Supporting Information

Nicolau et al. 10.1073/pnas.1102826108

SI Text

1. Microarray Data Analysis. We provide details for the microarray data analysis of the *Nederlands Kanker Instituut* (*NKI*) data (1) consisting of 295 tumors, the *Breast Cancer Normal* (*BCN*) data (2) consisting of 13 normal breast tissue samples, and the validation data sets *Ullevål University Hospital* (*ULL*) (3) consisting of 46 tumors of ductal histological type that had been in the study for longer than 10 mo and *HERSCH* (4) consisting of 188 primary breast tumors.

1.1. Data preprocessing. Data were retrieved, missing values imputed, then data were collapsed by UniGene cluster ID build 219, and genes present in both the tumor cohort and the normal data set were retained.

For *NKI*, data consisted of 24,479 GeneBank accession IDs on 295 tumor samples, all of which had at least 70% data. Missing data were imputed using a *knn* algorithm (5) with $k = 10$. Data were also transformed from the original log_{10} values to log_2. Data were then collapsed (mean) by UniGene to the mean. The resulting data set consisted of 18,970 UniGene clusters.

For *BCN*, data from 13 normal tissue samples (nine nonneoplastic tissue from cancer patients, four reduction mammoplasty tissue) were retrieved with quality filters for each spot: (*i*) spot regression correlation $r > 0.6$, or (*ii*) channel 1 mean intensity/median background intensity >1.5, or (*iii*) channel 2 normalized (mean intensity/median background intensity) >1.5. Clones with 70% data were retained: 32,644 clone IDs. Missing data were imputed using a *knn* algorithm (5) with $k = 10$. Data were then collapsed by UniGene to 18,971 UniGene clusters. Of these, 12,237 UniGene IDs were in common with the *NKI* data set, and 17,441 were in common with the *ULL* data set (see below).

For *ULL*, data from 46 tumors were retrieved with quality filters for each spot: (*i*) spot regression correlation $r > 0.6$, or (*ii*) channel 1 mean intensity/median background intensity >1.5, or (*i*) channel 2 normalized (mean intensity/median background intensity) >1.5. Only clones with 70% good data were retained: 31,667 clone IDs. Missing data were imputed using a *knn* algorithm (5) with $k = 10$. Data were then combined with normal tissue data *BCN* and collapsed by UniGene to 17,441 UniGene clusters.

For *HERSCH*, data from 188 primary tumors were retrieved with quality filters for each spot: (*i*) spot regression correlation $r > 0.6$, or (*ii*) channel 1 mean intensity/median background intensity >1.5, or (*iii*) channel 2 normalized (mean intensity/median background intensity) >1.5. Only clones with 70% good data were retained: 32,644 clone IDs. Missing data were imputed using a *knn* algorithm (5) with $k = 10$. Data were then combined with normal tissue data *BCN* and collapsed by UniGene to 18,896 UniGene clusters.

1.2. Disease-Specific Genomic Analysis (DSGA). For *NKI* and *BCN*, data from tumors and normal tissue were combined along the common 12,237 UniGenes, and columns were normalized to have the magnitude of the mean vector magnitude of 13 normal tissue samples. The *Healthy State Model (HSM)* was constructed from normal tissue data $\{\vec{N}_1, \ldots, \vec{N}_{13}\}$ as follows: *FLAT* construction (2) is a method to de-sparse the data in high dimensions by substituting for each normal tissue vector $\vec{N}_i$, its fit $\widehat{N}_i$ to a linear model in the other normal tissue vectors:

$$\widehat{N}_i = \sum_{\substack{1 \le j \le 13 \\ j \ne i}} \beta_j \vec{N}_j.$$

This was shown to decrease noise in simulated data and help identify a good dimension reduction for *Principal Component Analysis* (*PCA*). We use a method described in ref. 2 to compute the Wold invariant (6) designed to measure a version of signal-to-noise ratio:

$$W(l) = \left(\frac{\lambda_l^2}{\lambda_{l+1}^2 + \ldots + \lambda_{13}^2} \right) \frac{(n-l-1)(13-l)}{(n+13-2l)}.$$

Fig. S1 plots $W(l)$ vs. the dimension l and shows a jump at $l = 10$, indicating that signal-to-noise ratio is higher at dimension 10, thereby justifying *PCA* dimension reduction of the *FLAT* normal data to 10. This produced the 10 dimensional *HSM*. Linear models are then used to compute the fitted tumor data matrix to the *HSM* (normal component *Nc.mat*) and the residuals (disease component *Dc.mat*). Along with tumor data, a leave-one-out procedure gives an estimate of the deviation of normal tissue data from the model of the healthy state *HSM*. Details of this procedure are found in ref. 2.

The validation data sets *ULL* and *HERSCH* were similarly transformed using the same normal data set *BCN*.

For gene thresholding, the 12,237 genes in the disease component matrix *Dc.mat* of tumors were reduced to 262 through the following method of testing for significance in deviation from the null hypothesis space. For each gene we computed the 5th and 95th percentiles of values in the disease components of the 295 tumors, and we recorded the larger of the two in absolute value and denoted the collection of these gene-by-gene deviations from normal by *MaxAbs595*. A histogram of these values is seen in Fig. S2. We then computed the 85th and 98th percentiles of *MaxAbs595* and denoted these as *relaxed* threshold and *stringent* threshold, respectively. A total of 1,836 genes exceeded the relaxed threshold, and 245 genes exceeded the stringent threshold. Genes were retained for further analysis if they passed the relaxed threshold and if they were also highly correlated ($r > 0.6$) to at least three genes that passed the stringent threshold. A total of 262 genes satisfied the condition. This method ensures that genes are retained in the analysis if they not only (*i*) deviate significantly from the null hypothesis space *HSM* but (*ii*) do so in groups of highly correlated genes. We denote the reduced matrix of disease component of *NKI* data: *nkiDc.mat*. The result of clustering the *nkiDc.mat* array and gene mean-centered can be found in supplementary folder Dataset S1: *nkiDc.AGmc.cdt*. It can be explored with TreeView (7), and all of the known clusters of genes can be observed, but because this is not germane to our present study we forgo any in-depth analysis of this clustering.

We did not follow the same thresholding procedure for the validation data sets *ULL* and *HERSCH*; rather, we found that of the 262 genes retained in the *NKI* data set, 255 genes were present in the *ULL* data and 221 in the *HERSCH* data.

