palasso in comparison to the standard and the adaptive lasso in the tumour-normal paired data setting. In this process, we also investigate to which extent normal data is more predictive of survival than tumour data.

2 Methods

Paired normal-tumour TCGA RNA-Seq data sets of six cancer types were downloaded from the harmonised GDC database ([10]) using the R-package TCGAbiolinks ([4]). In each data set, only those genes were retained that had more than two counts per million in at least three samples. The data were normalised using the trimmed mean of M-values method (TMM) of the R-package edgeR ([11, 12]).

To explore the data, we produced multidimensional scaling plots, and examined the profile of p-values resulting from simple Cox proportional hazard regression on each gene, and the p-values obtained by applying the global test on the normal and tumour data separately ([6, 7]). Multiple testing correction was done using the Benjamini-Hochberg method ([2]).

We compared the Cox palasso to the standard Cox lasso and the adaptive Cox lasso ([5, 13]), with weights given by the inverse of the absolute values of the rescaled concordance indices of the standardised covariates. The models were each applied to the tumour tissue data, the normal tissue

data and the pooled data separately. The measure optimised by the cross-validation procedure used to select the regularisation parameter and weighting scheme is the log partial likelihood deviance. The cross-validated log partial likelihood for the k^{th} fold is given here by the difference of the log partial likelihood evaluated on the full data set, and that evaluated on the data set with the k^{th} fold excluded (see [13], equation (11)). To evaluate the performance of each of the models, we divided the data of each cancer type into 5 subsets, and used each of these subsets as a test set for the model trained on the remaining 4 subsets. We then computed the resulting log partial likelihood deviances as above.

The data were standardised prior to the analysis, and all cross-validation procedures were performed with five folds, which are fixed across models. For all models, we restricted the maximum number of non-zero coefficients allowed to 10.

3 Results

The summary of the data sets can be seen in Table 1, giving the number of patients in each cohort, the number of corresponding events (deaths), and the number of genes retained for analysis.

	Full Name of Study	Number of patients (Number of events)	Number of genes
TCGA-KIRC	Kidney renal clear cell carcinoma	71 (25)	18391
TCGA-LUAD	Lung adenocarcinoma	56 (22)	19376
TCGA-LIHC	Liver hepatocellular carcinoma	49 (30)	16729
TCGA-LUSC	Lung squamous cell carcinoma	47 (23)	18842
TCGA-HNSC	Head and Neck squamous cell carcinoma	43 (31)	17700
TCGA-BRCA	Breast invasive carcinoma	105 (37)	20936

Table 1: Data sets.

The primary tumour data is labelled as "TP", and the normal tissue data is labelled as "NT". As can be expected, there appears to be a substantial amount of differential expression between the tumour samples and the normal samples. This can be visualised with the help of multidimensional scaling plots, shown in Figure 3.

To evaluate the strength of the survival signal in the data, we performed simple Cox regression on each gene, in the tumour data and the normal data separately (Figures 4, 5, 6). We did not see any pattern as to which data set is more informative with regard to survival across cancer types (cf. [9]). In the case of TCGA-KIRC both data sets exhibit comparable, relatively strong, survival signal. For TCGA-LIHC and TCGA-HNSC, normal tissue seems to be more important for survival, while for TCGA-BRCA it seems to be the tumour tissue that is more important. For TCGA-LUAD and TCGA-LUSC, none of the data sets seem to contain any survival information. These results are corroborated to a great extent by the p-values of the global test.

	TP min adj p-value	NT min adj p-value	TP global test	NT global test
TCGA-KIRC	0.007	0.002	0.000	0.069
TCGA-LUAD	0.300	1.000	0.033	0.775
TCGA-LIHC	0.312	0.042	0.098	0.023
TCGA-LUSC	1.000	0.276	0.662	0.185
TCGA-HNSC	0.387	0.096	0.124	0.296
TCGA-BRCA	0.181	0.999	0.089	0.851

Table 2: The minimum adjusted p-values obtained from simple Cox regression and the p-values obtained from the global test.

