RNA-Seq For Identifying Gene Expression Changes Associated with Relapse in Acute Lymphoblastic Leukemia (ALL)

Dustin Holloway[1,2], Steve McLaughlin[2], Christopher Clouser[2], Elizabeth Levandowsky[2], Tamara Gilbert[2], Tristen Ross[2], Jessica Spangler[2], Letha Phillips[3], Sue Heatley[3], Lei Wei[3], Jinghui Zhang[3], Clarence Lee[2], Heather Peckham[2], Charles Mullighan[3,4]
[1] Presenting author [2] Life Technologies, Beverly, MA, USA [3] St. Jude Children's Research Hospital, Memphis, TN, USA [4] Principle investigator

ABSTRACT

In this study we have performed RNA-Seq using the SOLiD™ 3 system on samples from five children with acute lymphoblastic leukemia (ALL) with paired samples taken at the time of initial presentation with cancer ("initial" or "I") and at the time of relapse ("relapse" or "R"). Raw data was processed using the analysis pipelines in Bioscope and tertiary analysis was performed using several published statistical methods. The methods identified gene sets that can select cancer relapse samples with almost 98% accuracy, and have further identified genes that were found to be correlated with tumor relapse and poor prognosis in other studies. The analysis have also identified nucleotide variants unique to the samples at the time of relapse. We further note that processing of the data using a Bayesian method such as Bayseq appears to alleviate sample and run specific variation in the data. RNA-Seq is a powerful tool for exploring the biology of cancer, and these results point the way to important follow up studies that may further elucidate the causes of cancer relapse.

INTRODUCTION

Although the rate of cure for ALL has been improving, a substantial minority of those afflicted with the disease still have poor outcome. Up to one quarter of children with ALL fail therapy and relapse. The biologic determinants of disease relapse are poorly understood. Previous studies have identified structural genetic alterations acquired at the time relapse, and differences in gene expression patterns between matched samples obtained at the initial cancer and relapse conditions. However, a genome-wide analysis of sequence variation in relapsed ALL has not been performed. Moreover, next-generation sequencing approaches offer the opportunity to profile changes in gene expression patterns in great detail.

In this study we have performed RNA-Seq on samples from five children with ALL, with paired samples taken at the time of initial cancer presentation and at the time of relapse. 40 barcode samples (4 technical replicates per sample) were sequenced on four SOLiD system runs. Samples were analyzed using several statistical methods including principle component analysis (PCA), significance analysis of microarrays (SAM), weighted-voting (WV), BaySeq, and Support Vector Machines (SVM). These methods produce statistically significant gene sets that identify relapse samples with almost 98% accuracy, and have further identified genes that were found to be correlated with tumor invasiveness and poor prognosis in other studies. Further analysis of SNP results have also identified sequence alterations unique to the relapse state in some individuals, providing a pool of variations which may include potential drivers of metastasis.

