Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med 2013;368:2059-74. DOI: 10.1056/NEJMoa1301689

Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia

The Cancer Genome Atlas Research Network

Address correspondence to:
Timothy J. Ley, MD
Washington University School of Medicine
Division of Oncology, Stem Cell Biology Section
Campus Box 8007
660 South Euclid Avenue
St. Louis, MO 63110   USA
email: timley@wustl.edu

Author Contributions: The TCGA research network contributed collectively to this study. Annotated biospecimens were provided by Timothy J. Ley, John F. DiPersio, and Peter Westervelt. Data generation and analyses were performed by genome sequencing centers and cancer genome characterization centers. All data were released through the Data Coordinating Center. Project activities were coordinated by NCI and NHGRI project teams. We also acknowledge the following TCGA investigators of the AML Analysis Working Group who contributed substantially to the analysis and writing of this manuscript: Project leaders, Timothy J. Ley, Richard K. Wilson; manuscript coordinator, Timothy J. Ley; data coordinator, Eric E. Snyder; analysis coordinator, Timothy J. Ley; DNA sequence analysis, Christopher A. Miller, Michael D. McLellan, Timothy A. Graubert, Li Ding; mRNA microarray analysis, Katherine A. Hoadley; mRNA and miRNA sequence analysis, Andrew J. Mungall and Gordon A. Robertson; miRNA mutation analysis, Daniel Link; DNA methylation analysis, Timothy J. Triche, Jr., Peter W. Laird; copy number analysis, Christopher A. Miller, Li Ding; pathway/integrated analysis, Hsin-Ta Wu, Benjamin J. Raphael; pathology and clinical expertise, Jeffery Klco, John DiPersio, Timothy A. Graubert, and Peter Westervelt.

All of the primary sequence files are deposited in CGHub (https://cghub.ucsc.edu/); all other data including mutation annotation files are deposited at the Data Coordinating Center (http://cancergenome.nih.gov/). Sample lists, data matrices and supporting data can be found at http://tcga-data.nci.nih.gov/docs/publications/aml_2013/. The authors declare competing financial interests: details are available in the online version of the paper. Correspondence and requests for materials should be addressed to T.J.L. (timley@wustl.edu).

The Cancer Genome Atlas Network

Genome sequencing centers: Washington University in St Louis: Timothy J. Ley, Christopher A. Miller, Li Ding, Cyriac Kandoth, Charles Lu, Michael D. McLellan, Daniel C. Koboldt, David E. Larson, Ling Lin, John W. Wallis, Joshua McMichael, Ken Chen, Michael C. Wendl, Krishna-Latha Kanchi, Heather Schmidt, Joelle Kalicki-Veizer, Robert S. Fulton, Lucinda L. Fulton, Elaine R. Mardis, Richard K. Wilson. Broad Institute: Gaddy Getz, Stacy B. Gabriel, Carrie Sougnez, Lihua Zou.

Genome characterization centers: BC Cancer Agency: Andy Chu, Hye-Jung E. Chun, Richard Corbett, Andrew J. Mungall, Karen L. Mungall, Erin Pleasance, A. Gordon Robertson, Dominik Stoll, Adrian Ally, Miruna Balasundaram, Yaron S. N. Butterfield, Readman Chiu, Eric Chuah, Noreen Dhalla, Ranabir Guin, An He, Carrie Hirst, Martin Hirst, Robert A. Holt, Aly Karsan, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard A. Moore, Jeremy Parker, Patrick Plettner, Jacqueline E. Schein, Lucas Swanson, Angela Tam, Nina Thiessen, Richard J. Varhol, Natasja Wye, Yongjun Zhao, Inanc Birol, Steven J. M. Jones, Marco A. Marra; University of North Carolina, Chapel Hill: Katherine A. Hoadley; Brown

University: Hsin-Ta Wu, Fabio Vandin, Mark D.M. Leiserson, Benjamin J. Raphael; University of Southern California/Johns Hopkins: Timothy J. Triche, Jr, Daniel J. Weisenberger, Hui Shen, Phillip H. Lai, Moiz S. Bootwalla, David J. Van Den Berg, Stephen B. Baylin (Johns Hopkins), Peter W. Laird

Genome data analysis center: The University of Texas MD Anderson Cancer Center: Rehan Akbani, Nianxiang Zhang, Bradley M. Broom, Thomas C. Motter, John N. Weinstein

Biospecimen core resource: Nationwide Children's Hospital Biospecimen Core Resource: Tara M. Lichtenberg, Aaron D. Black, Christopher Adams, Thomas Grossman, Jay Bowen, Lisa Wise, Julie M. Gastier-Foster

Tissue source site: Washington University Medical School, Division of Oncology: Peter Westervelt, Timothy A. Graubert, Matthew J. Walter, John S. Welch, Lukas D. Wartman, Jeffrey M. Klco, David H. Spencer, Tamara L. Lamprecht, Jacqueline E. Payton, Mark A. Watson, Shashikant Kulkarni, Michael H. Tomasson, Sharon E. Heath, Jack D. Baty, Li Ding, Daniel C. Link, John F. DiPersio, Timothy J. Ley

Disease working group: John F. DiPersio, Timothy A. Graubert, James R. Downing, Michelle M. Le Beau, Khanh Ngyuen, Jean-Pierre Issa, Steven M. Kornblau, Hagop M. Kantarjian, Frederick R. Appelbaum, Martin Carroll, Richard K. Wilson, Timothy J. Ley

Data coordination centre: Eric E. Snyder, Shelley Alonso, Brenda Ayala, Julien Baboud, Mark Backus, Sean P. Barletta, Dominique Berton, Anna Chu, Stanley Girshik, Ari Kahn, Prachi Kothiyal, Matthew Nicholls, Todd D. Pihl, Rohini Raman, Rashmi N. Sanbhadti, Deepak Srinivasan, Jessica Walton, Yunha Wan, Mark A. Jensen, David A. Pot

Project team: National Cancer Institute: Margi Sheth, Kenna R. Mills Shaw, Greg Eley, Martin L. Ferguson, John A. Demchok, Tanja Davidsen, Liming Yang; National Human Genome Research Institute: Heidi J. Sofia, Zhining Wang, Bradley A. Ozenberger