1.3. Progression Analysis of Disease (PAD) on NKI. We give details of *PAD* on the reduced and *DSGA*-transformed *NKI* data matrix: *nkiDc.mat* of 295 tumors and 262 genes. First, this was combined with the leave-one-out matrix that estimates normal tissue: *bcnL1.mat*. The *Mapper* filter function was computed on each column vector, as explained in the main text (Eq. **2**). The image space was then fragmented into 15 intervals, with 80% overlap. Two outputs of mapper were obtained: the first, which included all of the bins, can be found in Fig. 3 (main text). The second provides the tighter streamlined subset of *Mapper* output, by excluding all bins with only one data point in them. The two outputs appear side by side in Fig. S3.

1.4. Comparison with clustering. Although *Mapper* incorporates clustering at the local level, the final output captures a wide

range of characteristics that are obfuscated by the standard methods of clustering the entire data. We provide in Fig. S4 an expanded version of the comparison presented in Fig. 4 (main text) between clustering and *PAD* analysis, complete with heat maps and progressions of bins. It is important to note that the comparison is performed after both *Mapper* and clustering were applied to exactly the same data matrix. Thus, whatever transformations one might perform on the data, for example *DSGA*, and however genes are thresholded to provide a reduced number of genes used in the analysis, the final step of clustering vs. *Mapper* generates very different outputs. Because both clustering and *Mapper* are methods that identify the shape of the data, this comparison highlights the fact that shape characteristics identified by *Mapper* can be lost by clustering. Note as well that clustering has scattered the *c-MYB*$^+$ tumor group among several clusters. This is a common problem known to clustering: data points will be segregated into separate clusters, and sometimes data points that are fairly close to one another will be torn apart and scattered into separate clusters. This is precisely what has happened with the *c-MYB*$^+$ group. Despite how similar the *c-MYB*$^+$ tumors are to one another, clustering has not kept them together.

2. Genes of Interest Analysis. We isolated a subgroup of tumors, *c-MYB*$^+$, through the use of *PAD*. We provide *Prediction Analysis of Microarrays* (*PAM*) (8) analysis outputs for comparison of this group to the normal tissue group. We provide *Significance of the Analysis of Microarrays* (*SAM*) (9) analysis for genes most significantly distinct between the *c-MYB*$^+$ group and the normal tissue group, as well as genes most significantly distinct between the *c-MYB*$^+$ group and the most adjacent tumors to the *c-MYB*$^+$ group in the *PAD* output, namely the tumors in the *ER*$^+$ arm that are not part of the *c-MYB*$^+$ group.

2.1. PAM. *PAM* finds a small set of *predictor genes* for distinguishing between two groups of tumors. Fig. S5 shows *PAM* output for comparing the *c-MYB*$^+$ group to the normal tissue group.

2.2. SAM. *SAM* finds a large number of *significant genes* that behave differently between two groups of tumors. Table S1 shows *SAM* output genes significantly distinct between the *c-MYB*$^+$ group and the rest of the *ER*$^+$ arm of *PAD* output. Table S2 shows *SAM* output genes significantly distinct between the *c-MYB*$^+$ group and the normal tissue group.

2.3. c-MYB signature. Genes that are believed to be downstream from *c-MYB* (10) were tested in the *c-MYB*$^+$ group vs. normal tissue using a one-sided Student *t* test. Results are listed in Table S3.

1. van de Vijver MJ, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 347:1999–2009.
2. Nicolau M, Tibshirani R, Børresen-Dale AL, Jeffrey SS (2007) Disease-specific genomic analysis: identifying the signature of pathologic biology. *Bioinformatics* 23:957–965.
3. Langerød A, et al. (2007) TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast Cancer Res* 9:R30.
4. Herschkowitz JI, He X, Fan C, Perou CM (2008) The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas. *Breast Cancer Res* 10:R75.
5. Troyanskaya O, et al. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* 17:520–525.
6. Wold S, et al. (1978) Cross-validatory estimation of the number of component in factor and principal components models. *Technometrics* 20:397–405.
7. Saldanha AJ (2004) Java Treeview—extensible visualization of microarray data. *Bioinformatics* 20:3246–3248.
8. Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* 99:6567–6572.
9. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98:5116–5121.
10. Ramsay RG, Gonda TJ (2008) MYB function in normal and cancer cells. *Nat Rev Cancer* 8:523–534.

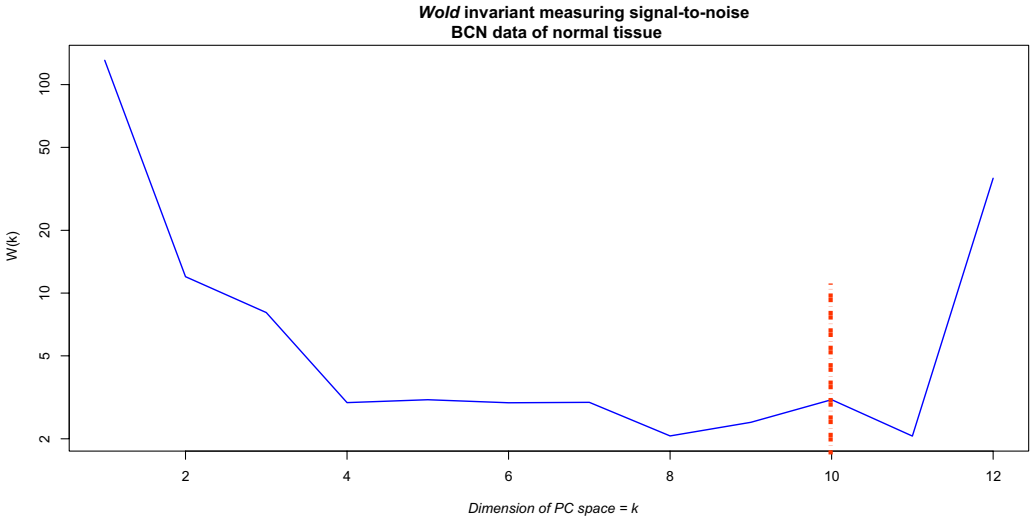

Fig. S1. The *wold* invariant is plotted as a function of the dimension reduction *K*. As the *wold* invariant is a measure of signal to noise, a local maximum in this plot indicates a good place to perform dimension reduction. In this case *K* = 10 is a good choice.