For each cancer type we compared palasso with 2x3 = 6 standard models: the ordinary lasso and adaptive lasso as applied on the tumour data, the normal data and the pooled data. We compared the results by ranking the log partial likelihood deviances, as computed on the left-out samples using five-fold cross-validation [14]. The model resulting from the fourth weighting scheme is unique for palasso. If palasso chooses this model during the training phase, we assess its performance across seven models. Otherwise, palasso is one of the standard adaptive models, in which case its performance is assessed across six models.

We also display the performance of palasso on training samples. This allows us to judge from the training data whether adaptation is deemed useful or not. Note that by definition of palasso, which contains the 4 adaptive models in its bag, it can only be outperformed by one of the three lasso models here, so the maximal rank equals 4. We display the results for all five folds (which are fixed for all models), because different folds may lead to different chosen models, plus it allows assessing the stability of the results.

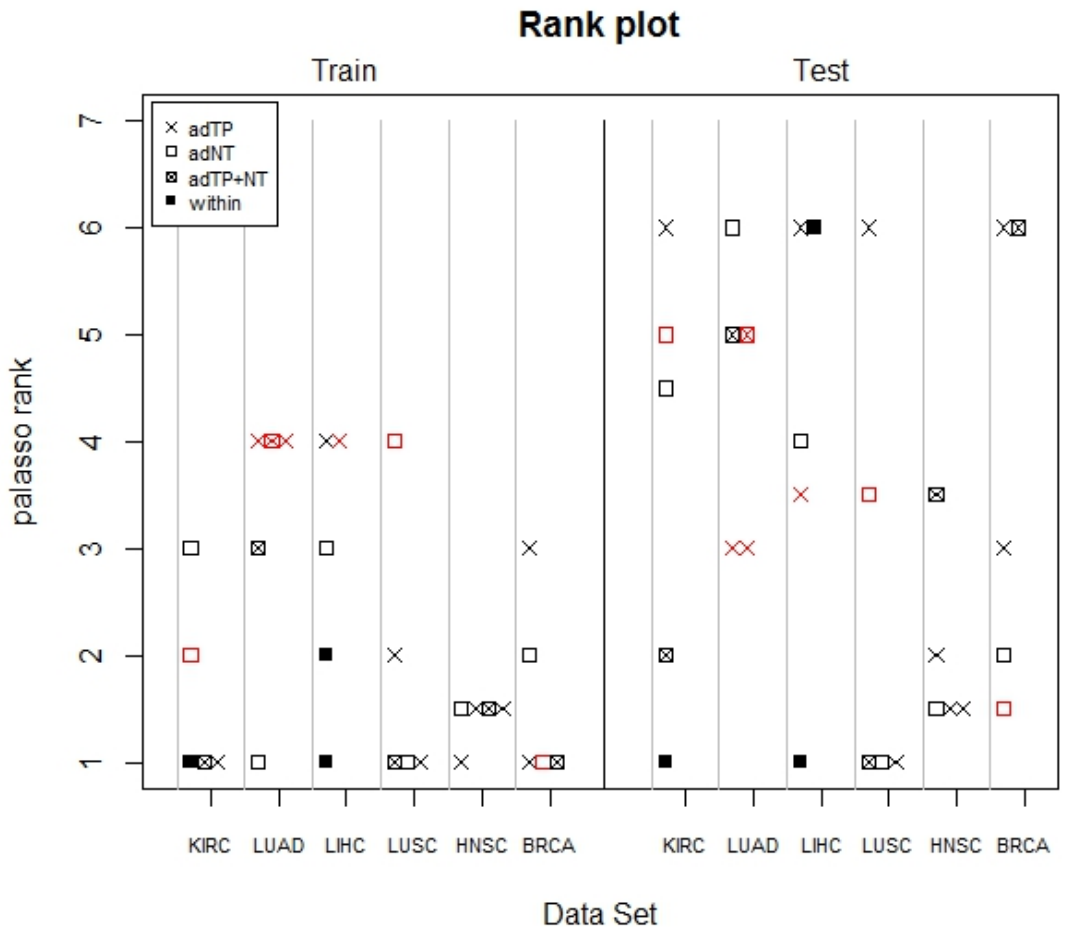

Figure 1: Ranks of the palasso models upon training (left), and testing (right). Each symbol represents the model chosen by palasso. The tumour tissue data is coded as TP and the normal tissue data is code as NT. The fourth weighting scheme of the palasso is the "within"-scheme. Ranks in the training are computed out of 4. Ranks in the testing are computed out of 6 or 7. Red symbols indicate that the model chosen by palasso was the empty model.

The number of non-zero coefficients estimated across all models and all training sets, for each cancer type, is summarised in Figure 2.

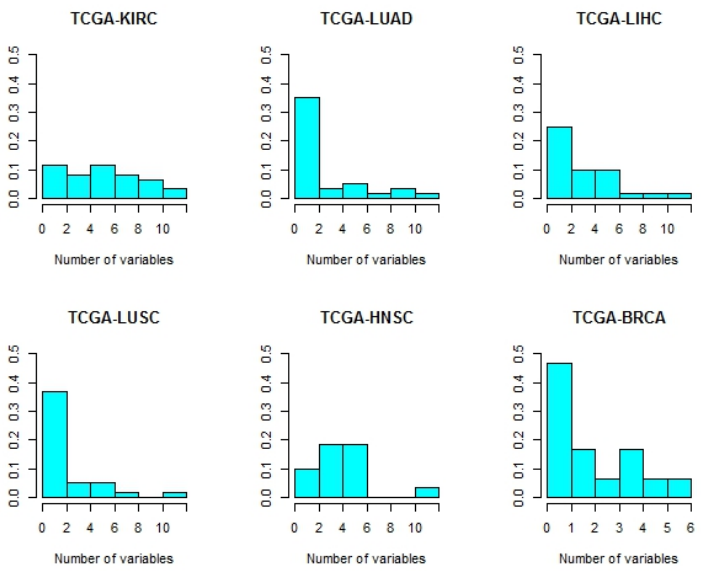

Figure 2: Histograms of the number of non-zero coefficients across all training sets and all models.

4 Discussion