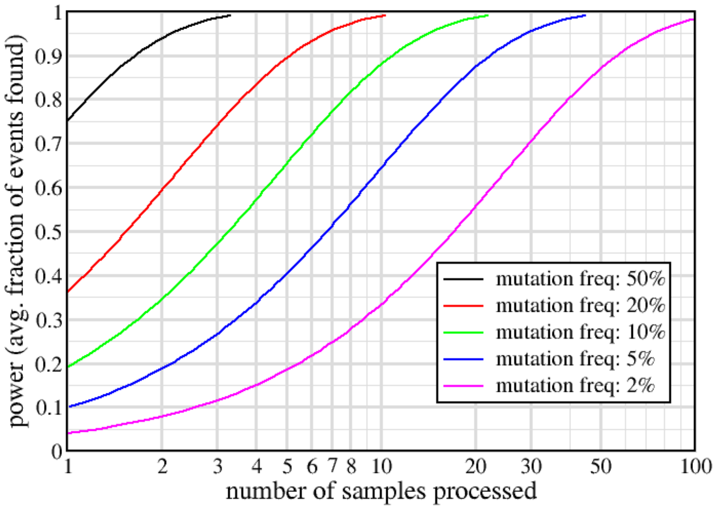
Table of Contents

A. Supplementary Materials and Methods

A.1 Patient Characteristics and Outcomes

A 1.1 Patient Characteristics

200 cases of *de novo* AML occurring in adults were selected from a larger cohort of AML patients enrolled in a single institution tissue banking protocol (with explicit consent for whole genome sequencing) that was approved by the Washington University Human Studies Committee (WU HSC #01-1014). All patients were between ages 18 and 88 and had previously untreated *de novo* AML. All samples were collected between November of 2001 and March of 2010. The study was designed and powered to detect >99% of mutations (at least once) that are present in at least 5% of all *de novo* AML cases; cases were chosen from a collection of more than 400 consented AML samples to represent the currently recognized subtypes of the disease (based on morphologic and cytogenetic criteria). Additionally, we required adequate sample quality and inventory for multiple testing platforms. Clinical characteristics at diagnosis, including peripheral blood white cell counts, blast percentages in blood and marrow, the distributions of FAB subtypes and cytogenetic risk groups, immunophenotypes, and the frequencies of known recurrently mutated genes (e.g., *FLT3, NPM1, DNMT3A, TP53,* etc.) were highly representative of adult patients with de novo AML (**Table 1)** (maintext references 8,15,16). Patients were treated in accordance with NCCN guidelines (www.nccn.org), with an emphasis on enrollment in therapeutic clinical trials wherever possible; however, patients were not treated uniformly within the intermediate and unfavorable cytogenetic risk groups, limiting the ability of this study to be used for outcomes predictions within these groups. Patients with unfavorable risk underwent allogeneic stem cell transplant if they were medically fit for the risks of transplantation, and if a suitably matched donor was available (22/43 patients: 8 matched sibling donors, 13 matched unrelated donors, and 1



haploidentical donor). Many intermediate risk patients also underwent allogeneic transplantation at some point in their disease course (48/115 patients: 24 matched sibling donors, 22 matched unrelated donors, and 2 haploidentical donors). Clinical data for all patients, including the treatment approach and outcomes data, are presented in Supplementary Table 1, and a multivariate analysis of outcomes is presented in the Supplementary Results (**Supplementary Table 2**). Overall and event- free survival data (classified by cytogenetic risk groups) are shown in **Supplementary Figure 1**.

A 1.2 Multivariate analysis of outcomes

The patients who participated in this study were not treated in a uniform fashion (**Table S1**). Many younger patients with intermediate and unfavorable risk cytogenetics were treated with allogeneic stem cell transplants at some point in their treatment course. Analysis of outcomes was therefore confounded by the non-uniform treatment of intermediate risk and unfavorable risk patients, and the relatively small sample size of the study.

For both event free and overall survival, the variables in the base model were the same: the total peripheral blood white blood cell (WBC) count at presentation, and cytogenetic risk. A white blood cell count (WBC) of > 16,000/ul in the peripheral blood at presentation was associated with poorer survival. The three cytogenetic-risk groups were associated with different survival outcomes in the expected directions: patients with favorable cytogenetic risk had better survival than the other two groups, and those with intermediate risk had better survival than those with unfavorable risk. These models incorporated age as a stratifying variable rather than as a covariate, for two reasons: 1) age is known to affect outcome, but our primary goal was to obtain estimates of the effects of the other variables, and 2) there was an indication that age (as a covariate) violated the proportional-hazards assumption. Using it as a stratification variable removed this problem.

The genes that were significant or nearly significant in univariate survival models were added separately to the base model. *TP53* was the only gene that had a significant effect when added to the base model for overall survival, with a hazard ratio of 2.61 (95% CI: 1.30 - 5.23.) *DNMT3A* and *FLT3* had p-values of 0.08 and 0.09, respectively. With *TP53* in the model, unfavorable cytogenetics was not a predictor of survival, since 14 of the 16 patients with *TP53* mutations had unfavorable risk, creating confounding variables. The same was true for the analysis of event-free survival (**Table S2**).

The same steps were used to create the model for event-free survival. *TP53* and *FLT3* were both associated with poorer survival when added to the base model. A final model with WBC, cytogenetic risk, *TP53*, and *FLT3* was therefore created; both genes were still significant when placed in the model together.

A.2. Copy number and LOH analysis

A.2.1 Copy Number and LOH Methods

Affymetrix Genome Wide SNP Arrays v6 were used to assay the data. Intensity values were normalized using Partek Genomics Suite. Segmentation and copy number calling were done using R (version 2.15.1): Log2 ratios for each probe were boosted by 25% to account for tumor/normal admixture, then segmented using Circular Binary Segmentation (CBS), as implemented in the DNACopy package (v1.24.0). Adjacent segments with similar copy number were then merged. Likely false positives were removed by requiring segments to contain at least 30 probes and be at least 35kb in length, criteria based on previous work[1]. Loss of heterozygosity/Uniparental Disomy was assayed by segmenting SNP data using CBS and manually reviewed to arrive at a high-confidence set of calls. Results are included in **Table S5** and shown in **Figure S2**.

A.2.2 Copy number alterations in AML genomes

Most de novo AML samples with intermediate and favorable risk cytogenetics had very few copy number events detected by high-resolution SNP arrays (**Figure S2** and **Table S5**). 133 of the 200 cases (67%) contained one or more detectable somatic copy-number amplifications or deletions. Overall, we detected a median of one somatic copy number alteration per genome, comparable to what was reported in Walter et al (2009) 3 ; 79/86 cases in that study were included in this dataset. In cases with unfavorable risk cytogenetics, the number of copy number variants was considerably higher, since all of these samples have cytogenetically detectable copy number alterations (median 6 copy number events per case, maximum of 57). These cases contain several expected copy number changes, including deletions that frequently involved large portions of chromosomes 5 and 7, and amplification of all or part of chromosome 8. No cases contained evidence of chromothripsis 18. Several small events affected recurrently mutated AML genes, including focal deletions containing *DNMT3A, STAG2, KDM6A,* and *NF1.* We also identified 22 copy number neutral LOH events (also known as partial uniparental disomy, or pUPD) in 19 samples; five of those events affected chromosome 17p in cases with mutations of *TP53*, leading to loss of heterozygosity. Other important genes in regions of pUPD included *RUNX1*, *WT1*, and *U2AF1*.

A.3. DNA sequencing and analysis

A.3.1. Illumina library construction and whole genome sequencing

The procedure described by Mardis *et al* [2] was followed for library construction and whole genome sequencing. Briefly, Illumina DNA sequencing was used to generate between 58.5 and 155.9 billion base pairs of sequence data for each of the 50 tumors and their matched normal samples, with haploid coverages ranging from 18.9 to 50.4 (**Table S3**). Comparison of heterozygous SNPs detected in the whole genome sequencing (WGS) data with SNP array genotypes confirmed bi-allelic detection of between 96.60 and 99.86 percent of the heterozygous array SNPs in the 50 cases. Detailed coverages for all cases are included in **Table S3**.