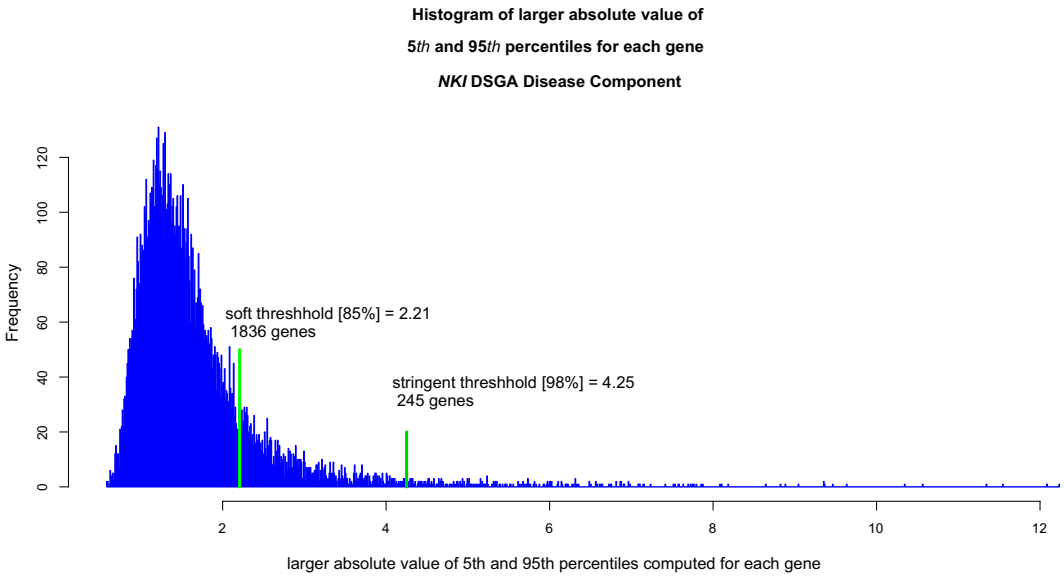

Fig. S2. For each gene, the 95th and 5th percentiles of expression levels in the disease component is computed. The larger of the two in absolute value denoted as Q_{gene} gives an estimate of the extent of deviation from normal for the gene. This deviation can be positive, indicating overexpression relative to normal levels, or negative, indicating underexpression relative to normal levels. The figure shows a histogram of the collection Q_{gene} of deviations from normal for the set of all genes. There are 1,836 genes for which this value exceeds the 85th percentile (*lax*-threshold genes) and 245 genes for which it exceeds the 95th percentile (*stringent*-threshold genes).

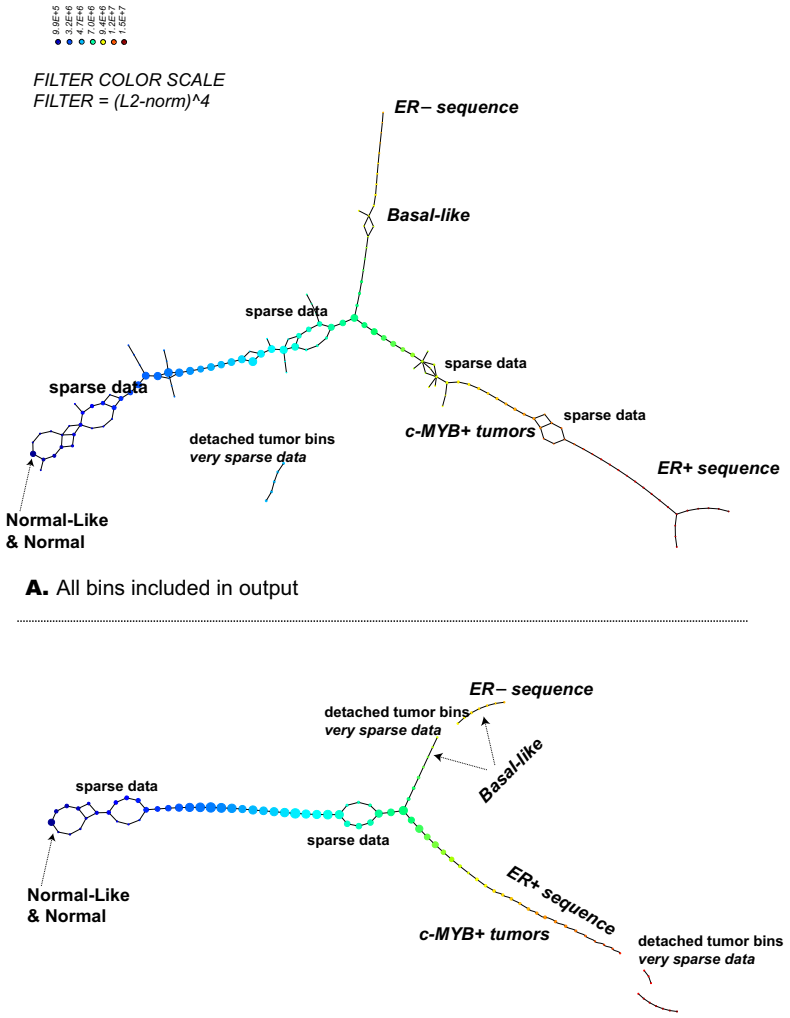

Fig. S3. (*A*) Complete output of the analysis. Each colored disk represents a bin containing several data points or patients. Thus, individual patients (data points) are not visible, and we only see bins containing collections of very similar points. This step provides a simplification of the original set of data points, because instead of showing a multitude of individual points, it shows a much smaller collection of bins, each bin containing a collection of very similar points. The size of bins relates to the number of data points contained in them. Thus, bins containing many data points appear as large discs, whereas bins with few points are drawn much smaller. When two bins have patients in common, an edge connects them. Thus, the bins provide a granularity to the overall set of data points, and the connections between bins, the edges that connect them, capture a rough shape of the data. Each data point has assigned to it a value of the *Mapper* xtit filter function, and the bins are colored by the average value of this function for the points in the bin. The legend with assigned colors is seen at the top. In this particular example, each data point is a tumor sample, its gene expression transformed by *DSGA* to measure deviation from the *HSM*. The filter function is the overall amount of deviation from the *HSM*. Thus, red bins contain patients whose overall molecular profiles deviate a lot from normal, whereas blue bins contain patients whose profile is very close to normal. Sometimes data points are quite sparse, and this sparseness is visible in the output as well. When the data points become some what sparse, we see the graph fan out in a slight web-like feature. When data becomes really sparse, pieces of the graph become completely disconnected. Areas of local data sparseness are indicated in the figure. Finally, some bins are very small, containing only a few data points. To get a more streamlined, simplified picture, we can choose to ignore bins that are very small. This is similar to ignoring outliers. (*B*) Same output, but with bins containing single points not shown. Notice that this more streamlined version loses some of the sparseness information (for example that the long ER^+ arm no longer exhibits the sparseness at the halfway point) and accentuates some sparseness areas by causing breaks in some places (for example the *Basal* arm now appears in two pieces).

