1 Genes identified for the DLBCL data

In this supplement, we perform AIMER on all four datasets (DLBCL [20], breast cancer [6], lung cancer [3], AML [22]) discussed in the manuscript. Rather than using training sets containing 50% of the data as in the main paper, we use all the observations here. We allow our method to select up to as many features as there are observations.

1.1 DLBCL

For the DLBCL data, we not only list the selected genes, but also attempt to find any discussion of those genes in existing literature. Our final estimated model uses 49 gene features, which correspond to 26 genes. To examine the relevance of each selected gene for DLBCL, we adopt two approaches. The first endeavors to find literature examining the biological connection of the identified gene to any type of lymphoma. The second lists any reference in the (rather lengthy) methodological literature in statistics, computer science, and bioinformatics that uses statistical or machine learning methods to examine the DLBCL dataset.

We display our findings for all 26 genes in Table 1. To summarize, 16 out of the 26 genes have been related to lymphoma in the biological literature, and 19 of them have already been identified via statistical techniques developed for the DLBCL dataset. While many of the 26 genes have been previously connected to lymphoma in general and DLBCL in particular, AIMER does identify 4 genes with symbols ALDH2, CELF2, COL16A1, and DHRS9 that have not been previously identified in the biological or methodological literature. We note that, while we have made every effort to locate each gene, given the large and evolving literature on this topic, those we have been unable to locate may have none-the-less been previously studied.

1.2 Genes identified for breast cancer, lung cancer and AML data

As before, we allow the maximum number of selected genes be the same as the total number of patients. AIMER identifies 78 genes with breast cancer data, 12 genes for lung cancer, and 50 genes for the AML dataset. We list the top 20 selected genes for breast cancer in Table 2, all 12 selected genes for lung cancer in Table 3, and the top 20 selected genes for AML in Table 4.

2 Alternative analysis for lung cancer data

Compared with the other three datasets, the public lung cancer data comes presents gene expression measurements for only patients who have been diagnosed with lung cancer. The other three datasets instead give the logarithm of the ratio between diseased sample expression measurements and a reference control group. To try to make the lung cancer dataset comparable to the others, we perform two separate transformations on the data. The first transformation is to take the base-2 logarithm of all the expression measurements. Because some measurements are negative, before taking the logarithm, we first add the negative of the minimum value plus one to each feature vector, making all measurements at least 1. This transformation mimics the standard process. The second transformation orthonormalizes the gene expression matrix.
| Symbol | In biology | Source(s) | In methodology | Source(s) | Name of gene |
|--------|------------|-----------|----------------|-----------|--------------|
| 1 ALDH2 | ×          |           | ×              |           | aldehyde dehydrogenase 2 family (mitochondrial) |
| 2 BCL2  | ✓          | [4, 9]    | ✓              | [4, 12]   | BCL2, apoptosis regulator |
| 3 CCND2 | ✓          | [4]       | ✓              | [4, 12, 16] | cyclin D2 |
| 4 CELF2 | ×          |           | ×              |           | CUGBP Elav-like family member 2 |
| 5 COL3A1 | ✓          | [4, 20]   | ✓              | [4]       | collagen type III alpha 1 chain |
| 6 COL16A1 | ×          |           | ×              |           | collagen type XVI alpha 1 chain |
| 7 CR2   | ×          |           | ✓              | [13, 16]  | complement C3d receptor 2 |
| 8 CYP27A1 | ×         |           | ✓              | [12]      | cytochrome P450 family 27 subfamily A member 1 |
| 9 Dhrs9 | ×          |           | ×              |           | dehydrogenase/reductase 9 |
| 10 EPHB1 | ✓          | [2]       | ✓              | [24]      | EPH receptor B1 |
| 11 ESTs | ✓          | [20]      | ✓              | [11]      | ESTs |
| 12 FN1  | ✓          | [4, 20]   | ✓              | [4, 12]   | fibronectin 1 |
| 13 FUT8 | ×          |           | ✓              | [10]      | fucosyltransferase 8 |
| 14 IGHM | ×          |           | ✓              | [4, 16, 24] | immunoglobulin heavy constant mu |
| 15 IGKC | ✓          | [23]      | ✓              | [16, 24]  | immunoglobulin kappa constant |
| 16 IRF4 | ✓          | [1, 18]   | ✓              | [4, 10]   | interferon regulatory factor 4 |
| 17 KIAA0233 | ✓ | [4, 20] | ✓ | [4] | KIAA0233 gene product |
| 18 LMO2 | ✓          | [1, 17]   | ✓              | [4, 11, 12] | LIM domain only 2 |
| 19 MAPK10 | ✓        | [23]      | ✓              | [4, 11, 24] | mitogen-activated protein kinase 10 |
| 20 MME  | ×          |           | ✓              | [4]       | membrane metalloendopeptidase |
| 21 MMP2 | ✓          | [8]       | ✓              | [13]      | matrix metallopeptidase 2 |
| 22 MMP7 | ✓          | [14]      | ×              |           | matrix metallopeptidase 7 |
| 23 MMP9 | ✓          | [1, 21]   | ✓              | [11]      | matrix metallopeptidase 9 |
| 24 MYB  | ✓          | [7]       | ✓              | [4]       | MYB proto-oncogene, transcription factor |
| 25 SPARC | ✓          | [5, 15]   | ×              |           | secreted protein acidic and cysteine rich |
| 26 VPREB3 | ✓        | [19]      | ×              |           | V-set pre-B cell surrogate light chain 3 |