A.3.2. Illumina library construction and exome sequencing

Libraries for whole exome sequencing were constructed and sequenced on either an Illumina HiSeq 2000 or Illumina GA-IIX using 76 bp paired-end reads. Details of whole exome library construction have been given elsewhere [3]. Standard quality control metrics, including error rates, percentage passing filter reads, and total Gb produced, were used to characterize process performance before downstream analysis. The Illumina pipeline generates data files (BAM files) that contain the reads together with quality parameters. Output from Illumina software was processed by the “Picard” data processing pipeline to yield BAM files containing aligned reads (via MAQ or BWA, to the NCBI Human Reference Genome Build hg18) with well-calibrated quality scores [4,5]. Exome sequencing coverage for 150 cases are included in **Table S3.**

A.3.3. Mutation detection pipeline

For each sample, reads were aligned using either Maq 0.6.8 or 0.7.1 or BWA 0.5.5 on a per-lane basis, merged into a single BAM file, and duplicate reads were removed using Picard 1.17, 1.22, or 1.25 (http://picard.sourceforge.net). Sample variants were called using Samtools (svn rev 320, 453, 544, or 599) [6]. Somatic single nucleotide variants were detected using SomaticSniper[7]. High quality somatic predictions were defined as those sites with a SomaticSniper somatic score greater than 40 and an average mapping quality greater than 40. Indels in all samples were called using a combination of Pindel [8] and GATK [9]._Somatic variants were grouped into tiers based on genome annotation as described previously[2]. Y chromosome variants were removed for all female patients.

A.3.4. Structural variant detection

Structural variants (SVs) in all samples were predicted by BreakDancer[10] and SquareDancer (unpublished). All SV predictions were filtered using TIGRA (Chen et al., in preparation) to identify assembled breakpoints in SV flanking reads. The same procedure as described in Ding *et al.* [11] for selecting somatic SVs was used.

A.3.5. Capture array design for WGS cases

We used custom sequence capture arrays from Roche Nimblegen to validate putative WGS mutations. To perform solid phase capture validation for individual samples, we included all tier 1 to 3 sites and all coding exons of 11 genes (*CBL*, *CEBPA*, *DNMT3A*, *GATA1*, *JAK1*, *JAK2*, *NOTCH1*, *PTPN11*, *RUNX1*, *TET2*, and *TYK2*).

For small insertions and SNVs, the targeted regions were exactly 200 bp centered on the variant. For small deletions, the deleted sequence plus 100 bp of sequence flanking each end of the deletion were selected.

For putative somatic SVs, we requested probes tiled across the predicted breakpoint flanking 100 bp of the outermost, predicted breakpoint. For larger insertions a single region was requested, but for translocations, deletions, and inversions etc, we requested two targets, one for each breakpoint. Roche Nimblegen design parameters allowed for probes with up to five additional sequence matches elsewhere in the genome.

A.3.6. Capture validation of exome data

Exome probes were designed as described for whole genome, except the targets consisted of the following: a) all putative SNVs, indels, and SVs identified in the exome sequencing, b) all tier 1 SNVs and indels found in the 50 cases sequenced by WGS, c) all exons from significantly mutated genes found in the 50 tumors sequenced by WGS, d) all recurrent (within 200bp) tier 2/3 mutations found in the WGS.

A.3.7. Alignment of solid and liquid phase capture validation data

We generated 100 bp paired-end Illumina sequence data for 200 tumor or normal genomes. Illumina reads were mapped to the NCBI Build 36 reference sequence (BWA v0.5.5 or 0.5.9), merged into BAM files (SAMtools v1 r544 or r599), and duplicate reads were tagged (Picard v1.17 or 1.29). Coverage of target sequences was assessed using RefCov software (T. Wylie et al, unpublished). We obtained greater than 20X haploid reference coverage for 66.0 to 98.6% (median 94.5%) of the targeted sites in each sample pair.

A.3.8. Validation of SNVs and dinucleotide variants

Putative SNVs and dinucleotide variants were validated using VarScan 2 (http://varscan.sourceforge.net) with the following parameters:

-min-coverage 30
-min-var-freq 0.08
-normal-purity 1
-p-value 0.10
-somatic-p-value 0.001
-validation 1

Based on the allele frequency and reads supporting reference and variant alleles at the position of each predicted variant in the tumor and normal BAMs, VarScan classifies each putative somatic event as Reference (wildtype), Germline, Somatic, or LOH. Validated somatic mutations are further filtered with additional filters that removes false positives supported by strand specific artifacts, read position artifacts, or poorly mapped reads. Potentially ambiguous sites were further resolved with additional visualization of the primary and validation data.

Validation data was supplemented with information from previous studies that used the same tumors[11,12]. *CEBPA* had low coverage in validation sequencing data, so RNA-sequencing was used to provide additional evidence for the presence of somatic mutations. samtools mpileup (parameters: -B -q 1) was run on the CEBPA locus and the results analyzed with Varscan2 (params --min-coverage 3 --min-var-freq 0.05 --p-value 0.10). Extensive manual curation was performed to remove false positive variants and rescue false-negatives.

A.3.9. Validation of indels

<u>Small (1-2 bp) Indel Validation with liquid- and Solid-Phase Capture Validation Data</u>

Putative indels 1-2 bp in size were converted to BED format and provided as the target intervals for the GATK IndelRealigner algorithm. BAM files for the tumor and matched normal were re-aligned independently using this set of target intervals.

To validate the original predictions, we developed a matching algorithm that attempts to match VarScan validation calls with the original indel predictions. Specifically, the algorithm searched for a validated indel of same type (insertion or deletion) and similar size (within 1 bp). To allow for differences in gapped alignments, the algorithm allowed matches at slightly different genomic positions, so long as the validated indel mapped within a specified interval (indel_size + 2bp) of the original prediction. Matched indels reported "Somatic" in the tumor sample were manually reviewed in the re-aligned BAM files using IGV to visualize the data.

Medium (3-100 bp) Indel Validation with liquid and Solid-Phase Capture Validation Data

Sample validation data for indels 3-100 bp in size were assembled using the TIGRA assembler (Chen et al. unpublished). Breakpoints and microhomology were identified using Crossmatch alignments (version 1.080721, Green unpublished). We then sized the chosen contigs to 500 bp length by trimming excess sequence or padding from the reference sequence and compared overlapping contigs using the dpAlign module of BioPerl (http://www.bioperl.org/wiki/Main_Page) to generate an "Ends-free" alignment between the two pairs. If an alignment contained no gaps, shared at least 98% sequence identity and had a length of at least 95 bp then the leftmost contig that aligned to the reference was retained. Contigs that remained after merging were concatenated to the NCBI Build36 reference sequence as additional novel contigs and the validation reads were mapped back to the expanded reference using BWA and depduplicated using Picard (http://sourceforge.net/apps/mediawiki/picard). Those with a mapping quality greater than 0 that completely spanned the established indel breakpoints without gaps in the alignment were identified. Variants with greater than 30 reads aligning to either the reference or the indel contig and a variant allele frequency difference of greater than 10% between any two samples were manually reviewed. Additionally, 454 and 3730 sequencing were used to resequence the NPM1 and FLT3 recurrent mutation sites in some samples as previously reported[11].