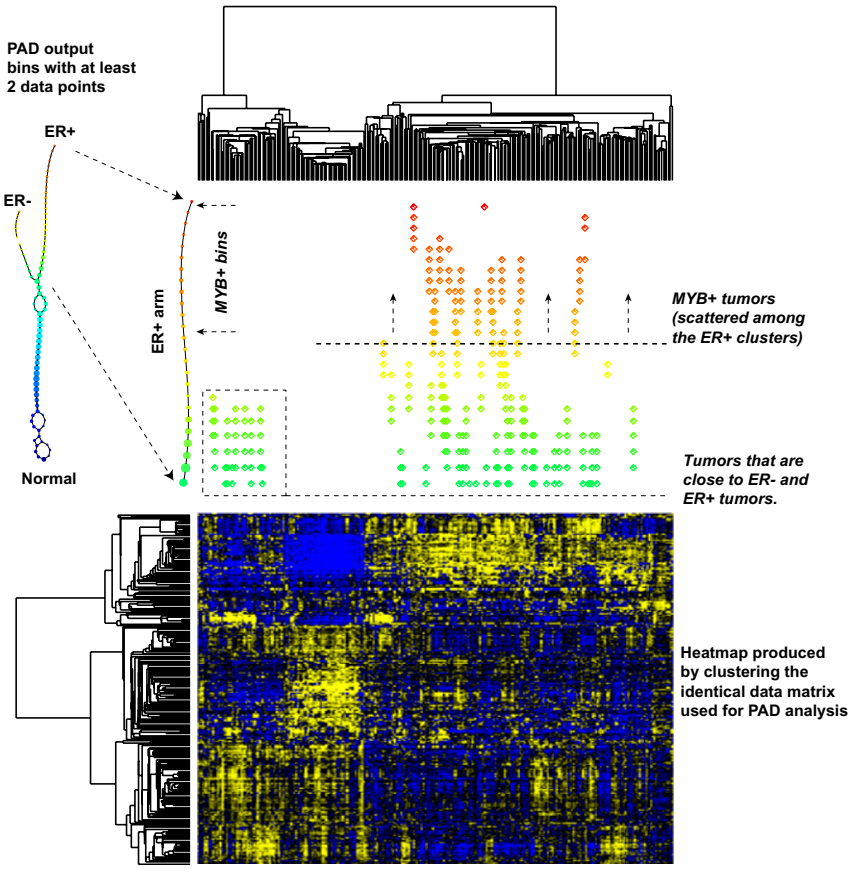

Fig. S4. Comparison between cluster analysis and *PAD*. Specifically, *PAD* consists of two major steps: the first step, *DSGA*, defines a transformation of the original data to detect extent of deviation from normal. It also provides a means to threshold genes so that only genes that deviate significantly from normal are retained. The second step, *Mapper*, involves detecting the shape of the data points in space. *Cluster analysis* is a different method to detect the shape of the data in space. This figure shows the difference between using cluster analysis as opposed to using *Mapper* to detect the shape of the same data matrix. We took the matrix whose columns are the disease components of the *DSGA*-transformed data, with only the 262 genes obtained by thresholding genes according to deviation from normal. This matrix was analyzed to detect its shape in space in two distinct ways: (*i*) it was clustered with associated heatmap and dendrograms shown, and (*ii*) it was processed with *Mapper*, with the output shown. The ER^+ arm is magnified, and the position of each tumor in each consecutive bin is shown relative to its placement in the clustering dendrogram. It is easily visible that whereas the *c-MYB*$^+$ group of tumors are close to one another in the *PAD* output, they are scattered throughout the ER^+ portion of the clustering diagrams. It is important to note that the same matrix was fed into the *Mapper* and the cluster analysis. The figure shows these outputs to be very distinct. The figure does not and cannot identify which output is identifying features that deserve to be noticed: cluster analysis did not identify the *c-MYB*$^+$ group, but it is not clear, simply on the basis of this figure, that the group is a real feature rather than an artifact of *Mapper*. It is through subsequent analysis methods that we see that the *c-MYB*$^+$ group is indeed both mathematically and biologically distinct. Thus, the *PAM* analysis shows the group to be mathematically coherent and easily distinct, and functional exploration of the genes identified by *SAM* analysis, along with survival analysis of the group, show it to be a biologically coherent and meaningful group of tumors. This figure shows that the shape analysis provided by clustering is different from that provided by *Mapper*.

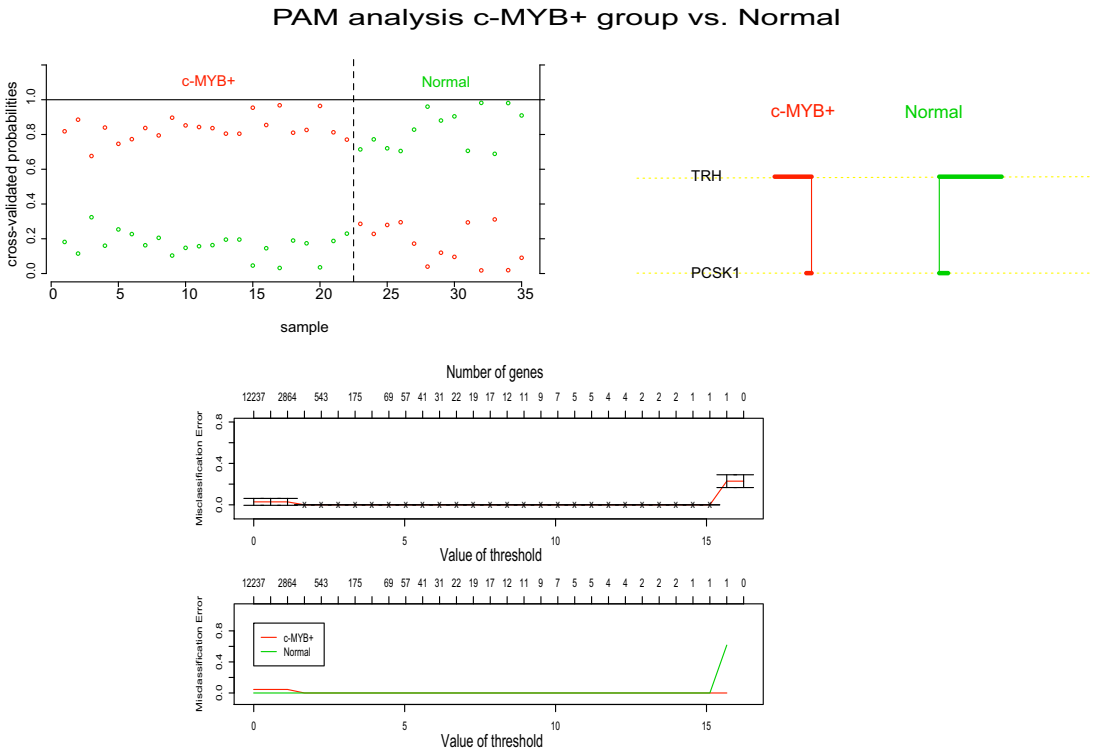

Fig. S5. Output of *PAM* analysis on the *c-MYB*$^{+}$ group vs. *Normal* data. Two genes provide class prediction with *error rate* = 0: *TRH*,*TSH-releasing hormone*, and *PCSK1*, *proprotein convertase subtilisin kexin type 1*. The centroids, cross-validation probabilities, and misclassification error plots are shown.