MATERIALS AND METHODS

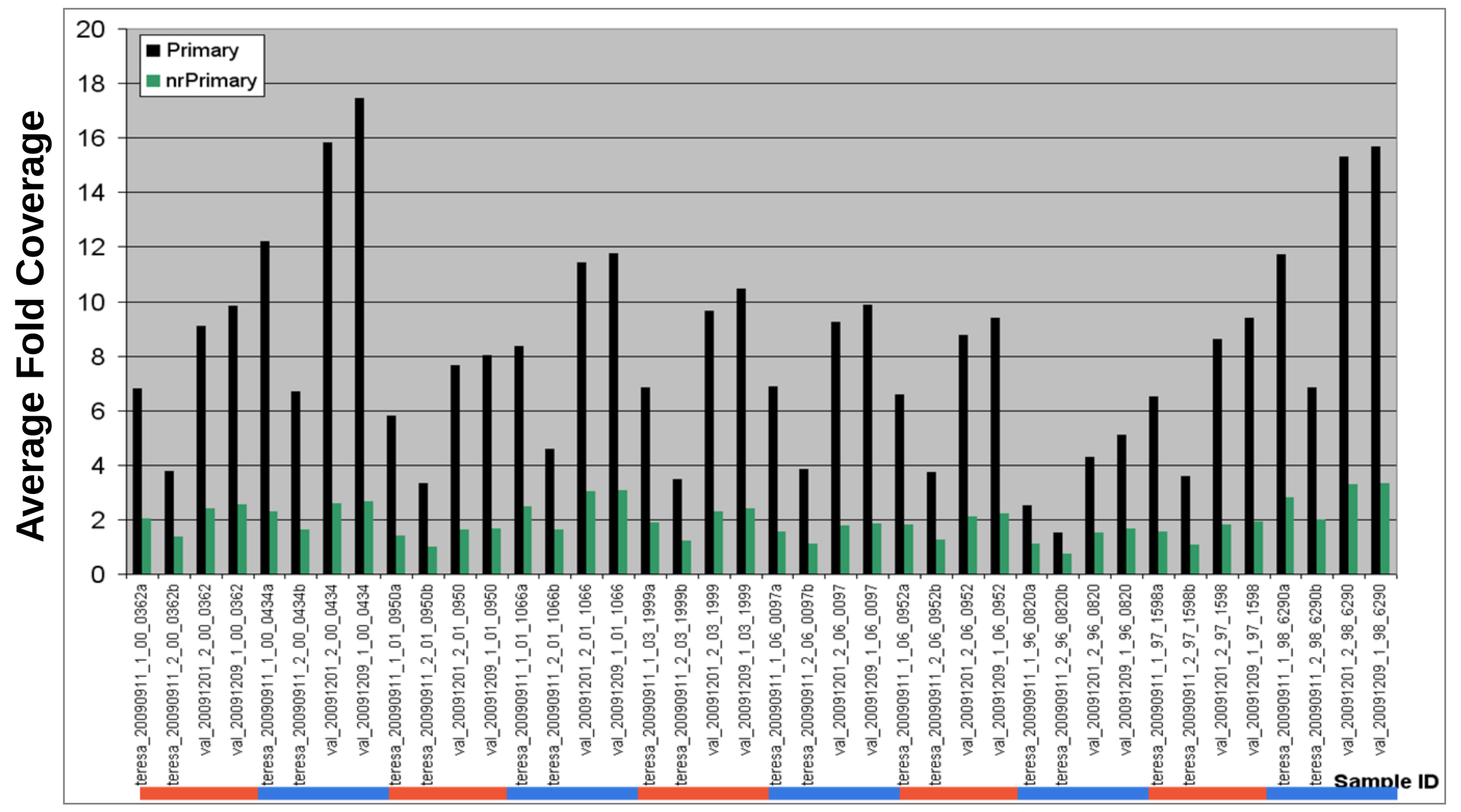
RNA samples were extracted using the TriZOL method from bone marrow samples obtained at diagnosis and relapse from five children with B-progenitor ALL (10 samples in total). Samples were prepared with the SOLiD Total RNA-Seq kit [1] and barcoded for multiplex fragment sequencing. Each of the 10 samples were barcoded and sequenced on 4 slides on SOLiD 3 instruments (10 samples x 4 slides = 40 barcode sequencing samples) as 50-mer fragments.

SOLiD system results were processed using the Bioscope™ v1.2.1 Whole Transcriptome pipeline. Analysis consisted of mapping the short reads to the genome and to annotated exon regions. Read counting was performed with custom scripts and the HTSeq [2] module in R. Principle component analysis, hierarchical clustering [3,4], and weighted-voting [5] methods were performed in GenePattern [6]. SAM [7,8,9,10] was performed in the TM4 Analysis Suite [11,12]. SVM-RFE [13,14] was performed in the SPIDER [15] machine learning toolbox in MATLAB® software [16]. BaySe q[17] analysis was conducted in the R language.