A.3.10. Validation of structural variants

All BWA-aligned capture reads and their mates that mapped within 1000 bp of the structural variant breakpoints were realigned by CrossMatch (version 1.080721) to the assembled SV contigs and to the reference. The threshold for an acceptable alignment was ≤ 1 unaligned base at either end, ≤ 1% substitutions, ≤ 1% indels and a CrossMatch score ≥ 50. An SV-supporting read was required to span the breakpoint on the SV contig, align to 10 bases of flanking on each side of the breakpoint, and have no alignment to the reference above minimum alignment criteria. SV-supporting reads were tabulated in the tumor and normal sample separately, and Fisher's exact test was applied to these counts to determine the somatic status of each variant. The same method for determining SV-supporting reads was applied to the WGS alignment data for those calls deemed somatic by all other criteria. Variants with any SV-supporting reads in the normal WGS sample were filtered out as potential germline variants or alignment artifacts. An additional filter was put in place to filter ALU sequences and the remaining high confidence SV events were manually reviewed based on BWA mapping of supporting capture validation data to the assembled SV contigs spanning the breakpoint.

A.3.11. Kernel density analysis for identifying clusters and estimating allele frequencies for each tumor

Tumor clonality estimates were determined using the mutation allele frequencies from sites with deep coverage from capture validation data. To minimize the effect of coverage on allele frequency estimations, only mutations with >100x coverage in both the normal and tumor validation data were included in this analysis. Varscan 2 was utilized on whole-genome sequencing data to eliminate all LOH SNV calls. For each chromosome, the variant allele frequencies were plotted from both the tumor and normal in copy-number neutral regions. A kernel density estimate (KDE) plot was drawn for tumor variant allele frequencies using the density function in R. A peak-finding function evaluated each KDE plot to determine the number of peaks. The clusters identified served as an estimation of the number and relative composition of clones and subclones present in each tumor.

A.3.12. Significantly mutated gene analysis

We used components of the Mutational Significance in Cancer (MuSiC) package to determine significantly mutated genes (SMG) and pathways. The SMG test in MuSiC assigns mutations to seven categories, including AT transition, AT transversion, CG transition, CG transversion, CpG transition, CpG transversion, and indel, and then uses statistical tests including convolution, Fisher’s test, and a likelihood test to combine the category-specific binomials to obtain an overall p value All P-values were combined using the same methods as described in Dees *et al*.[13]. SMGs are listed in **Table S7**.

A.3.13. Recurrent mutations in non-genic regions

We screened for recurrent non-coding mutations in the 50 samples with whole genome sequencing data, and identified 69 tier 2 or 3 regions with two mutations within 200 base pairs of each other (**Table S8**). These regions were then annotated with ENCODE chromatin state segmentation data from the K562 erythroleukemia cell line, which expresses *BCR-ABL,* and which is the only myeloid leukemia line that has been fully analyzed to date[14]. 87% of the regions were annotated as heterochromatin, repressed, or repetitive, and are unlikely to have any functional role. Of the remaining mutations, most have low mammalian conservation scores. Only one site was associated with an active promoter, which was upstream from a pseudogene. We compared these regions to results from whole genome sequencing of an additional 50 AML genomes (unpublished data), and identified no overlapping regions with recurrent mutations between the two sets, suggesting that most of these mutations are random.

A.3.14. Mitochondrial variants

Human cells contain dozens to hundreds of mitochondria, and each mitochondrion contains multiple copies of DNA (mtDNA). Somatic mutations have been identified in mtDNA in a variety of cancers and can be homoplastic (all of the mtDNA copies are identical in all cells) or heteroplastic (only a fraction of the mtDNA in cells carry the mutation)[15]. We identified 40 variants (in 31 patients) that were enriched in the tumor samples relative to the skin (**Table S9**). Many of these variants were heteroplastic in the skin at low levels (range 0.1-7.7%). The low level of heteroplasty in the skin samples may represent tumor cell contamination, or inherited variants that are variable from tissue to tissue, as previously described in normal individuals[16]. The functional significance of these variants is unknown, and several have been previously described in normal individuals[17]. These results are similar to the mtDNA mutations independently identified and reported for several of these AML samples, extracted from this dataset[18].

A.3.15. Germline Variant Calling

Germline SNPs and indels were identified in GRCh37 aligned tumor-normal BAM pairs using VarScan 2.2.6 (http://varscan.sourceforge.net) with the following parameters: min-coverage 30 -min-var-freq 0.08 -normal-purity 1 -p-value 0.10 -somatic-p-value 0.001 -validation 1). Additional germline SNPs were identified using Samtools (v1.1.16-(r963:234)) and additional germline indels were extracted using GATK (v3 http://genome.cshlp.org/cgi/reprint/gr.107524.110v1). Predicted variants were filtered to remove false positives from homopolymer repeats, strand-specific artifacts, ambiguously mapped reads, and variants supported exclusively by low quality data at the beginning or end of reads. Additionally, those variants with a supporting variant allele frequency <8% were filtered. Variants were annotated using a combination of NCBI Refseq and Ensembl transcripts and only truncating variants defined as nonsense, frameshift, or disrupting the canonical splice donor/acceptor were retained. Additional filters were applied to remove variants affecting olfactory receptors, annotated noncoding/RNA-genes, annotated pseudogenes, transcripts with incomplete open reading frames lacking a start or stop codon, predicted genes, hypothetical genes, transcripts exclusive to Ensembl, as well as genes suspected to have missing paralogs in the human reference. Common truncation variants with a reported frequency >1% from the 1000 Genomes project or the Caucasian population in the NHLBI GO Exome Sequence Project data were filtered, as well as any germline variants that were recurrent at the same position in more than 2% of the cohort. Sequence data supporting all remaining germline truncating variants were visually examined with Integrative Genomics Viewer and any data that appeared to be an artifact, somatic mutation, or complex in-frame indels were discarded.

We interrogated the variants found in normal skin samples from 200 subjects, and identified 1,547 predicted truncating variants (1,364 unique alleles) in 1,219 genes (**Table S11**). Predicted truncating variants included frameshift-deletions (30.9%), frameshift-insertions (11.5%), stop-gain (37.0%) and splice site alterations (20.6%). All 1,364 alleles have been previously identified in other populations (82 in dbSNP build 135, the remainder in the NHLBI Exome Sequencing Project). Truncating variants found in the skin samples were common (**Figure S6a**, median of 7 per subject), consistent with recently reported findings[19]. The burden of truncating variants was not associated with age at AML diagnosis (**Figure S6b**). Among the 154 genes with nonsynonymous somatic mutations in at least 2 subjects, 58 truncating mutations in 30 genes were detected in the skin samples; 14 genes were mutated more than once (**Figure S6c**). This proportion is greater than expected by chance (*P* =10-50, compared to 1,489 germline truncations in the remaining 23,491 genes with adequate sequence coverage). For example, we previously reported that two of the patients in this cohort with early-onset AML (diagnosed at age 25) had novel germline truncating mutations in *WT1* (R430*) or *PTPN11* (Y197*)[12]. However, a minority of genes with germline truncating alleles are expressed in AML cells (23.1% of variants with 10 or more reads in tumor by RNA-seq), and only a small fraction of these (16.2%) demonstrate mutant allele-specific loss of expression in the tumors, suggesting that the vast majority of the truncating variants are irrelevant for AML pathogenesis.

