SUPPLEMENTARY APPENDIX

Trisomy 12 chronic lymphocytic leukemia expresses a unique set of activated and targetable pathways

Lynne V. Abruzzo,[1*] Carmen D. Herling,[2] George A. Calin,[3] Christopher Oakes,[4] Lynn L. Barron,[5] Haley E. Banks,[5] Vikram Katju,[6] Michael J. Keating[7] and Kevin R. Coombes[6*]

[1]Department of Pathology, The Ohio State University, Columbus, OH, USA; [2]Department I for Internal Medicine and Center of Integrated Oncology, University of Cologne, Germany; [3]Department of Experimental Therapeutics, The University of Texas, MD Anderson Cancer Center, Houston, TX, USA; [4]Department of Internal Medicine, The Ohio State University, Columbus, OH, USA; [5]Department of Hematopathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA; [6]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA and [7]Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

**LVA and KRC contributed equally to this study*

 doi:10.3324/haematol.2018.190132

Received: February 5, 2018.

Accepted: June 29, 2018.

Pre-published: July 5, 2018.

Correspondence: lynne.abruzzo@osumc.edu

SUPPLEMENTAL METHODS

Sample collection and preparation

The discovery and validation sets were formed using different criteria. In essence, the experiment performed on the discovery set used a “cohort design” while the experiment performed on the validation set used a “case-control design”. During the period when many of the samples were collected, i.e., the early 2000's, the now-standard FISH panel for the common recurrent CLL cytogenetic abnormalities was not part of the routine care of CLL patients at M.D. Anderson. The Illumina microarray gene expression profiling study of the "discovery set" was performed using 114 patient samples, which were selected primarily from early samples with the longest clinical follow-up. At the time of collection, their cytogenetic status was unknown. We later performed Illumina SNP array profiling on these samples, which provided adequate resolution to identify their cytogenetic status. To perform the analysis described in this manuscript, we then restricted our attention to the subset of patients who had at most one of the common abnormalities, yielding the 97 patients in the discovery set. After analyzing the data to identify differentially expressed genes, we selected specific genes for validation using QRT-PCR, which is more sensitive and has a wider dynamic range than microarrays. At this point, the cytogenetic status of many more cases was already known, and we selected validation samples in such a way as to balance the number of cases between cytogenetic categories in order to get a more robust validation of differential expression.

Detection of genomic gains and losses by single nucleotide polymorphism (SNP) genotyping

We determined genomic copy number variations (CNV) by SNP genotyping.[17,18] We divided cases into cytogenetic subsets defined by abnormalities that would be detected using a panel of FISH probes to the common CLL cytogenetic abnormalities: del(11q), del(13q), del(17p), +12, and diploid. Segments of constant copy number in the LRR data were identified by applying the circular binary segmentation (CBS) algorithm, as implemented in the DNAcopy package (version 1.24.0).[19] Segments with mean LRR ≤ -0.15 and two bands in the BAF plot were called “deleted”; segments with mean LRR

>0.15 and four bands in the BAF plot were called “gained”. We identified 147 that contained single abnormalities or were diploid: 27 with +12 as the sole abnormality (18%), 49 with del(13q) as the sole abnormality (33%), and 47 diploid cases (32%). These were divided randomly between the discovery and validation sets. Because there were relatively few cases with del(11q) as the sole abnormality, to achieve statistical significance we used 10 cases with del(11q) as the sole abnormality for the discovery set, but included 14 cases with del(11q) and del(13q) in the validation set.

Transcriptional profiling of protein-coding genes using microarrays

For discovery, we performed transcriptional profiling using gene expression microarrays. We hybridized synthesized biotin-labelled cRNA prepared from purified CLL cells to Illumina® HumanHT-12 v4 arrays (www.illumina.com). The arrays contain 47,231 probes that target 32,378 unique REFSEQ entities representing 24,741 distinct mRNAs, 2,876 distinct non-coding RNAs (ncRNAs), and 2,976 distinct expressed sequence tags (ESTs). After RNA extraction, *in vitro* transcription of cDNA was performed to incorporate biotin-labelled nucleotides into cRNA (Illumina® TotalPrep™ RNA Amplification kit, Applied Biosystems). We performed hybridization, washing, blocking, streptavidin-Cy3-staining, and scanning following the manufacturer’s protocol. We acquired and analysed data using Illumina® GenomeStudio software. Quality control, pre-processing, and normalization were performed using version 2.2.1 of the lumi package[27] in version 2.12.0 of the R statistical software environment (http://cran.r-project.org). Batch effects were removed by aligning the mean and standard deviation of each gene. To identify differentially expressed (DE) genes, we compared the +12 cohort to the others individually and jointly. We assessed DE genes between cytogenetic subtypes by performing probe-by-probe ANOVA (for multiple subtypes) or t-tests (to compare two subtypes). We fit a beta-uniform-mixture (BUM) model[28] to the set of p-values in order to bound the false discovery rate (FDR).[29] Microarray data are available at http://silicovore.com/CLL/Trisomy12.