Table S1. Genes significantly up-regulated and down-regulated in *MYB+ vs.* the rest of *ER*$^+$ sequence

UniGene build 219	Gene symbol	*q* value	Gene information	MYB level vs. rest of *ER*$^+$ sequence
Hs.654446	*MYB*	0	MYB‖v-myb myeloblastosis viral oncogene homolog(avian)	Up
Hs.88417	*SUSD3*	0	SUSD3‖Sushi domain containing 3	Up
Hs.414028	*C9orf116*	0	C9orf116‖Chromosome 9 ORF 116	Up
Hs.532634	*IFI27*	5.15	IFI27‖IFN, α-inducible protein 27‖Hs.532634	Down
Hs.477891	*CPB1*	5.15	CPB1‖Carboxypeptidase B1 (tissue)‖Hs.477891	Down
Hs.49760	*ORC6L*	5.15	ORC6L‖Origin recognition complex, subunit 6 like (yeast)‖Hs.49760	Down
Hs.517307	*MX1*	5.15	MX1‖Myxovirus (influenza virus) resistance 1, IFN-inducible protein p78 (mouse)‖Hs.517307	Down
Hs.77367	*CXCL9*	5.15	CXCL9‖Chemokine (C-X-C motif) ligand 9‖Hs.77367	Down
Hs.501778	*TRIM22*	5.15	TRIM22‖Tripartite motif-containing 22‖Hs.501778	Down
Hs.521459	*ADAMDEC1*	5.15	ADAMDEC1‖ADAM-like, decysin 1‖Hs.521459	Down
Hs.458485	*ISG15*	5.15	ISG15‖ISG15 ubiquitin-like modifier‖Hs.458485	Down
Hs.109225	*VCAM1*	5.15	VCAM1‖Vascular cell adhesion molecule 1‖Hs.109225	Down
Hs.17518	*RSAD2*	5.15	RSAD2‖Radical S-adenosyl methionine domain containing 2‖Hs.17518	Down
Hs.7155	*CMPK2*	5.15	CMPK2‖Cytidine monophosphate (UMP-CMP) kinase 2, mitochondrial‖Hs.7155	Down
Hs.20315	*IFIT1*	6.51	IFIT1‖IFN-induced protein with tetratricopeptide repeats 1‖Hs.20315	Down
Hs.306777	*GSDMB*	6.51	GSDMB‖Gasdermin B‖Hs.306777	Down
Hs.715518	*STAT1*	6. 51	STAT1‖Signal transducer and activator of transcription 1, 91kDa‖Hs.715518	Down
Hs.709313	*B2M*	6. 51	B2M‖Beta-2-microglobulin‖Hs.709313	Down
Hs.584823	*PLA2G7*	6. 51	PLA2G7‖Phospholipase A2, groUp VII (platelet-activating factor acetylhydrolase, plasma)‖ Hs.584823	Down
Hs.181244	*HLA-A*	6. 51	HLA-A‖Major histocompatibility complex, class I, A‖Hs.181244	Down
Hs.473341	*SAMSN1*	6. 51	SAMSN1‖SAM domain, SH3 domain and nuclear localization signals 1‖Hs.473341	Down
Hs.523847	*IFI6*	6. 51	IFI6‖IFN, α-inducible protein 6‖Hs.523847	Down
Hs.504641	*CD163*	6. 51	CD163‖CD163 molecule‖Hs.504641	Down
Hs.250615	*CYP2A6*	15.08	CYP2A6‖Cytochrome P450, family 2, subfamily A, polypeptide 6‖Hs.250615	Down
Hs.655652	*LILRB2*	15. 08	LILRB2‖Leukocyte Ig-like receptor, subfamily B (with TM and ITIM domains), member 2‖Hs.655652	Down
Hs.459265	*ISG20*	15. 08	ISG20‖IFN stimulated exonuclease gene 20kDa‖Hs.459265	Down
Hs.926	*MX2*	15. 08	MX2‖Myxovirus (influenza virus) resistance 2 (mouse)‖Hs.926	Down
Hs.525157	*TNFSF13B*	15. 08	TNFSF13B‖Tumor necrosis factor (ligand) sUperfamily, member 13b‖Hs.525157	Down
Hs.86859	*GRB7*	15. 08	GRB7‖Growth factor receptor-bound protein 7‖Hs.86859	Down
Hs.352018	*TAP1*	15. 08	TAP1‖Transporter 1, ATP-binding cassette, subfamily B (MDR/TAP)‖Hs.352018	Down
Hs.32763	*GRIA2*	15. 08	GRIA2‖Glutamate receptor, ionotropic, AMPA 2‖Hs.32763	Down
Hs.654585	*PSMB9*	15. 08	PSMB9‖Proteasome (prosome, macropain) subunit, β type, 9 (large multifunctional peptidase 2)‖Hs.654585	Down
Hs.718626	*KIF20A*	15. 08	KIF20A‖Kinesin family member 20A‖Hs.718626	Down
Hs.474787	*IL2RB*	15. 08	IL2RB‖Interleukin 2 receptor, β‖Hs.474787	Down
Hs.650174	*HLA-E*	15. 08	HLA-E‖Major histocompatibility complex, class I, E‖Hs.650174	Down

Table S1. Cont.