A.3.15 Allele-specific expression

Tumor and RNA-seq readcounts were extracted and used to determine the ratios of mutant-allele to wild-type allele expression for each validated SNV. Six putative tumor-suppressor genes are shown in **Figure S4**.[20]. Only sites with at least 10x coverage were considered to minimize sampling error. On the plots, shifts away from the diagonal indicate allelic expression bias. Some sites appear at or near 100% in the RNA-seq. Hemizygosity due to copy number changes, UPD, or presence on sex chromosomes in males explain some, but not all of these events.

A.3.16 Outlier analysis

The number of tier 1 mutations in each sample was calculated and outlier samples were identified as those with a number of mutations 1.5 times the interquartile distance away from the quartiles.

A.3.17 Comparison to other cancers

Data on tier 1 (coding) mutations was obtained from four published TCGA projects: breast cancer[21], squamous cell lung cancer[22], colorectal cancer[23], and ovarian cancer[24]. A Student's t-test was used to compare the tier 1 mutation counts for all tumors in these sets to the tier 1 counts for all AML tumors. (**Figure S18**)

A.4. Affymetrix Expression Array data generation and analysis

The experiments and analysis were conducted by using approaches described in Payton et al.[25].

A.5. Pathway analysis

A.5.1. Mutation Data for Pathway Analysis

We analyzed combined mutation, fusion gene, and copy number data for 200 samples annotated in **Supplementary Tables 5, 6, and 13**. For somatic mutations, we considered single nucleotide variants and small indels, ignoring mutations marked as silent, germline or with validation status of wild type. In total, 1,476 genes contained at least one such somatic mutation in at least one sample. Moreover, chromosomal rearrangements in AML patients produce fusion proteins, therefore in total 32 in-frame and 38 out-of-frame fusion genes identified by RNA-seq are included in the analysis. We also included *MLL* partial tandem duplications (*MLL*-PTD) in 9 samples and micro-deletion in 3 samples. For copy number data, we used focal copy number and uniparental disomy (UPD). In focal copy number data, there are 292 altered genes in 76 samples. A total of 6,859 genes marked as containing UPD in 19 samples. (More details are given in **Table S5**) The pathway picture (**Figure 2**) uses this comprehensive mutation matrix on a reduced set of genes, as described below.

For the HotNet[26] subnetwork analysis below, we used a reduced mutation matrix that included the mutations and aberrations above, with the exception of UPD. Regions of UPD generally span large regions of the genome, making it difficult to identify the target gene of each such aberration. In addition, we restricted attention to four out-of-frame fusion genes: two *RUNX1* out-of-frame fusions (e.g. *RUNX1-ADAMTS19* and *RUNX1-PRRC1*) in sample TCGA-AB-2854 were annotated as mutations (loss of function) in *RUNX1*. Out-of-frame fusions *MDM4-DNMT3B* (TCGA-AB-2904) and *MLL3-ACTR3B* (TCGA-AB-2932) are marked as mutations in *DNMT3B* and *MLL3*, respectively. Finally, the out-of-frame fusion *MLLT10-CEP164* (TCGA-AB-2985) was annotated as a mutation in *MLLT10*, a myeloid transcription factor.

A.5.2. HotNet analysis

We used HotNet[26] to identify subnetworks of a large protein-protein interaction network that contain genes with significant numbers of aberrations, using the reduced mutation matrix described in previous section. HotNet considers each mutation or copy number alteration (CNA) in each sample as a unit heat source, and uses a diffusion process to derive "hot" subnetworks that contain more alterations than expected by chance. Therefore the significance of a subnetwork is determined by both the frequency of alteration of genes in the subnetwork and the local topology of the subnetwork. HotNet returns a list of subnetworks, each containing at least *s* genes, and employs a two-stage statistical test to assess the significance of the list of subnetworks. The first stage of the test computes a *p-value* for the number of subnetworks in the list, for different values of *s,* under a suitable null hypothesis. The second stage estimates the false discovery rate (FDR) of the *list* of subnetworks, providing a bound on the number of subnetworks in the list that are expected to be significant. Finally, we assess the significance of each individual subnetwork in the list by comparing to known pathways and protein complexes.

We analyzed the combined mutation and copy number for the 200 samples. For each sequenced gene, we defined the gene as *altered* in a sample if the gene had an aberration in the sample, where the aberrations considered are described in previous section. Moreover, we discarded CNAs for which the sign of the aberration was not the same in at least 90% of altered samples. The resulting alteration data on 200 samples was input to HotNet. We used the interaction network derived from the iRefIndex[27]. For the HotNet statistical test, we generated random datasets in the following manner. We simulated mutations using an estimated background mutation rate (2.97×10^{-7}). This rate is higher than the background rate observed in the 200 AML samples analyzed. Therefore, our analysis is conservative. We simulated CNAs using the observed distribution of CNA lengths, permuting their genomic positions. The latter minimizes potential artifacts resulting from functionally related genes that are both neighbors on the interaction network and close enough on the genome that they are affected by the same CNA. We also removed genes that are potentially biased toward a higher number of silent mutations than expected (because of their length or higher background mutation rate).

Using this approach HotNet identified 4 subnetworks containing at least 6 genes ($P < 0.001$) with a corresponding FDR <= 0.1 for the list of subnetworks (**Table S17**). To gain additional support for individual subnetworks and to focus attention on subnetworks with known biological function, we computed the overlap between the genes in candidate subnetworks and: (i) pathways from the KEGG database[28]; (ii) protein complexes from PINdb[29]. All of the subnetworks reported by HotNet have statistically significant (corrected $P \leq 0.05$) overlap with at least one KEGG pathway or PINdb protein complex (**Table S18**). In particular, HotNet identifies: a subnetwork containing genes in the cohesin complex, a subnetwork that overlaps part of the acute myeloid leukemia KEGG pathway; a subnetwork containing some genes in the polycomb complex; and part of the PTIP histone methyltransferase complex. The identification of the KEGG name “Acute Myeloid Leukemia” pathway is obviously not a new discovery. However, we note that this pathway was not pre-selected but rather automatically identified by HotNet in a large protein-protein interaction network containing more than 9000 proteins. Thus, this identification serves as a positive control that HotNet identifies mutated subnetworks that are meaningful for AML

A.5.3. Defining functional gene groups from biological knowledge

Many genes are mutated at very low frequencies in the dataset: for example, of the 1,779 genes marked with a mutation in the reduced matrix we used in HotNet, 1,555 are mutated in only a single sample. To increase our sensitivity in annotating these rare mutations, we combined some of the genes with rare mutations into 9 functional groups, based on prior knowledge of the biological function of these genes and/or prior reports of their role in AML. We combined all fusions involving the *MLL* gene into a group called “MLL-X fusions” **(Figure 2** and **Figure S9 and S10)**. We combined genes in the cohesin complex (**Figure S7**), spliceosome (**Figure S11a**), chromatin modifier (**Figure S11b**), and myeloid transcription factors (**Figure 10c**) into fS7our additional groups. We also formed groups (**Figure S9**) for tyrosine kinases, serine/threonine kinases, protein tyrosine phosphatases (PTP), and Ras proteins. We trimmed each of these gene groups to the subset of genes exhibiting approximate mutual exclusivity in their mutations with the following approach. (1) Define the initial gene set as all genes in the group having at least one exclusive mutation. (2) Remove all genes that are mutated in only one sample, if this sample also contains another mutation in the same group. (3) Add the remaining genes in the initial gene set.