Whole Transcriptome Analysis is available in LifeScope™

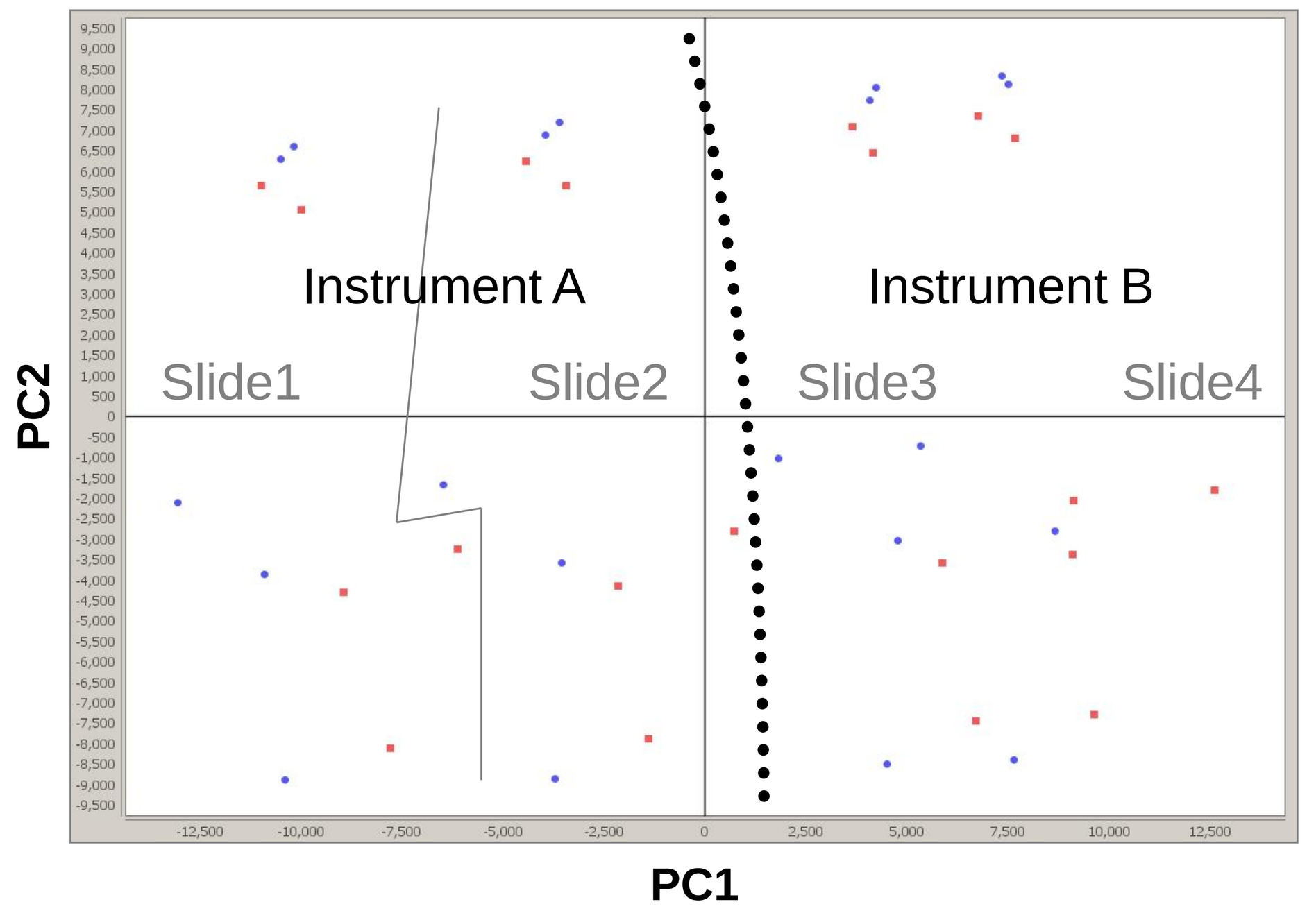
RESULTS

Figure 1. Average Exon Coverage per Sample: Aggregate ~6X-8X non-redundant Exome coverage, and ~15X primary Exome coverage per child.



Average fold coverage of each sequenced sample. 10 barcode samples were included in each of 4 SOLiD 3 system slides, yielding 40 total samples. There are 5 individuals each represented by 4 samples at the initial stage and 4 at the relapse stage. By aggregating samples we see ~6X-8X non-redundant Exome coverage, and ~15X primary Exome coverage per child.
Black bars show primary coverage (top-hit), green bars show non-redundant primary (top-hit, filtered for potential duplicate reads coverage.
Red bars: initial (I)
Blue bars: relapse (R)

Figure 2. PCA on samples with standard gene-wise normalization



Principle Component Analysis on gene RPKM values after standard gene-wise normalization across samples. Each point on the graph represents a single sample (8 samples per individual: 4 initial, 4 relapse).
Appropriate normalization of gene RPKM values removes variation amongst instrument runs

Blue: relapse
Red: initial

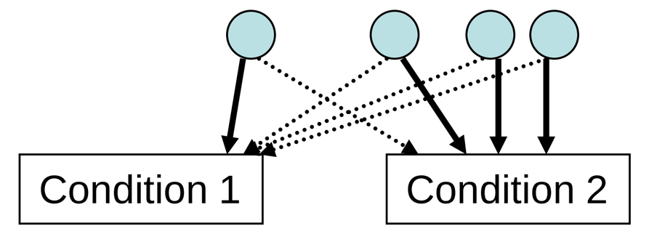
Figure 3. Relapse classifier genes can be selected with a variety of methods.