UniGene build 219	Gene symbol	*q* value	Gene information	MYB level vs. rest of ER^+ sequence
Hs.143961	*CCL18*	15. 08	CCL18‖Chemokine (C-C motif) ligand 18 (pulmonary and activation-regulated)‖ Hs.143961	Down
Hs.81337	*LGALS9*	15. 08	LGALS9‖Lectin, galactoside-binding, soluble, 9‖Hs.81337	Down
Hs.474217	*CDC45L*	15. 08	CDC45L‖CDC45 cell division cycle 45-like (*S. cerevisiae*)‖Hs.474217	Down
Hs.301921	*CCR1*	15. 08	CCR1‖Chemokine (C-C motif) receptor 1‖Hs.301921	Down
Hs.16362	*P2RY6*	15. 08	P2RY6‖Pyrimidinergic receptor P2Y, G protein coUpled, 6‖Hs.16362	Down
Hs.419259	*REC8*	15. 08	REC8‖REC8 homolog (yeast)‖Hs.419259	Down
Hs.591742	*IL7R*	15. 08	IL7R‖Interleukin 7 receptor‖Hs.591742	Down
Hs.647962	*ZIC1*	18.67	ZIC1‖Zic family member 1 (odd-paired homolog, *Drosophila*)‖Hs.647962	Down
Hs.43388	*RTP4*	18. 67	RTP4‖Receptor (chemosensory) transporter protein 4‖Hs.43388	Down
Hs.376208	*LTB*	18. 67	LTB‖Lymphotoxin β (TNF sUperfamily, member 3)‖Hs.376208	Down
Hs.14623	*IFI30*	18. 67	IFI30‖IFN, γ-inducible protein 30‖Hs.14623	Down
Hs.660866	*CTSL2*	18. 67	CTSL2‖Cathepsin L2‖Hs.660866	Down
Hs.278658	*KRT86*	18. 67	KRT86‖Keratin 86‖Hs.278658	Down
Hs.1051	*GZMB*	18. 67	GZMB‖Granzyme B (granzyme 2, cytotoxic T lymphocyte-associated serine esterase 1)‖ Hs.1051	Down
Hs.1594	*CENPA*	18. 67	CENPA‖Centromere protein A‖Hs.1594	Down
Hs.161985	*TMPRSS4*	18. 67	TMPRSS4‖Transmembrane protease, serine 4‖Hs.161985	Down
Hs.153752	*CDC25B*	18. 67	CDC25B‖Cell division cycle 25 homolog B (S. pombe)‖Hs.153752	Down
Hs.446352	*ERBB2*	18. 67	ERBB2‖V-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)‖ Hs.446352	Down
Hs.497599	*WARS*	18. 67	WARS‖Tryptophanyl-tRNA synthetase‖Hs.497599	Down
Hs.182231	*TRH*	18. 67	TRH‖TSH-releasing hormone‖Hs.182231	Down
Hs.521903	*LY6E*	20.44	LY6E‖Lymphocyte antigen 6 complex, locus E‖Hs.521903	Down
Hs.370036	*CCR7*	20. 44	CCR7‖Chemokine (C-C motif) receptor 7‖Hs.370036	Down

Table S2. Genes significantly up-regulated and down-regulated in *MYB+ vs. Normal* tissue

UniGene build 219	Gene symbol	*q* value	Gene information	*MYB* level vs. normal
Hs.414028	*C9orf116*	0	C9orf116‖Chromosome 9 ORF 116‖Hs.414028	Up
Hs.406050	*DNALI1*	0	DNALI1‖Dynein, axonemal, light intermediate chain 1‖Hs.406050	Up
Hs.163484	*FOXA1*	0	FOXA1‖Forkhead box A1‖Hs.163484	Up
Hs.76704	*[Hs.76704]*	0	NA‖Transcribed locus‖Hs.76704	Up
Hs.654446	*MYB*	0	MYB‖V-myb myeloblastosis viral oncogene homolog (avian)‖Hs.654446	Up
Hs.88417	*SUSD3*	0	SUSD3‖Sushi domain containing 3‖Hs.88417	Up
Hs.494496	*FBP1*	0	FBP1‖Fructose-1,6-bisphosphatase 1‖Hs.494496	Up
Hs.448520	*SLC7A2*	0	SLC7A2‖Solute carrier family 7 (cationic amino acid transporter, y+ system), member 2‖Hs.448520	Up
Hs.534847	*C4A*	0	C4A‖Complement component 4A (Rodgers blood groUp)‖Hs.534847	Up
Hs.496240	*AR*	0	AR‖Androgen receptor‖Hs.496240	Up
Hs.631650	*GLT8D2*	0	GLT8D2‖Glycosyltransferase 8 domain containing 2‖Hs.631650	Up
Hs.91109	*PRR15*	0	PRR15‖Proline rich 15‖Hs.91109	Up
Hs.387057	*THSD4*	0	THSD4‖Thrombospondin, type I, domain containing 4‖Hs.387057	Up
Hs.98265	*ST6GAL2*	0	ST6GAL2‖ST6 β-galactosamide α-2,6-sialyltranferase 2‖Hs.98265	Up
Hs.208124	*ESR1*	0	ESR1‖Estrogen receptor 1‖Hs.208124	Up
Hs.111779	*SPARC*	0	SPARC‖Secreted protein, acidic, cysteine-rich (osteonectin)‖Hs.111779	Up
Hs.480819	*TBC1D9*	0	TBC1D9‖TBC1 domain family, member 9 (with GRAM domain)‖Hs.480819	Up
Hs.437638	*XBP1*	0	XBP1‖X-box binding protein 1‖Hs.437638	Up
Hs.444414	*AFF3*	0	AFF3‖AF4/FMR2 family, member 3‖Hs.444414	Up
Hs.524134	*GATA3*	0	GATA3‖GATA binding protein 3‖Hs.524134	Up
Hs.467733	*GREB1*	0	GREB1‖GREB1 protein‖Hs.467733	Up
Hs.458573	*PDGFRL*	0	PDGFRL‖Platelet-derived growth factor receptor-like‖Hs.458573	Up
Hs.210995	*CA12*	0	CA12‖Carbonic anhydrase XII‖Hs.210995	Up
Hs.523468	*SCUBE2*	0	SCUBE2‖Signal peptide, CUB domain, EGF-like 2‖Hs.523468	Up
Hs.654370	*FAP*	0	FAP‖Fibroblast activation protein, α‖Hs.654370	Up
Hs.489142	*COL1A2*	0	COL1A2‖Collagen, type I, α 2‖Hs.489142	Up
Hs.416108	*CRKRS*	0	CRKRS‖Cdc2-related kinase, arginine/serine-rich‖Hs.416108	Up
Hs.371147	*THBS2*	0	THBS2‖Thrombospondin 2‖Hs.371147	Up
Hs.519601	*ID4*	0	ID4‖Inhibitor of DNA binding 4, dominant negative helix–loop–helix protein‖Hs.519601	Up
Hs.100686	*AGR3*	0	AGR3‖Anterior gradient homolog 3 (*Xenopus laevis*)‖Hs.100686	Up
Hs.435655	*ASPN*	0	ASPN‖Asporin‖Hs.435655	Up
Hs.425777	*UBE2L6*	0	UBE2L6‖Ubiquitin-conjugating enzyme E2L 6‖Hs.425777	Up
Hs.659093	*[Hs.659093]*	0	NA‖Transcribed locus‖Hs.659093	Up
Hs.93764	*CPA4*	0	CPA4‖Carboxypeptidase A4‖Hs.93764	Up
Hs.719277	*SLC39A6*	0	SLC39A6‖Solute carrier family 39 (zinc transporter), member 6‖Hs.719277	Up
Hs.604376	*[Hs.604376]*	0	NA‖Transcribed locus‖Hs.604376	Up
Hs.95612	*DSC2*	0	DSC2‖Desmocollin 2‖Hs.95612	Up
Hs.8059	*SYT4*	0	SYT4‖Synaptotagmin IV‖Hs.8059	Up
Hs.1925	*DSG3*	0	DSG3‖Desmoglein 3 (pemphigus vulgaris antigen)‖Hs.1925	Up
Hs.8786	*CHST2*	0	CHST2‖Carbohydrate (*N*-acetylglucosamine-6-O) sulfotransferase 2‖Hs.8786	Up
Hs.24950	*RGS5*	0	RGS5‖Regulator of G protein signaling 5‖Hs.24950	Up
Hs.19492	*PCDH8*	0	PCDH8‖Protocadherin 8‖Hs.19492	Up
Hs.520339	*COL10A1*	0	COL10A1‖Collagen, type X, α 1‖Hs.520339	Up
Hs.5210	*GMFG*	0.46	GMFG‖Glia maturation factor, γ‖Hs.5210	Up
Hs.497636	*LAMB3*	0.46	LAMB3‖Laminin, β 3‖Hs.497636	Up
Hs.6360	*TMCC2*	0. 46	TMCC2‖Transmembrane and coiled-coil domain family 2‖Hs.6360	Up
Hs.34526	*CXCR6*	0. 46	CXCR6‖Chemokine (C-X-C motif) receptor 6‖Hs.34526	Up
Hs.504115	*TRIM29*	0.85	TRIM29‖Tripartite motif-containing 29‖Hs.504115	Up