Table S18 lists the groups, with the corresponding mutation matrices in the above-referenced figures.

A.5.4. One-Hit and Two-Hit mutation matrices

We created two mutation matrices using the functional gene groups described above. The first is a “one-hit” mutation matrix, that marks a gene, or functional gene group, as altered in a sample if it has at least one mutation from the comprehensive mutation matrix. The second “two-hit” mutation matrix annotates when both homologs of a gene are mutated. In most cases, we cannot uniquely identify which homolog is mutated, and so we mark a gene as a “two-hit” if it has at least two mutations, of any type, in the sample. Otherwise, if there exist only one mutation in the sample, we still treat the mutation on the gene as one-hit. **Figure 2** in the main text shows the “two-hit” mutation matrix.

A.5.5. Pathways identified by computational approach

We applied our Dendrix++ algorithm (Wu et al., *in preparation*), for identifying mutual exclusive sets of mutations to the one-hit mutation matrix. Dendrix++ extends our earlier De novo Driver Exclusivity (Dendrix) algorithm[30], and both algorithms aim to find sets of mutations that occur in many patients and are (approximately) mutually exclusive. The key difference between the Dendrix algorithm and Dendrix++ is that Dendrix++ uses a statistical score that conditions on the observed frequency of each mutation. In particular, if a gene is represented with two states (mutated or not), then the frequencies of each combination of states of *k* genes can be recorded in a 2 × 2 × … × 2 = 2^k contingency table. We derive an exclusivity statistic *T* that sums the counts in exclusive cells in the table and then perform an exact test (analogous to Fisher's exact test for independence of categories) by enumerating tables with larger values of the test statistic *T*. We define the score of a gene set as –log(*p*), where *p* is the p-value of the observed value of the test statistic. Dendrix++ generally has higher sensitivity than Dendrix to detect sets of mutations (or mutated genes) that are strongly exclusive, but are mutated at lower frequencies. The maximum scoring set of genes of a given size is identified using a Markov Chain Monte Carlo (MCMC) approach, as described in the Dendrix paper[30].

We run Dendrix++ in an iterative fashion to discover multiple sets of mutually exclusive mutations/mutated genes (**Figure S9 and S10**) and perform the permutation test by generating 1,000 random mutation datasets where the mutation frequencies of genes are preserved. The highest scoring set found by Dendrix++ contains eight genes whose mutations are perfectly exclusive. We refer to this set of genes as Group A. This gene set is composed of *NPM1*, one of the most frequently mutated genes in AML, with *TP53*, *RUNX1*, *MLL*-X fusions and four additional fusion genes. The Group A gene set covers 143 (71.5%) samples and is significant by a permutation test (*P* < 0.001). In the second iteration, we identified a significant (*P* < 0.021) gene set containing *FLT3*, *KRAS*/*NRAS*, other the groups "Tyrosine kinases" and "Other serine/threonine kinases". This set, Group B, covers 104 (52%) samples with only 7 samples having more than one mutation. In the third iteration Dendrix++ identified a gene set (*P* < 0.103) containing *ASXL1*, the cohesin complex, the group "Myeloid transcription factors", and the group "Other epigenetic modifiers". This set, Group C, covers 71 (35.5%) samples and contains only one co-occurrence. **Figure S10** shows the mutation matrix for each of these groups.

A.5.6. Patterns of pairwise co-occurrence and exclusivity among genes and gene groups

We examined pairwise co-occurrence and exclusivity among somatic mutations. Given a pair of genes, we construct a 2 × 2 contingency table:

$$T = \begin{bmatrix} x_{00} & x_{01} \\ x_{10} & x_{11} \end{bmatrix},$$

where x_{00} is the number of samples with no mutations in either gene, x_{11} is the number of samples with mutations in both genes, and x_{01} and x_{10} are the number of samples with a mutation in one gene or the other, respectively. We use the one-tailed Fisher's exact test to compute the statistical significance of mutual exclusivity (left tail) or co-occurrence (right tail) between mutations in the pair of genes..

We performed this analysis on two somatic mutation datasets. The first includes: 27 genes and gene groups (as defined in A.5.3) with at least 4 mutations, and two of the cytogenetic risk categories (intermediate and unfavorable). We did not include the favorable cytogenetic risk category because the fusion genes *PML-RARA*, *RUNX1-RUNX1T1*, and *MYH11-CBFB* define the category. We performed Fisher's exact test for each pair of genes, gene groups, and cytogenetic risk categories. We found 31 pairs among genes and gene groups with $P < 0.04$, and 4 pairs including cytogenetic risk categories with $P < 10^{-7}$. **Figure 2** and **Figure S8** summarize the results. We also list the pairs and their respective p-values in **Table S19**.

The second somatic mutation dataset includes: all genes or gene groups with at least 1 mutation and where metagenes that include at least one gene with more than 2 mutations were completely split, and the four cytogenetic risk categories. We used Fisher's one-tailed exact test: a pair of genes is significantly co-occurring or exclusive by Fisher's exact test if the value is less than 0.05 for the right or left tails, respectively. We found 115 statistically significant ($P < 0.05$) mutually exclusive or co-occurring pairs (**Table S20)**