First, we observe that for most cancer types there is no clear connection between the p-values from the profiles and the global test (see Table 2) on one side and the predictive performance of the palasso, lasso and adaptive lasso models on the other side. It seems that for at least 4 out of 6 cancer types, lasso is not able to detect the signal, which is corroborated by Figure 2 showing that many of the trained models are empty for those types. These results show that palasso is not a 'miracle tool': if the standard lasso models do not perform well, it is unlikely that palasso will enhance the results.

For KIRC, both the p-value results and the variable selection by lasso models do indicate a presence of signal. Here, however, palasso does not improve the test performance of the other lasso models. When comparing the test set results of the adaptive lasso models with those of their standard counterparts (results not shown) this seems mostly a consequence of the fact that adaptation does not improve the results. Given that palasso is essentially a bag of adaptive models, it is no surprise palasso does not outperform the other models for KIRC. For HNSC, the palasso shows test set results that are somewhat better than average. Here, the adaptive lasso on TP (tumour tissue) also performs well (average rank across 5 folds: 2; data not shown), so there is no advantage of pairing. Nevertheless, this can only be observed a posteriori, so the fact that palasso *automatically* chooses a model is a practical advantage.

For most of the cancer types, the p-value results suggest that dense models may be a better alternative in this setting. Indeed, there is substantial literature devoted to the comparison of different prediction models for survival from gene expression data, and how to evaluate their performances [3, 15, 16]. In these comparisons, the lasso is typically outperformed by ridge regression, regardless of the measure of evaluation. Alternatively, prior aggregation of the gene expression values (e.g. by using prior information and/or dimension reduction techniques like principle component analysis or partial least squares) may improve survival prediction [15].