BaySeq
51 genes at 95% significance

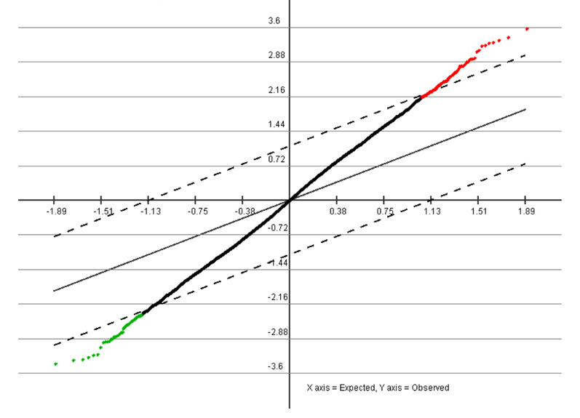
$$\mathbb{P}(D_c \mid M) \approx \prod_q \frac{1}{|\Theta_q|} \sum_{\Theta_q} \left[\prod_{\{i: A_i \in E_q\}} \frac{\Gamma(u_{ic}+\phi_q^{-1})}{\Gamma(\phi_q^{-1}) u_{ic}!} \left(\frac{1}{1+l_i\mu_q\phi_q}\right)^{\phi_q^{-1}} \left(\frac{l_i\mu_q}{\phi_q^{-1}+l_i\mu_q}\right)^{u_i} \right]$$

from Hardcastle and Kelley [17]

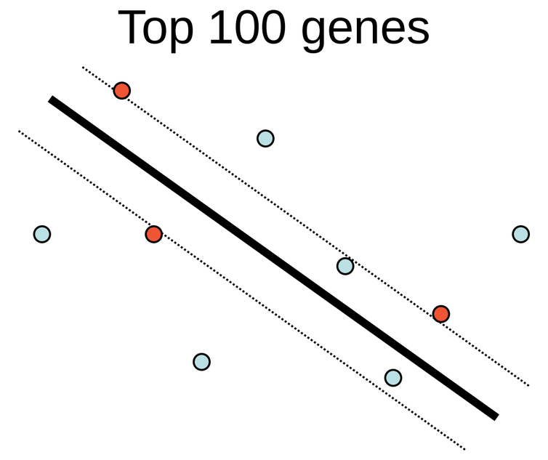
Weighted-Voting
36 genes with high prediction accuracy



SAM
142 genes (FP rate=0)