A.6. mRNA and miRNA sequencing and analysis

A.6.1. Messenger RNA library construction and sequencing

Two micrograms of total RNA samples were arrayed into a 96-well plate and polyadenylated (PolyA+) messenger RNA (mRNA) was purified using the 96-well MultiMACS mRNA isolation kit on the MultiMACS 96 separator (Miltenyi Biotec, Germany) with on-column DNaseI-treatment as per the manufacturer's instructions. The eluted polyA+ mRNA was ethanol precipitated and resuspended in 10µL of DEPC treated water with 1:20 SuperaseIN (Life Technologies, USA). Double-stranded cDNA was synthesized from the purified polyA+ RNA using the Superscript Double-Stranded cDNA Synthesis kit (Life Technologies, USA) and random hexamer primers at a concentration of 5µM. The cDNA was quantified in a 96-well format using PicoGreen (Life Technologies, USA) and VICTOR^{3}V Spectrophotometer (PerkinElmer, Inc. USA). The quality was checked on a random sampling using the High Sensitivity DNA chip Assay (Agilent). The cDNA was fragmented by a Covaris E210 (Covaris, USA) for 55 seconds, using a Duty cycle of 20% and Intensity of 5. Plate-based libraries were prepared following the BC Cancer Agency, Genome Sciences Centre (BCGSC) paired-end (PE) protocol on a Biomek FX robot (Beckman-Coulter, USA). Briefly, the cDNA was purified in 96-well format using Ampure XP SPRI beads, and was subject to end-repair and phosphorylation by T4 DNA polymerase, Klenow DNA Polymerase, and T4 polynucleotide kinase respectively in a single reaction, followed by cleanup using Ampure XP SPRI beads and 3' A-tailing by Klenow fragment (3' to 5' exo minus). After cleanup using Ampure XP SPRI beads, picogreen quantification was performed to determine the amount of Illumina PE adapters used in the next step of adapter ligation reaction. The adapter-ligated products were purified using Ampure XP SPRI beads, then PCR-amplified with Phusion DNA Polymerase (Thermo Fisher Scientific Inc. USA) using Illumina's PE primer set, with cycle conditions of 98˚C 30sec followed by 10-15 cycles of 98˚C 10 sec, 65˚C 30 sec and 72˚C 30 sec, and then 72˚C 5min. The PCR products were purified using Ampure XP SPRI beads, and checked with a Caliper LabChip GX for DNA samples using the High Sensitivity Assay (PerkinElmer, Inc. USA). PCR products with a desired size range were purified using a 96-channel size selection robot developed at the BCGSC, and the DNA quality was assessed and quantified using an Agilent DNA 1000 series II assay and Quant-iT dsDNA HS Assay Kit using Qubit fluorometer (Invitrogen), then diluted to 8nM. The final concentration was verified by Quant-iT dsDNA HS Assay prior to Illumina HiSeq2000 PE 75 base sequencing.

A.6.2. MicroRNA library construction and sequencing

Two micrograms of total RNA per sample was arrayed into 96-well plates, as above, with controls as described below. From the poly(A) selection flow-through, small RNAs, including miRNAs, were recovered by ethanol precipitation. Flow-through RNA quality was checked for a subset of 12 samples using an Agilent Bioanalyzer RNA Nano chip.

miRNA-Seq libraries were constructed using a plate-based protocol developed at the BCGSC. Negative controls were added at three stages: elution buffer was added to one well when the total RNA was loaded onto the plate, water to another well just before ligating the 3' adapter, and PCR brew mix to a final well just before PCR. A 3' adapter was ligated using a truncated T4 RNA ligase2 (NEB Canada, cat. M0242L) with an incubation of 1 hour at 22^{o}C. This adapter is adenylated, single-strand DNA with the sequence 5' /5rApp/ ATCTCGTATGCCGTCTTCTGCTTGT /3ddC/, which selectively ligates miRNAs. An RNA 5' adapter was then added, using a T4 RNA ligase (Ambion USA, cat. AM2141) and ATP, and was incubated at 37^{o}C for 1 hour. The sequence of the single strand RNA adapter is 5'GUUCAGAGUUCUACAGUCCGACGAUCUGGUCAA3'.

When ligation was complete, first-strand cDNA was synthesized using Superscript II Reverse Transcriptase (Invitrogen, cat.18064 014) and RT primer (5'-CAAGCAGAAGACGGCATACGAGAT-3'). This was the template for the final library PCR, into which we introduced index sequences to enable libraries to be identified from a sequenced pool that contains multiple libraries. Briefly, a PCR brew mix was made with the 3' PCR primer (5'-CAAGCAGAAGACGGCATACGAGAT-3'), Phusion Hot Start High Fidelity DNA polymerase (NEB Canada, cat. F-540L), buffer, dNTPs and DMSO. The mix was distributed evenly into a new 96-well plate. A Biomek FX robot (Beckman Coulter, USA) was used to transfer the PCR template (first-strand cDNA) and indexed 5' PCR primers into the brew mix plate. Each indexed 5' PCR primer, 5'-AATGATACGGCGACCACCGACAGNNNNNNGTTCAGAGTTCTACAGTCCGA-3', contains a unique six-nucleotide 'index' (shown here as N's), and was added to each well of the 96-well PCR brew plate. PCR was run at 98°C for 30 sec, followed by 15 cycles of 98°C for 15 sec, 62°C for 30 sec and 72°C for 15 sec, and finally a 5 min incubation at 72^{o}C. Quality was checked across the whole plate using a Caliper LabChipGX DNA chip. PCR products were pooled, then size-selected to remove larger cDNA fragments and smaller adapter contaminants, using the 96-channel automated size-selection robot noted above. After size-selection, each pool was ethanol precipitated, quality checked using an Agilent Bioanalyzer DNA1000 chip and quantified using a Qubit fluorometer (Invitrogen, cat. Q32854). Each pool was then diluted to a target concentration for cluster generation and was loaded into a single lane of an Illumina GAIIx or HiSeq 2000 flow cell. Clusters were generated, and lanes were sequenced with a 31-bp main read for the insert and a 7-bp read for the index.

A.6.3. Alignment and coverage analysis of RNA-seq data

Using BWA version 0.5.7[31], we aligned chastity-passed reads to an extended human reference genome consisting of hg18/GRCh36 plus exon junction sequences constructed from all known transcript models in RefSeq, EnsEMBL and UCSC genes, as described[32]. We used default BWA parameter settings but disabled Smith-Waterman alignment. After alignment, the reads that aligned to exon junctions were repositioned in the genome as large-gapped alignments, using repositioning software developed in-house. We removed adapter dimer sequences and soft-clipped reads that contained adapter sequences. The unambiguously aligned, filtered reads were then analyzed by in-house gene coverage analysis software to calculate the coverage over the total collapsed exonic regions in each gene as annotated in EnsEMBL (version 59), and RPKM values[33] were calculated to represent the normalized expression level of exons and genes.

A.6.4. Preprocessing, alignment and annotation of miRNA

Briefly, the sequence data were separated into individual samples based on the index read sequences, read quality was assessed, adapter sequences were trimmed, and trimmed reads were aligned to the NCBI GRCh36 reference genome. Below we describe these steps in more detail.

For routine sequence quality checks (QC), a subset of raw sequences from each pooled lane were taken to assess the abundance of reads from each indexed sample in the pool, the proportion of reads that may originate from adapter dimers (i.e. a 5' adapter joined to a 3' adapter with no intervening sequence) and for the proportion of reads that map to miRBase human miRNA sequences. Sequencing error is estimated by a method originally developed for SAGE[34]. Libraries that pass this QC stage are preprocessed for alignment. While the size-selected miRNAs vary somewhat in length, they are typically ~21 bp long, and so are shorter than the 31-bp read length; given this, each read sequence extends some distance into the 3' sequencing adapter. Because this non-biological sequence can interfere with aligning the read to the reference genome, 3’ adapter sequence is identified and removed (trimmed) from a read. The adapter-trimming algorithm identifies as long an adapter sequence as possible, allowing a number of mismatches that depends on the adapter length found. A typical sequencing run yields several million reads; using only the first (5’) 15 bases of the 3’ adapter in trimming makes processing efficient, while minimizing the chance that an miRNA read will match the adapter sequence.