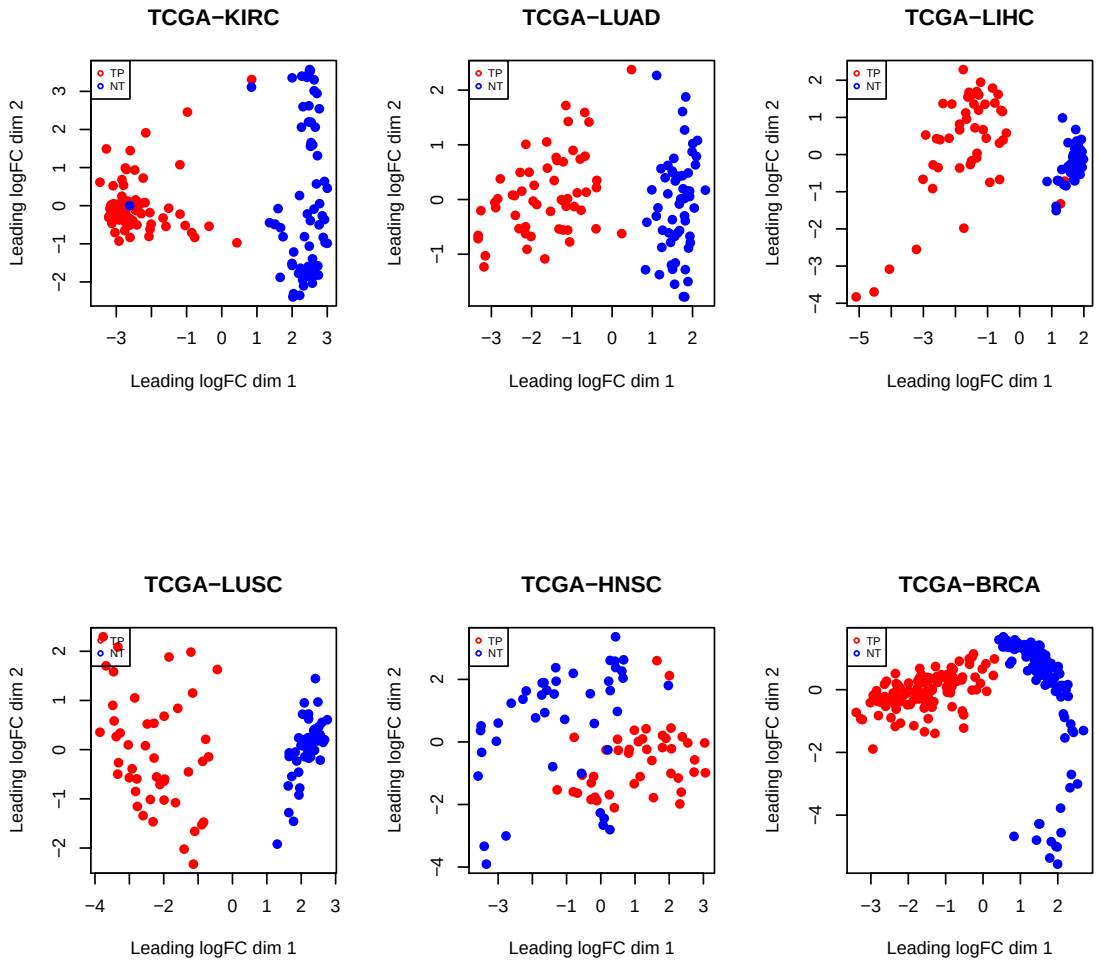

Figure 3: MDS plots of the pooled data sets: in blue the data points corresponding to normal tissue and in red the data points corresponding to the tumour tissue.