SVM-RFE
Top 100 genes

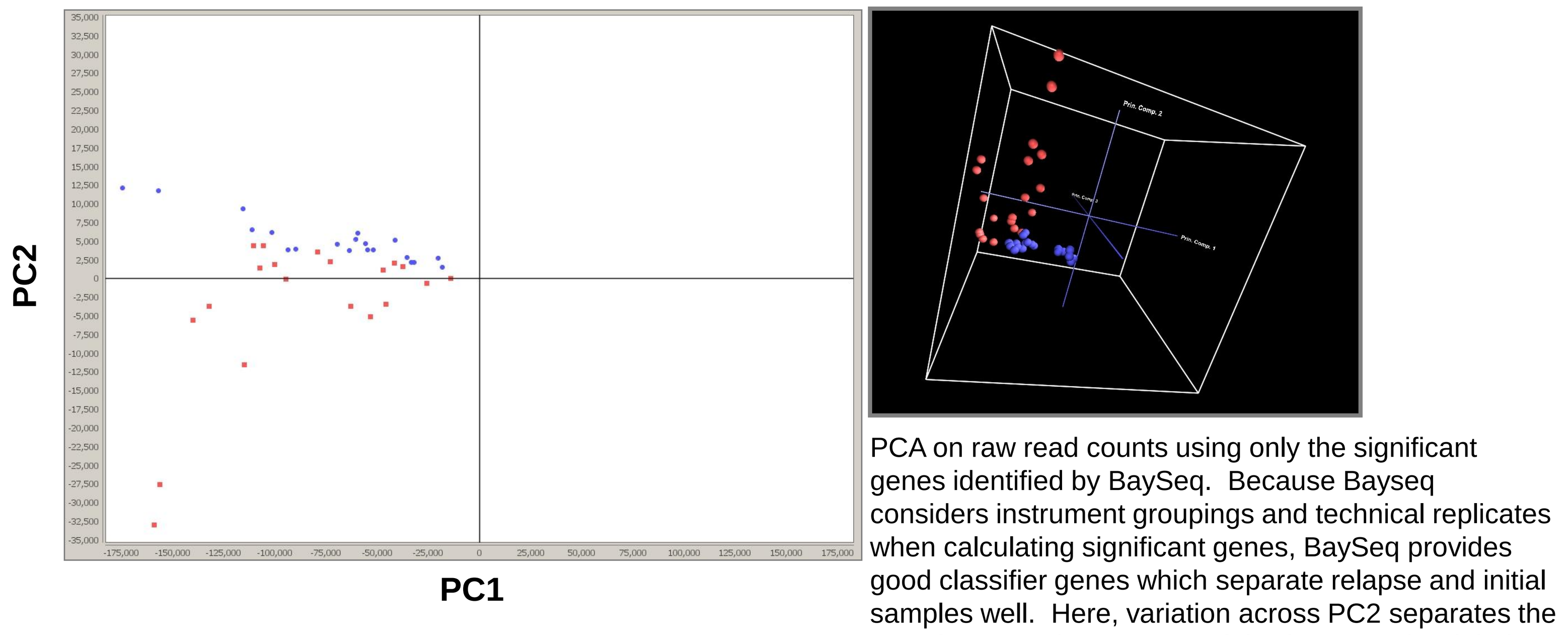


Several methods including BaySeq, Weighted-voting, SAM, and SVM-RFE were employed to discover gene panels:

35 genes selected by at least 2 of the 4 methods.

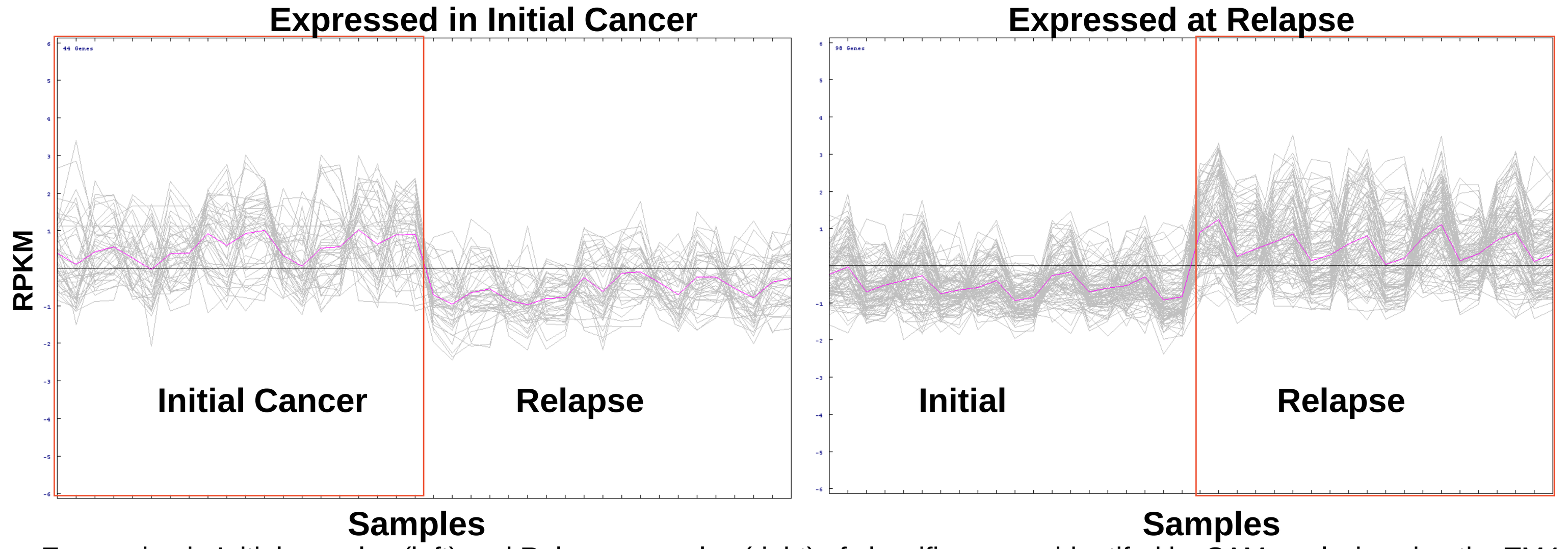
7 genes selected by at least 3 of the 4 methods.