Table 1 DLBCL Predictive Genes. AIMER selected 26 genes. We note that while we have made every effort to locate all 26 genes in the literature, a × should be taken to indicate that we were unable to locate a reference for that gene rather than the stronger conclusion that no one has yet investigated it.

| Gene | Symbol |
|------|--------|
| 1    | Contig47405_RC |
| 2    | NM_002964  |
| 3    | NM_002965  |
| 4    | NM_005980  |
| 5    | Contig43983_RC |
| 6    | NM_017422  |
| 7    | NM_002963  |
| 8    | NM_020974  |
| 9    | Contig50360_RC |
| 10   | Contig55725_RC |
| 11   | NM_018265  |
| 12   | NM_006115  |
| 13   | AK001423   |
| 14   | NM_004525  |
| 15   | Contig38438_RC |
| 16   | AL050227   |
| 17   | NM_014479  |
| 18   | NM_002421  |
| 19   | NM_000266  |
| 20   | NM_006419  |

Table 2 Top 20 selected genes for breast cancer dataset by AIMER.

| Gene | Symbol |
|------|--------|
| 1    | D49824_s_at |
| 2    | X57809_s_at |
| 3    | M17886_at |
| 4    | S71043_rna1_s_at |
| 5    | M87789_s_at |
| 6    | V00594_s_at |
| 7    | X98482_r_at |
| 8    | M34516_at |
| 9    | hum_alu_at |
| 10   | HG2873-HT3017_at |
| 11   | HG3364-HT3541_at |
| 12   | HG3549-HT751_at |

Table 3 12 selected genes for lung cancer dataset by AIMER.
Table 4 Top 20 selected genes for AML dataset by AIMER.

| Methods                   | original dataset | log2 transformation | normalization |
|---------------------------|------------------|--------------------|---------------|
|                           | MSE   | # genes | d | MSE   | # genes | d | MSE   | # genes | d |
| lasso                     | 0.8159 | 22    |   | 0.8722 | 16  |    | 0.7921 | 20  |    |
| ridge                     | 0.7713 | 7129   |   | 0.7594 | 7129 |    | 0.7687 | 7129 |    |
| SPC                       | 0.8344 | 19    | 3 | 0.8268 | 32  | 5  | 0.7799 | 22  | 3  |
| SPC+lasso                 | 0.8436 | 9    | 4 | 0.8376 | 25  | 4  | 0.7864 | 19  | 3  |
| AIMER(b = 0)              | 0.9444 | 7129  | 1 | 0.9570 | 7129 | 1  | 4.5202 | 7129 | 3  |
| AIMER                     | 1.0203 | 13    | 1 | 0.8901 | 13  | 2  | 0.8244 | 42  | 4  |

Table 5 The MSE on the test set, the number of selected genes, and the number of principal components used (d if relevant), each averaged across the 10 random training-testing splits on the three datasets respectively.