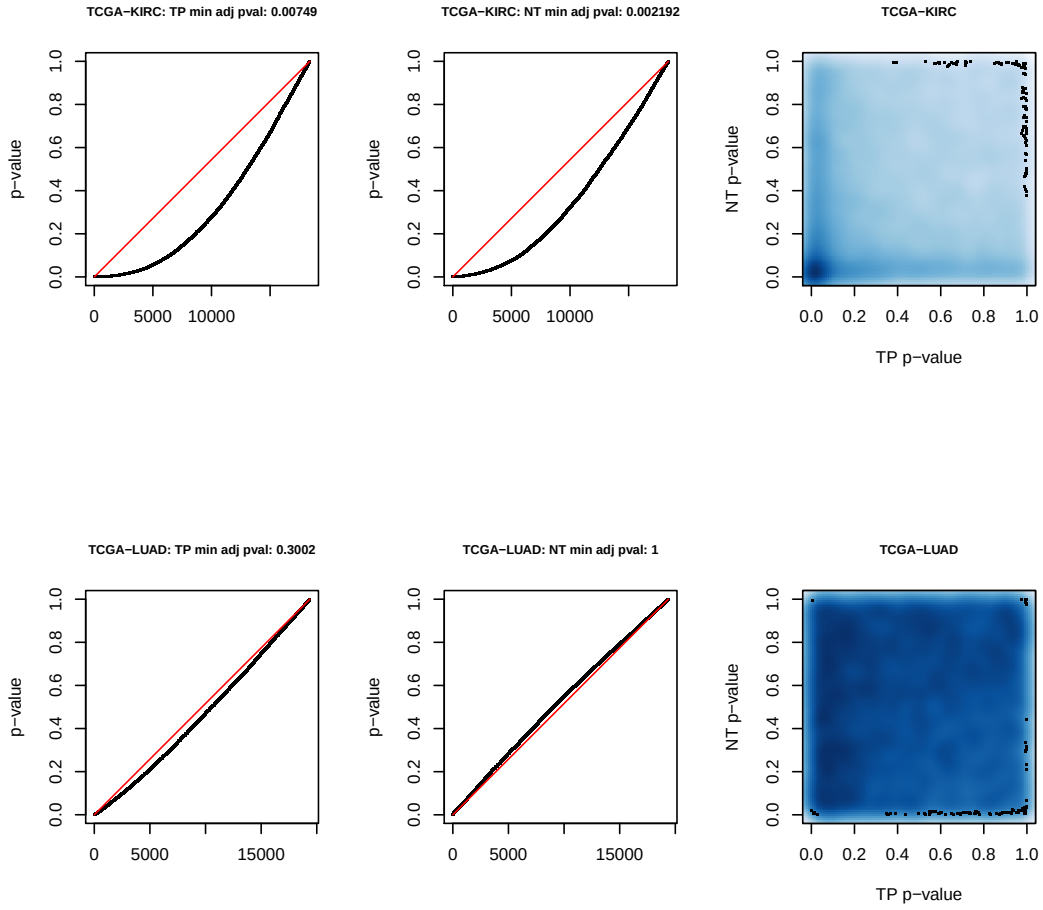

Figure 4: Plot of ordered p-values obtained from simple Cox regression for each gene, together with a scatterplot of the p-values of the genes in the tumour data against those in the normal data.