Figure 4. BaySeq uses raw read counts to separate relapse and initial samples. Bayseq alone does a good job at identifying differentially expressed genes.



PCA on raw read counts using only the significant genes identified by BaySeq. Because Bayseq considers instrument groupings and technical replicates when calculating significant genes, BaySeq provides good classifier genes which separate relapse and initial samples well. Here, variation across PC2 separates the sample sets. The same result is shown on the right in three principle components. Blue: relapse; Red: initial

Figure 5. 142 classifier genes identified by SAM are differentially expressed in the initial and relapse states



Expression in Initial samples (left) and Relapse samples (right) of classifier genes identifed by SAM analysis using the TM4 analysis suite [11, 12]. Expression is shown as normalized RPKM values. SAM was conducted using a two-sample paired statistic. The 40 samples (including replicates) are listed at the bottom of each expression graph.

Figure 6. 35 top classifier genes identifed by at least 2 methods cluster according to initial and relapse states.

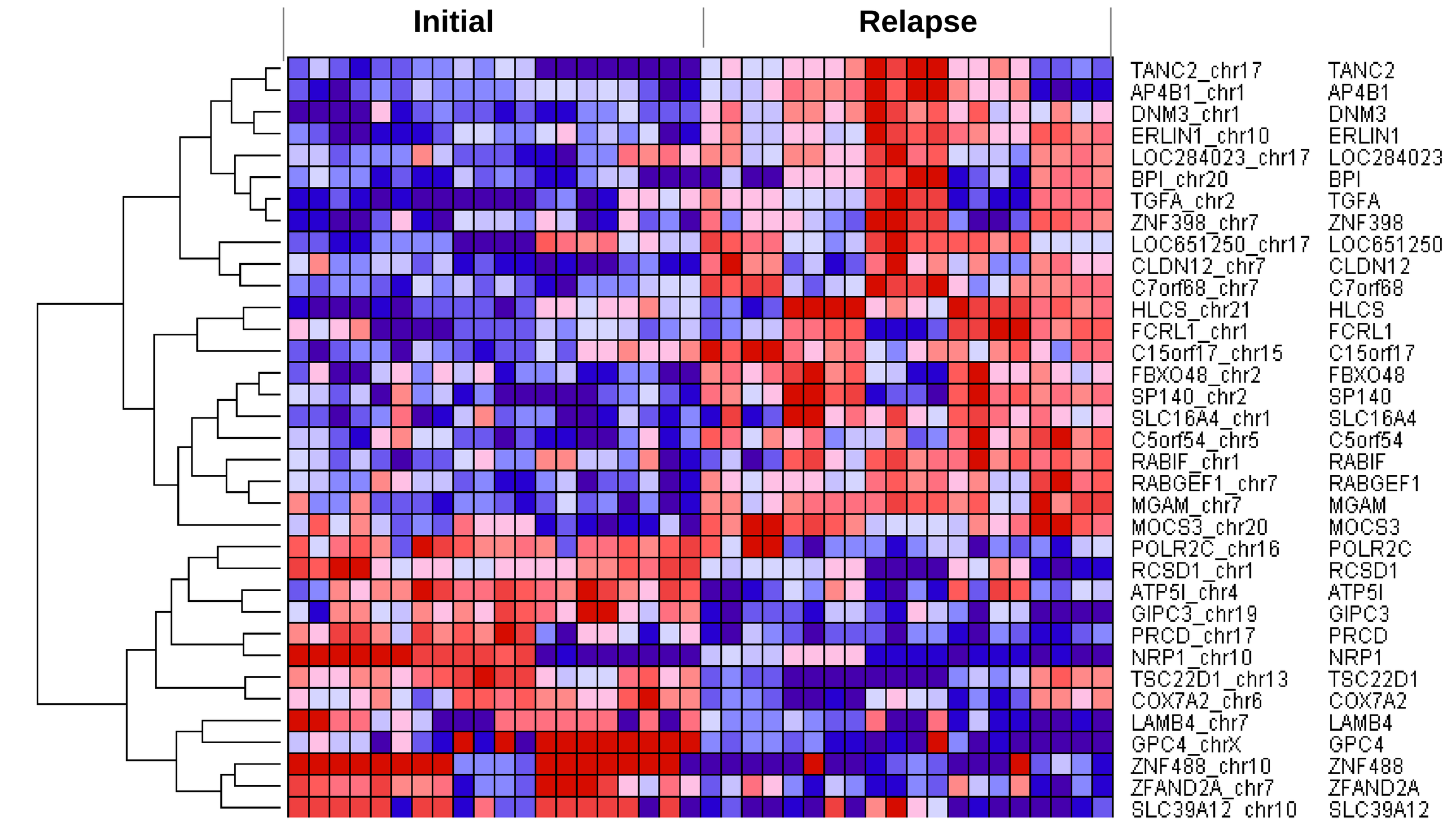
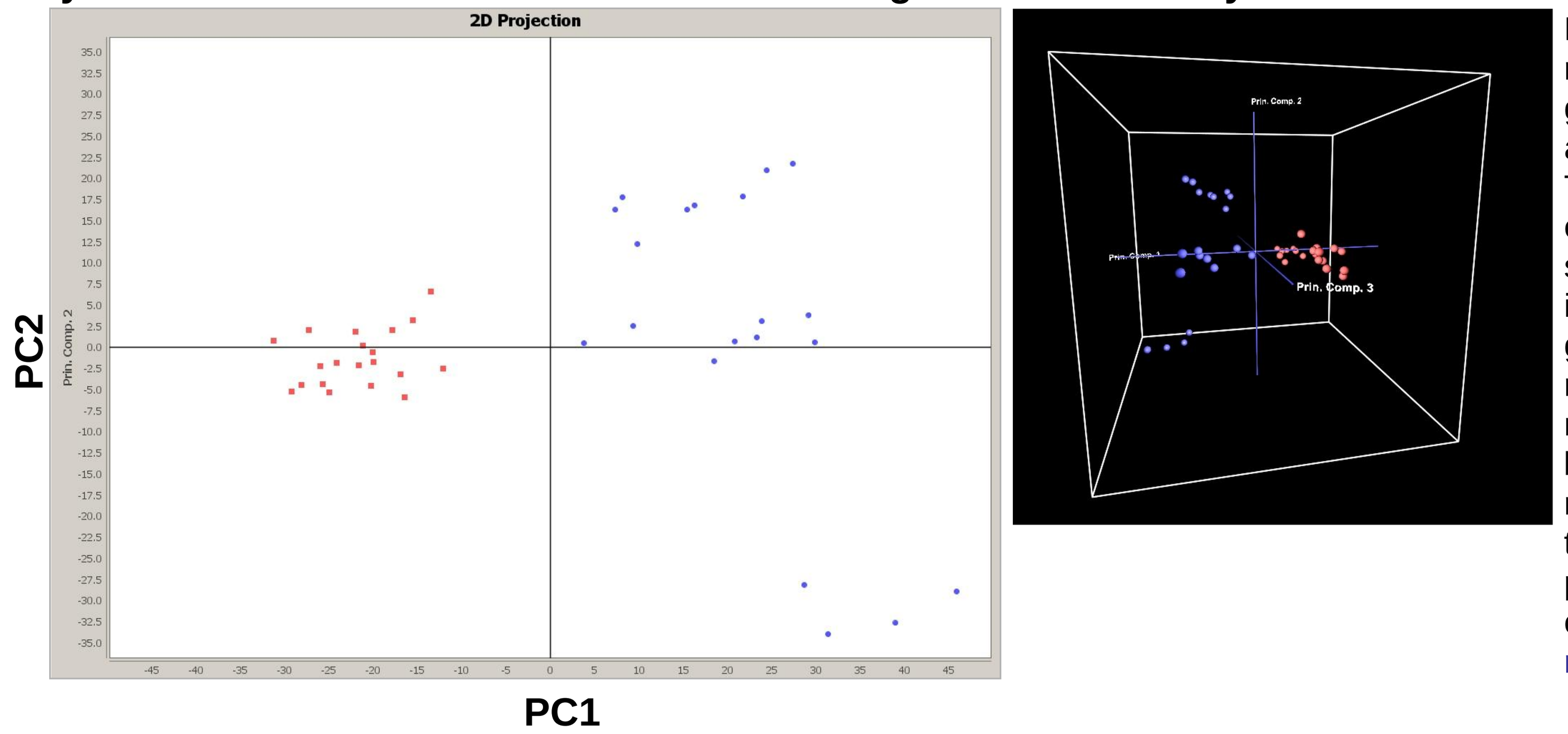


Figure 7. Combining methods perform best at separating initial and relapse states (shown here by PCA on normalized read counts for the 35 genes identified by at least 2 statistical methods).