Table S2. Cont.

UniGene build 219	Gene symbol	*q* value	Gene information	*MYB* level vs. normal
Hs.1787	*PLP1*	0. 85	PLP1‖Proteolipid protein 1‖Hs.1787	Up
Hs.523500	*CD2*	0. 85	CD2‖CD2 molecule‖Hs.523500	Up
Hs.131431	*EIF2AK2*	0. 85	EIF2AK2‖Eukaryotic translation initiation factor 2-α kinase 2‖Hs.131431	Up
Hs.136348	*POSTN*	0. 85	POSTN‖Periostin, osteoblast specific factor‖Hs.136348	Up
Hs.193235	*CPLX2*	0. 85	CPLX2‖Complexin 2‖Hs.193235	Up
Hs.438	*MEOX1*	1.94	MEOX1‖Mesenchyme homeobox 1‖Hs.438	Up
Hs.405614	*CTHRC1*	1.94	CTHRC1‖Collagen triple helix repeat containing 1‖Hs.405614	Up
Hs.182231	*TRH*	0	TRH‖TSH-releasing hormone‖Hs.182231	Down
Hs.477891	*CPB1*	0	CPB1‖Carboxypeptidase B1 (tissue)‖Hs.477891	Down
Hs.78977	*PCSK1*	0	PCSK1‖Proprotein convertase subtilisin/kexin type 1‖Hs.78977	Down
Hs.250615	*CYP2A6*	0	CYP2A6‖Cytochrome P450, family 2, subfamily A, polypeptide 6‖Hs.250615	Down
Hs.26770	*FABP7*	0	FABP7‖Fatty acid binding protein 7, brain‖Hs.26770	Down
Hs.516874	*CHGB*	0	CHGB‖Chromogranin B (secretogranin 1)‖Hs.516874	Down
Hs.150793	*CHGA*	0	CHGA‖Chromogranin A (parathyroid secretory protein 1)‖ Hs.150793	Down
Hs.77367	*CXCL9*	0	CXCL9‖Chemokine (C-X-C motif) ligand 9‖Hs.77367	Down
Hs.496843	*VGLL1*	0	VGLL1‖Vestigial like 1 (*Drosophila*)‖Hs.496843	Down
Hs.268728	*TTYH1*	0	TTYH1‖Tweety homolog 1 (*Drosophila*)‖Hs.268728	Down
Hs.416073	*S100A8*	0	S100A8‖S100 calcium binding protein A8‖Hs.416073	Down
Hs.473341	*SAMSN1*	0	SAMSN1‖SAM domain, SH3 domain and nuclear localization signals 1‖Hs.473341	Down
Hs.517307	*MX1*	0	MX1‖Myxovirus (influenza virus) resistance 1, IFN-inducible protein p78 (mouse)‖Hs.517307	Down
Hs.532634	*IFI27*	0	IFI27‖IFN, α-inducible protein 27‖Hs.532634	Down
Hs.143961	*CCL18*	0	CCL18‖Chemokine (C-C motif) ligand 18 (pulmonary and activation-regulated)‖Hs.143961	Down
Hs.458485	*ISG15*	0	ISG15‖ISG15 ubiquitin-like modifier‖Hs.458485	Down
Hs.192859	*PCDH10*	0	PCDH10‖Protocadherin 10‖Hs.192859	Down
Hs.419259	*REC8*	0	REC8‖REC8 homolog (yeast)‖Hs.419259	Down
Hs.470654	*CDCA7*	0	CDCA7‖Cell division cycle associated 7‖Hs.470654	Down
Hs.32763	*GRIA2*	0	GRIA2‖Glutamate receptor, ionotropic, AMPA 2‖Hs.32763	Down
Hs.415762	*LY6D*	0	LY6D‖Lymphocyte antigen 6 complex, locus D‖Hs.415762	Down
Hs.119689	*CGA*	0	CGA‖Glycoprotein hormones, α polypeptide‖Hs.119689	Down
Hs.278658	*KRT86*	0	KRT86‖Keratin 86‖Hs.278658	Down
Hs.17518	*RSAD2*	0	RSAD2‖Radical S-adenosyl methionine domain containing 2‖Hs.17518	Down
Hs.7155	*CMPK2*	0	CMPK2‖Cytidine monophosphate (UMP-CMP) kinase 2, mitochondrial‖Hs.7155	Down
Hs.20315	*IFIT1*	0	IFIT1‖IFN-induced protein with tetratricopeptide repeats 1‖Hs.20315	Down
Hs.418167	*ALB*	0	ALB‖Albumin‖Hs.418167	Down
Hs.372578	*FAM65C*	0	FAM65C‖Family with sequence similarity 65, member C‖Hs.372578	Down
Hs.26225	*GABRP*	0	GABRP‖Gamma-aminobutyric acid (GABA) A receptor, pi‖Hs.26225	Down
Hs.151254	*KLK7*	0	KLK7‖Kallikrein-related peptidase 7‖Hs.151254	Down
Hs.161985	*TMPRSS4*	0	TMPRSS4‖Transmembrane protease, serine 4‖Hs.161985	Down
Hs.376208	*LTB*	0	LTB‖Lymphotoxin β (TNF sUperfamily, member 3)‖Hs.376208	Down
Hs.414629	*CCL13*	0	CCL13‖Chemokine (C-C motif) ligand 13‖Hs.414629	Down
Hs.521459	*ADAMDEC1*	0	ADAMDEC1‖ADAM-like, decysin 1‖Hs.521459	Down
Hs.79361	*KLK6*	0	KLK6‖Kallikrein-related peptidase 6‖Hs.79361	Down
Hs.112405	*S100A9*	0	S100A9‖S100 calcium binding protein A9‖Hs.112405	Down
Hs.49760	*ORC6L*	0	ORC6L‖Origin recognition complex, subunit 6 like (yeast)‖ Hs.49760	Down
Hs.647962	*ZIC1*	0	ZIC1‖Zic family member 1 (odd-paired homolog, *Drosophila*)‖ Hs.647962	Down
Hs.30743	*PRAME*	0	PRAME‖Preferentially expressed antigen in melanoma‖Hs.30743	Down

Table S2. Cont.