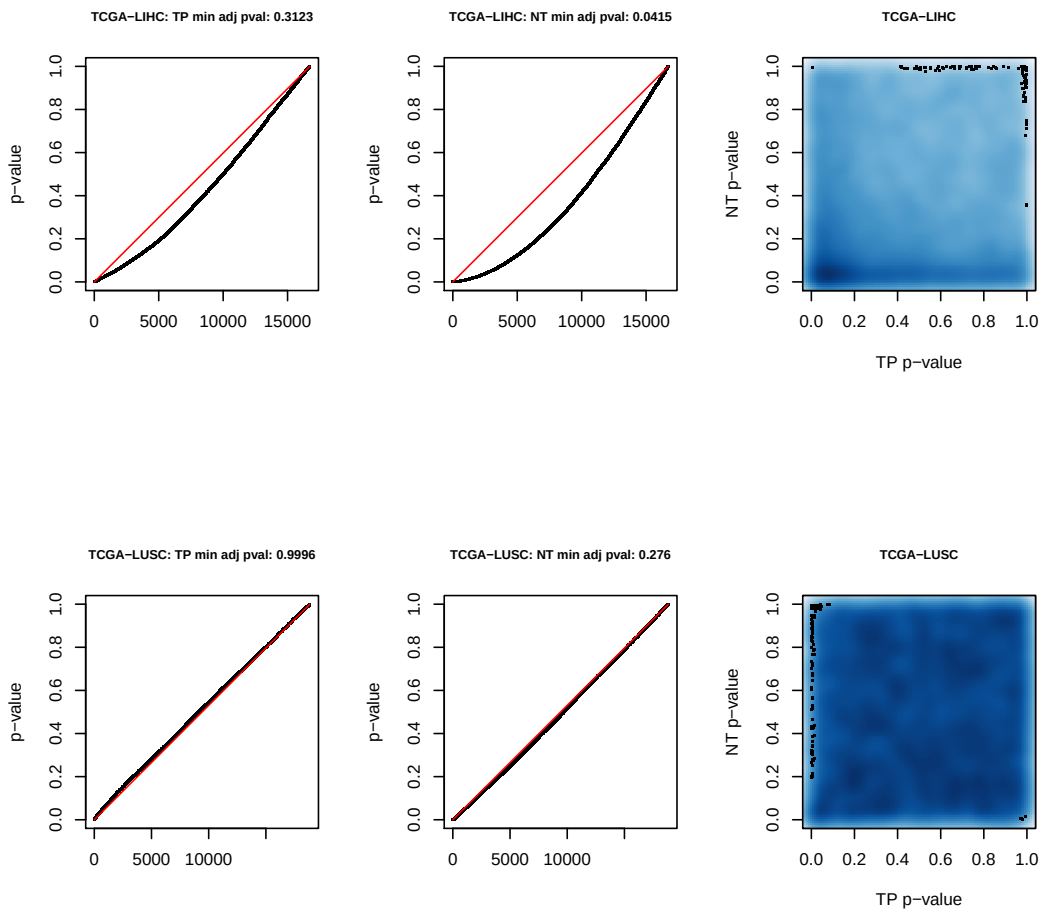

Figure 5: Plot of ordered p-values obtained from simple Cox regression for each gene, together with a scatterplot of the p-values of the genes in the tumour data against those in the normal data.