PCA on normalized read counts using 35 genes identified by at least 2 methods. These genes provide excellent sample separation and include several genes known to be related to cancer, malignancy, or leukemia. The same result is shown on the right in three principle components. Blue: relapse; Red: initial

Table 1. The relapse state contains many SNPs not seen in the initial state.

Individual	SNPs Initial	SNPs Relapse	SNPs Unique to Relapse
1	43,918	39,728	**5,356**
2	51,113	79,917	**11,891**
3	87,217	109,401	**14,154**
4	71,902	68,249	**8,420**
5	37,874	43,127	**5,978**

Table 2. Multiple individuals share SNPs that are unique to the relapse state

Relapse SNP is shared across any:	# SNPs shared across *x* individuals
1	**43,932**
2	**911**
3	**15**
4	**0**
5	**0**

SNPs were detected in all RNA-Seq samples using Bioscope v1.3. Total (all SNP) dbSNP concordance ranged from ~80% to ~90% depending upon the desired stringency level in diBayes. SNPs in the relapse cases were compared to the initial samples to identify SNPs unique to the relapse condition for each individual. These "relapse unique" SNPs were examined to determine whether any were shared across individuals in the study. SNPs shared by 2 or more ALL sufferers fall into many genes which are annotated in known cancer pathways or associated with cancer-related annotation terms according to the DAVID genome analysis system.

CONCLUSIONS

•RNA-seq on the SOLiD system is a powerful tool for exploring gene expression and for uncovering polymorphisms in whole transciptomes.
•Bayesian systems such as BaySeq are able to normalize for potential run-time or instrument bias to select gene panels which yield excellent sample separation.
•Combined analysis with multiple techniques is successful at selecting gene panels that help us to better understand cancer and leukemia.
•SNP analysis in WT samples detects polymorphisms unique to the relapse state, suggesting that cancers refractive to treatment may develop common polymorphisms that potentially cause malignancy.

REFERENCES

[1] Total RNA-Seq kit: https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=607144
[2] HTSeq: Anders, Simon. EMBL Heidelberg (Genome Biology Unit). http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html#
[3] Eisen MB, et al. Cluster Analysis and Display of Genome-Wide Expression Patterns, PNAS, 14863-14868 (1998).
[4] de Hoon MJL, et al. Open Source Clustering Software. Bioinformatics, 20 (9): 1453--1454 (2004).
[5] Golub et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 1999 286:531-537. (1999).
[6] Reich M, et al. GenePattern 2.0 Nature Genetics 38(5):500-501. (2006)
[7] Tusher, Tibshirani and Chu. Significance analysis of microarrays applied to the ionizing radiation response. PNAS 2001 98: 5116-5121. (2001).
[8] Storey. A direct approach to false discovery rates. J.R. Stat. Soc. B, 64, part3:479-498. 2002
[9] Efron, B., et al. Empirical Bayes Analysis of a Microarray Experiment, JASA, 96: 1151-1160. (2001)
[10] Efron and Tibshirani. Microarrays, Empirical Bayes Methods, and False Discovery Rates" Genet. Epidemiol. Jun;23(1):70-86. (2002)
[11] Saeed AI, et al. TM4: a free, open-source system for microarray data management and analysis. Biotechniques. Feb;34(2):374-8. (2003).
[12] Saeed AI, et al. TM4 microarray software suite. . Methods in Enzymology. 411:134-93. (2006)
[13] C.Cortes and V.Vapnik. Support vector networks. Machine Learning 20:273-297. (1995).
[14] Guyon, et al. Machine Learning 46, 389–422. (2002).
[15] Weston J, Elisseeff A, BakIr G , Sinz F. http://www.kyb.tuebingen.mpg.de/bs/people/spider/main.html
[16] MATLAB (2006b, The MathWorks, Natick, MA)
[17] Hardcastle and Kelly BMC Bioinformatics 2010, 11:422

For Research Use Only. Not intended for any animal or human therapeutic or diagnostic use.

Life Technologies • 5791 Van Allen Way • Carlsbad, CA 92008 • www.lifetechnologies.com