SUPPLEMENTARY TABLE LEGENDS

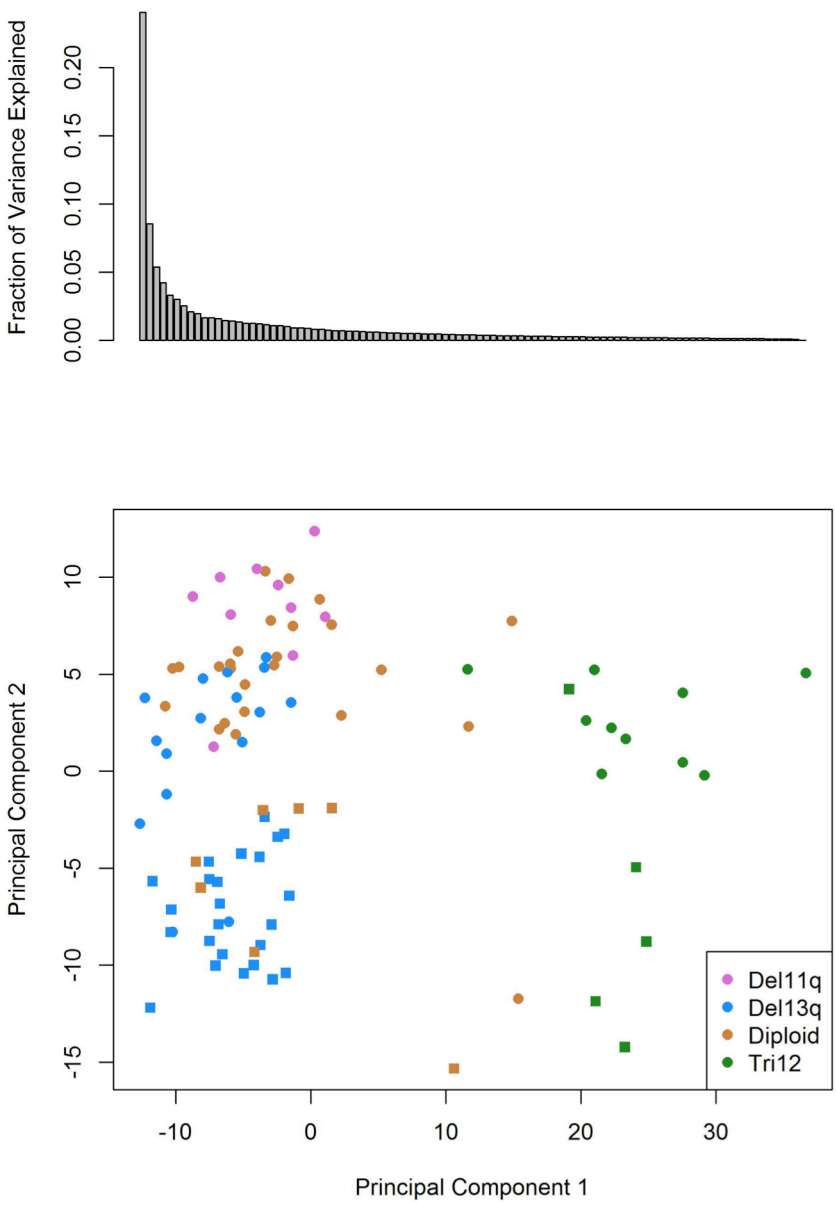
Supplementary Table 1: Analysis of differential expression in the Illumina microarray training set. Columns A-I are the probe annotations. Columns J-M are the mean log (base 2) expression values for each of the cytogenetic groups. Columns N-O are the results of probe-by-probe analysis of variance. Columns P-R come from t-tests comparing +12 to diploid cases. Columns S-U come from t-tests comparing +12 to del(13q) cases. Columns Y-AA come from t-tests comparing +12 to del(11q) cases. Column AB contains logical values indicating which genes are known targets of miR-15 or miR-16.

Supplementary Table 2: Analysis of differential expression between +12 cases of CLL with or without *NOTCH1* mutation. Columns A-G are probe annotations. Columns H-I are the results of probe-by-probe t-tests. Columns J-L are mean log (base 2) expression values in *NOTCH1*-mutated, *NOTCH1*-wild type, and all cases, respectively. Columns M-N represent the difference in expression between the two groups as fold change (on the raw scale) or as log ratios.

Supplementary Table 3: Results of an Ingenuity Pathways Analysis assessing canonical pathways. Each tab reports on the pathways associated with sets of differentially expressed genes in a different comparison. Negative log p-values, log ratios, z-scores, and lists of molecules are all as reported and exported by Ingenuity.

Supplementary Table 4: Analysis of differential expression in the QRT-PCR microfluidics card validation data set. Column A is the gene symbol. Columns B-F contain the mean log (base 2) expression of the five groups of samples. Columns G-H are the results of gene-by-gene analysis of variance. Columns I-J are t-test results comparing +12 to diploid. Columns K-L compare +12 to del(13q). Columns M-N compare +12 to normal B cell (NBC)l controls. Columns O-P compare +12 to del(11q). Columns Q-V compare del(11q) cases to other groups.

SUPPLEMENTARY FIGURE 1



Supplementary Figure 1: Principal components analysis of the training set using genes selected to be differentially expressed by ANOVA to be differentially expressed between at least two of the four groups. (**Top**) Scree plot showing the percentage of variance explained by each component. (**Bottom**) Scatter plot showing the first two components. The first principal component appears to represent differences between +12 and the other three groups, which is the strongest signal in the data