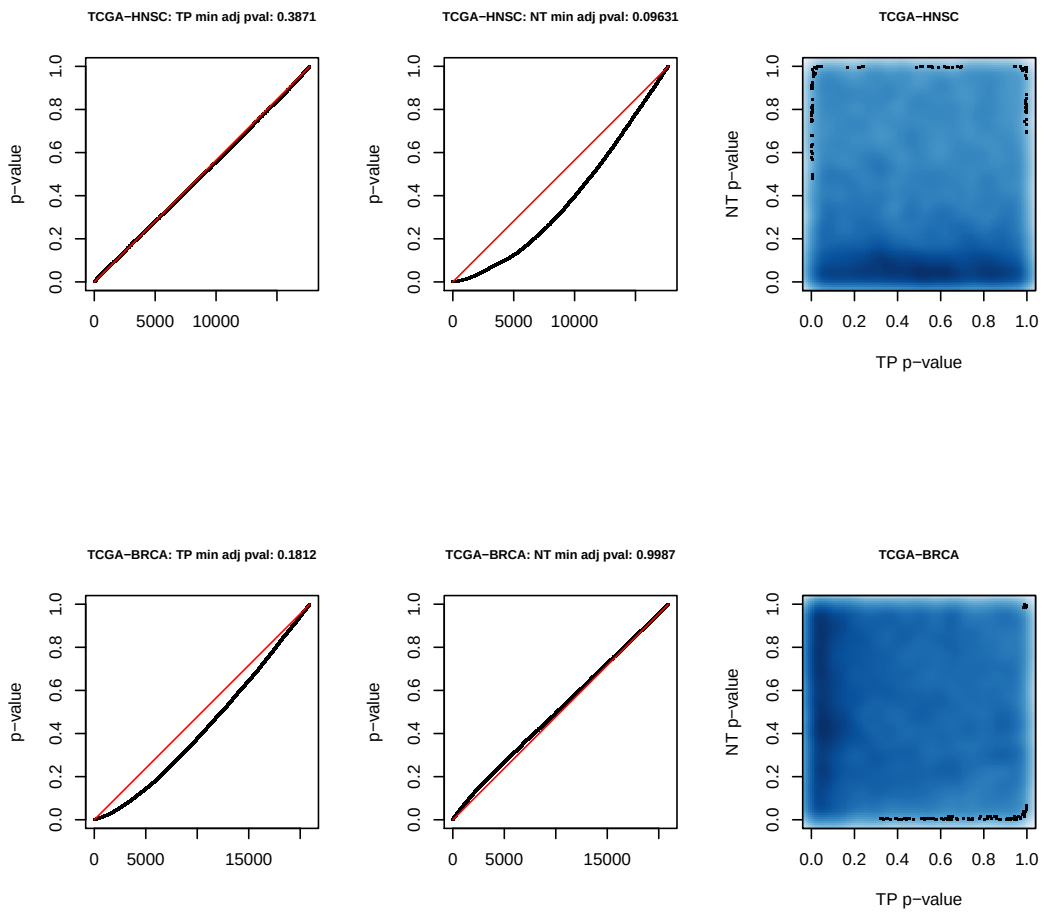

Figure 6: Plot of ordered p-values obtained from simple Cox regression for each gene, together with a scatterplot of the p-values of the genes in the tumour data against those in the normal data.

References

[1] Dvir Aran, Roman Camarda, Justin Odegaard, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nature Communications*, 8(1):1077, 2017.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[3] Hege M. Bøvelstad and Ørnulf Borgan. Assessment of evaluation criteria for survival prediction from genomic data. *Biometrical Journal*, 53(2):202–216, 2011.

[4] Antonio Colaprico, Tiago C. Silva, Catharina Olsen, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8):e71, 2016.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[6] Jelle J. Goeman, Sara A. van de Geer, Floor de Kort, and Hans C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[7] Jelle J. Goeman, Sara A. van de Geer, and Hans C. van Houwelingen. Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):477–493, 2006.

[8] Patrick J. Heagerty and Yingye Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, 2005.

[9] Xiu Huang, David F. Stern, and Hongyu Zhao. Transcriptional profiles from paired normal samples offer complementary information on cancer patient survival evidence from TCGA pan-cancer data. *Scientific Reports*, 6:20567, 2016.

[10] The Cancer Genome Atlas Research Network. `http://cancergenome.nih.gov/`.

[11] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[12] Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

[13] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011.

[14] Hans C. Van Houwelingen, Tako Bruinsma, Augustinus A. Hart, Laura J. Van 't Veer, and Lodewyk F. Wessels. Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine.*, 25:3201–3216, 2006.

[15] Wessel N. van Wieringen, David Kun, Regina Hampel, and Anne-Laure Boulesteix. Survival prediction using gene expression data: A review and comparison. *Computational Statistics & Data Analysis*, 53(5):1590–1603, 2009.

[16] Daniela M. Witten and Robert Tibshirani. Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19(1):29–51, 2010.