We use the same training and testing procedure as in the main paper on the original dataset and the two transformed datasets. Table 5 shows the corresponding prediction MSE, the number of selected genes, and the number of components used (when necessary) averaged over 10 training-testing splits. It turns out that both the log2 transformation and normalization improves AIMER relative to the other methods. The number of components used in AIMER also increases. The number of selected genes for AIMER on the log2 transformed dataset is the same as with the original dataset, but AIMER selects more genes on the normalized dataset. However, even after these two transformations, AIMER is still not quite as accurate as SPC. We posit that using the conventional transformation with a control group may enhance the results for AIMER.
References

[1] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X. et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403 (6769), 503–511.

[2] Asmar, F., Punj, V., Christensen, J., Pedersen, M.T., Pedersen, A., Nielsen, A.B., Høther, C., Ralfkiaer, U., Brown, P., Ralfkiaer, E. et al. (2013) Genome-wide profiling identifies a DNA methylation signature that associates with TET2 mutations in diffuse large B-cell lymphoma. *Haematologica*, 98 (12), 1912.

[3] Beer, D.G., Kardia, S.L., Huang, C.C. et al. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine*, 8 (8), 816–824.

[4] Blenk, S., Engelmann, J., Weniger, M., Schultz, J., Dittrich, M., Rosenwald, A., Müller-Hermelink, H.K., Müller, T. and Dandekar, T. (2007) Germinal center B cell-like (GCB) and activated B cell-like (ABC) type of diffuse large B-cell lymphoma (DLBCL): analysis of molecular predictors, signatures, cell cycle state and patient survival. *Cancer Informatics*, 3, 399–420.

[5] Brandt, S., Montagna, C., Georgis, A., Schüßler, P.G., Bühler, M.M., Seifert, B., Thiesler, T., Curioni-Fontecedro, A., Hegyi, I., Dehler, S. et al. (2013) The combined expression of the stromal markers fibronectin and SPARC improves the prediction of survival in diffuse large B-cell lymphoma. *Experimental Hematology & Oncology*, 2 (1), 27.

[6] Bullinger, L., Döhner, K., Bair, E. et al. (2004) Gene expression profiling identifies new subclasses and improves outcome prediction in adult myeloid leukemia. *The New England Journal of Medicine*, 350 (16), 1605–1616.

[7] Dai, Y.-H., Hung, L.-Y., Chen, R.-Y., Lai, C.-H. and Chang, K.-C. (2016) ON 01910. Na inhibits growth of diffuse large B-cell lymphoma by cytoplasmic sequestration of sumoylated C-MYB/TRAF6 complex. *Translational Research*, 175, 129–143.

[8] Gouda, H.M., Khorsheid, M.M., El Sissy, M.H., Shaheen, I.A.M. and Molsen, M.M.A. (2014) Association between matrix metalloproteinase 2 (MMP2) promoter polymorphisms and the susceptibility to non-Hodgkin’s lymphoma in Egyptians. *Annals of Hematology*, 93 (8), 1313–1318.

[9] Kramer, M., Hermans, J., Parker, J., Krol, A., Kluin-Nelemans, J., Haak, H., Van Groningen, K., Van Krieken, J., De Jong, D. and Kluin, P.M. (1996) Clinical significance of bcl2 and p53 protein expression in diffuse large B-cell lymphoma: a population-based study. *Journal of Clinical Oncology*, 14 (7), 2131–2138.

[10] Li, C., Zhu, B., Chen, J. and Huang, X. (2015) Novel prognostic genes of diffuse large B-cell lymphoma revealed by survival analysis of gene expression data. *OncoTargets and Therapy*, 8, 3407.