UniGene build 219	Gene symbol	*q* value	Gene information	*MYB* level vs. normal
Hs.2256	*MMP7*	0	MMP7‖Matrix metallopeptidase 7 (matrilysin, uterine)‖ Hs.2256	Down
Hs.523847	*IFI6*	0	IFI6‖IFN, α-inducible protein 6‖Hs.523847	Down
Hs.75285	*ITIH2*	0	ITIH2‖Interα (globulin) inhibitor H2‖Hs.75285	Down
Hs.654550	*KRT13*	0	KRT13‖Keratin 13‖Hs.654550	Down
Hs.532635	*SERPINA6*	0	SERPINA6‖Serpin peptidase inhibitor, clade A (α-1 antiproteinase, antitrypsin), member 6‖Hs.532635	Down
Hs.86859	*GRB7*	0.46	GRB7‖Growth factor receptor-bound protein 7‖Hs.86859	Down
Hs.514527	*BIRC5*	0.85	BIRC5‖Baculoviral IAP repeat-containing 5‖Hs.514527	Down
Hs.370036	*CCR7*	0.85	CCR7‖Chemokine (C-C motif) receptor 7‖Hs.370036	Down
Hs.22905	*RP13-102H20.1*	0.85	RP13-102H20.1‖Hypothetical protein FLJ30058‖Hs.22905	Down
Hs.660866	*CTSL2*	0.85	CTSL2‖Cathepsin L2‖Hs.660866	Down
Hs.315	*MUC2*	1.94	MUC2‖Mucin 2, oligomeric mucus/gel-forming‖Hs.315	Down
Hs.63287	*CA9*	1.94	CA9‖Carbonic anhydrase IX‖Hs.63287	Down
Hs.109225	*VCAM1*	1.94	VCAM1‖Vascular cell adhesion molecule 1‖Hs.109225	Down
Hs.501778	*TRIM22*	1.94	TRIM22‖Tripartite motif-containing 22‖Hs.501778	Down

Table S3. Testing the *MYB* signature genes

Gene symbol	Gene name	UniGene build 219	pval MYB$^+$group_UP Normal_LO
MYC	V-myc myelocytomatosis viral oncogene homolog	Hs.202453	0.24
MYB	V-myb myeloblastosis viral oncogene homolog	Hs.654446	4.70E-05
ADA	Adenosine deaminase	Hs.654536	1.60E-10
CDK1	Cyclin-dependent kinase 1	Hs.334562	0.00019
POLD1	Polymerase (DNA directed), δ 1	Hs.279413	2.60E-11
PRTN3	Myeloblastin‖proteinase 3	Hs.928	0.00014
CD4	T-cell surface antigen T4/Leu-3	Hs.631659	1
VEGF	Vascular endothelial growth factor A	Hs.73793	0.62
BCL2	B-cell CLL/lymphoma 2	Hs.150749	0.97
KIT	Proto-oncogene c-Kit‖mast/stem cell growth factor receptor	Hs.479754	1
CD34	Hematopoietic progenitor cell antigen CD34	Hs.374990	1
GATA3	Transacting T-cell-specific transcription factor GATA-3	Hs.524134	0.00048
MPO	Myeloperoxidase	Hs.458272	0.012
HSP70	HSPA4‖heat shorck 70kDa protein 4	Hs.90093	0.00064
H2A.Z	H2AZ histone	Hs.119192	0.00028
Adora2B	Adenosine receptor 2B – chicken	Hs.167046	0.01
Mcm4	CDC21; CDC54; MGC33310; P1-CDC21; hCdc21	Hs.460184	2.20E-05
GAS41	YEATS4;Yeats domeain containing 4	Hs.4029	0.00078
NMU	Neuromedin U	Hs.418367	0.38
CCNE1	Cyclin E1	Hs.244723	0.00049
CCNB1	cyclin B1	Hs.23960	0.021
CA1	Carbonic anhydrase 1	Hs.23118	0.00037
PDCD4	Programmed cell death 4(neoplastic transformation inhibitor)	Hs.711490	0.019
COL1A1	Collagen type I, α 1	Hs.172928	1
COL1A2	Collagen type I, α 2	Hs.489142	1
CD13 ‖ ANPEP	Ananyl (membrane) animopeptidase	Hs.1239	0.96
GBX2	Gastrulation brain homeobox 2	Hs.184945	0.61
Actn1	Actinin, α 1	Hs.509765	0.9
Birc3	Baculoviral IAP repeat-containing 3	Hs.127799	1
Casp6	caspase 6, apoptosis-related cysteine peptidase	Hs.654616	3.60E-06
Cbx4	Chromobox homolog 4 (Pc class homolog, *Drosophila*)	Hs.714363	0.00073
Copa	coatomer protein complex, subunit α	Hs.162121	0.00017
Hspa8	Heat shock 70kDa protein 8	Hs.702021	1.80E-05
Iqgap1	IQ motif containing GTPase activating protein 1	Hs.430551	0.0047
Lca ‖ CLTA	Clathrin, light chain A	Hs.522114	9.00E-07
Mad1l1	MAD1 mitotic arrest deficient-like 1 (yeast)	Hs.654838	7.30E-10
Ppp3ca	Protein phosphatase 3, catalytic subunit, α isozyme	Hs.435512	0.42
SLC1A5	Solute carrier family 1 (neutral amino acid transporter), member 5	Hs.631582	0.032
Cox-2 ‖ PTGS2	Prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	Hs.196384	0.28
TCRd ‖ TRD@	T-cell receptor δ locus	Hs.74647	0.79
FABP5	Fatty acid binding protein 5 (psoriasis-associated)	Hs.408061	1
DHRS2	Dehydrogenase/reductase (SDR family) member 2	Hs.272499	0.19
TGFB1	Transforming growth factor, β 1	Hs.645227	0.63
CTNNAL1	Catenin (cadherin-associated protein), α-like 1	Hs.58488	0.00059