[11] Liu, Z., Chen, D., Tan, M., Jiang, F. and Gartenhaus, R.B. (2010) Kernel based methods for accelerated failure time model with ultra-high dimensional data. *BMC bioinformatics*, 11 (1), 606.

[12] Lossos, I.S., Czerwinski, D.K., Alizadeh, A.A., Wechsler, M.A., Tibshirani, R., Botstein, D. and Levy, R. (2004) Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *New England Journal of Medicine*, 350 (18), 1828–1837.

[13] Ma, S. and Huang, J. (2007) Additive risk survival model with microarray data. *BMC bioinformatics*, 8 (1), 192.

[14] Matsumoto, T., Kumagai, J., Hasegawa, M., Tamaki, M., Aoyagi, M., Ohno, K., Mizusawa, H., Kitagawa, M., Eishi, Y. and Koike, M. (2008) Significant increase in the expression of matrix metalloproteinase 7 in primary CNS lymphoma. *Neurology*, 28 (3), 277–285.

[15] Meyer, P.N., Fu, K., Greiner, T., Smith, L., Delabie, J., Gascoyne, R., Ott, G., Rosenwald, A., Braziel, R., Campo, E. et al. (2011) The stromal cell marker SPARC predicts for survival in patients with diffuse large B-cell lymphoma treated with rituximab. *American Journal of Clinical Pathology*, 135 (1), 54–61.
[16] Miyazaki, K., Yamaguchi, M., Suguro, M., Choi, W., Ji, Y., Xiao, L., Zhang, W., Ogawa, S., Katayama, N., Shiku, H. et al. (2008) Gene expression profiling of diffuse large B-cell lymphoma supervised by CD21 expression. *British journal of haematology*, 142 (4), 562–570.

[17] Natkunam, Y., Zhao, S., Mason, D.Y., Chen, J., Taidi, B., Jones, M., Hammer, A.S., Dutoit, S.H., Lossos, I.S. and Levy, R. (2007) The oncoprotein LMO2 is expressed in normal germinal-center B cells and in human B-cell lymphomas. *Blood*, 109 (4), 1636–1642.

[18] Radivojac, P., Peng, K., Clark, W.T., Peters, B.J., Mohan, A., Boyle, S.M. and Mooney, S.D. (2008) An integrated approach to inferring gene–disease associations in humans. *Proteins: Structure, Function, and Bioinformatics*, 72 (3), 1030–1037.

[19] Rodig, S.J., Kutok, J.L., Paterson, J.C., Nitta, H., Zhang, W., Chapuy, B., Tumwine, L.K., Montes-Moreno, S., Agostinelli, C., Johnson, N.A. et al. (2010) The pre-B-cell receptor associated protein VpreB3 is a useful diagnostic marker for identifying c-MYC translocated lymphomas. *Haematologica*, 95 (12), 2056–2062.

[20] Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.L., Gascoyne, R.D., Muller-Hermelink, H.K., Smeland, E.B., Giltnane, J.M. et al. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine*, 346 (25), 1937–1947.

[21] Sakata, K., Satoh, M., Someya, M., Asanuma, H., Nagakura, H., Oouchi, A., Nakata, K., Kogawa, K., Koito, K., Hareyama, M. et al. (2004) Expression of matrix metalloproteinase 9 is a prognostic factor in patients with non-Hodgkin lymphoma. *Cancer*, 100 (2), 356–365.

[22] Van’t Veer, L.J., Dai, H., Van De Vijver, M.J. et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415 (6871), 530–536.

[23] Ying, J. and Gao, Z. (2010) Frequent epigenetic silencing of proapoptotic gene MAPK10 by methylation in B-cell lymphoma. *Journal of Leukemia & Lymphoma*, 19 (5), 272–275.

[24] Zhao, Y. and Wang, G. (2010) Additive risk analysis of microarray gene expression data via correlation principal component regression. *Journal of Bioinformatics and Computational Biology*, 8 (04), 645–659.