The algorithm first determines whether a read sequence should be discarded as an adapter dimer by checking whether the 3’ adapter sequence occurs at the start of the read. For reads passing this stage, the algorithm then tries to identify an exact 15-bp match anywhere within the read sequence. If it cannot, it retries, starting from the 3' end, and allowing up to 2 mismatches. If the full 15bp is not found, decreasing lengths of adapter are checked, down to the first 8 bases, allowing one mismatch. If a match is still not found, from 7 bases down to 1 base is checked, with an exact match required. Finally, the algorithm will trim 1 base off the 3’ end of a read if it matches the first base of the adapter. This is based on two considerations. First, it is preferable to get a perfect alignment than an alignment that has a potential one-base mismatch. Second, if only 1 base of adapter is found in the read sequence, the read is likely too long to be from a miRNA and the effect of the trimming on its alignment would not affect this sample’s overall miRNA profiling result.

After each read has been processed, a summary report is generated that contains the number of reads at each trimmed read length. Because the shortest mature human miRNA in miRBase v16 is 15 bp, any trimmed read that is shorter than this is discarded; remaining reads are submitted for alignment to the reference genome. BWA alignment(s) for each read are checked with a series of three filters: a) a read with more than 3 alignments is discarded as too ambiguous; b) for TCGA quantification reports, only perfect alignments with no mismatches are used; c) based on comparing expression profiles of test libraries (data not shown), reads that fail the Illumina basecalling chastity filter are retained, while reads that have soft-clipped CIGAR strings are discarded.

For reads retained after filtering, each coordinate for each read alignment is annotated using the reference databases in **Table S21**, and requiring a minimum 3-bp overlap between the alignment and an annotation. In annotating reads we address two potential issues. First, a single read alignment can overlap feature annotations of different types; second, a read can have up to three alignment locations, and each alignment location can overlap a different type of feature annotation. We resolve the first issue by using heuristically determined priorities to assign a single annotation to each alignment. We resolve the second by collapsing multiple annotations to a single annotation, as follows.

If a read has more than one alignment location, and the annotations for these are different, we use the priorities from **Table S21** to assign a single annotation to the read, as long as only one alignment is to a miRNA. When there are multiple alignments to different miRNAs, we flag the read as cross-mapped[35], and preserve all of its miRNA annotations while discarding all of its non-miRNA annotations. This approach ensures that all annotation information about ambiguously mapped miRNAs is retained, and allows annotation ambiguity to be addressed in downstream analyses. Note that we consider miRNAs to be cross-mapped only if they map to different miRNAs, not to functionally identical miRNAs that are expressed from different locations in the genome. Such cases are indicated by miRNA miRBase names, which can have up to 4 separate sections separated by "-", e.g. hsa-mir-26a-1. A difference in the final (e.g. '-1') section denotes functionally equivalent miRNAs expressed from different regions of the genome, and we consider only the first 3 sections (e.g. 'hsa-mir-26a') when comparing names. As long as a read maps to multiple miRNAs for which the first 3 sections of the name are identical (e.g. hsa-mir-26a-1 and hsa-mir-26a-2), it is treated as if it maps to only one miRNA, and is not flagged as cross-mapped.

From the profiling results for a tumor type, for a minimum of approximately 100 samples, we identify the depth of sequencing required to detect the miRNAs that are expressed in a sample by considering a graph of the number of miRNAs detected in a sample as a function of the number of reads aligned to miRNAs (data not shown). For the current work, a library from a sequenced pool was required to have at least 750,000 reads mapped to miRBase annotations. For any sequencing run that fails to meet this threshold, we sequence the sample again to achieve at least the minimum number of miRNA-aligned reads.

Finally, for each sample, the reads that correspond to miRNAs are summed and normalized to a million miRNA-aligned reads to generate the quantification files that are submitted to the DCC. Two quantification files are submitted. A 'precursor' file reports miRNA abundance for pre-miRNAs, while an 'isoform' file reports mature and star strands separately (i.e. 5p and 3p strands), and includes information on variable 5' and 3' read alignment locations, which can reflect isoforms, adapter trimming and RNA degradation.

A.6.5. Gene fusion detection and verification

RNA-seq libraries were assembled with ABySS (version 1.3.2 - http://www.bcgsc.ca/platform/bioinfo/software/abyss/releases) using k-mer values of 26 to 50 as previously described[36]. The contigs from assemblies were filtered, merged, aligned and post-processed using the Trans-ABySS pipeline (version 1.3.2 - http://www.bcgsc.ca/platform/bioinfo/software

/trans-abyss/releases)[37]. Contigs were aligned against the reference genome using BLAT with the following parameters: stepSize=5, repMatch=2253, minScore=0 and minIdentity=0. Potential fusion candidates were identified as contigs that could not be mapped to a single unique location for at least 95% of the sequence. Such contig alignments with the following characteristics were considered candidates: both alignments have percentage of identity of at least 98%, one of the alignments does not reside entirely within its partner in terms of genomic coordinates, the alignments do not overlap by more than 5% in terms of contig coordinates, the alignments do not overlap in terms of genome coordinates, and the total coverage of the two alignments, in terms of contig sequence, is at least 90%. To further filter the candidate events from contig alignments, we use alignment of sequence reads to both contigs and genome. Reads were aligned to the contigs using BWA 0.5.9-r16. As noted above, reads were aligned to the reference genome with exon-exon junctions using BWA 0.5.7, and reads that mapped across exon junctions were repositioned to their original genomic positions. Candidate fusion cases were then filtered by requiring at least 2 reads spanning the contig breakpoint with at least 4 flanking base pairs on either side, and at least 4 read pairs flanking the genomic breakpoint and pointing towards each other.

PCR primers were designed to flank the gene fusion breakpoints in the ABySS-assembled sequence contigs, and were used to amplify cDNA prepared from 100ng of total RNA for each sample with Accuscript High Fidelity reverse transcriptase (Agilent) (SuperScript II (Invitrogen/Life) had been used for library construction[38]). Successful PCR amplicons were purified then confirmed by Sanger capillary sequencing.

A.6.6. Partial and internal tandem duplications

Partial and internal tandem duplications were reported from the RNA-seq data by Barnacle v0.1.2 (Swanson et al. submitted). http://www.bcgsc.ca/platform/bioinfo/software/barnacle). The Barnacle pipeline has five stages.
1) Examine contig-to-genome alignments and identify anomalous or non-reference (candidate) contigs. These can have a variety of alignment topologies. 2) Examine transcriptome read alignments to assembled contig sequences and calculate read support for the candidate contigs. 3) Apply user-specified filters to the candidate contigs and retain only sufficiently confident candidates. 4) Identify PTDs, ITDs, and fusions from the filtered candidates. 5) Compare the coverage of the predicted chimeric transcripts to their corresponding wild-type transcripts.

A.6.7. Unsupervised consensus clustering

For mRNA-seq data, we removed genes expressed at or below a noise threshold of RPKM≤0.2 in at least 75% of samples, then identified the most-variant 25% of genes (N = 1728) by ranking expressed genes having a median RPKM of at least 10 by the coefficient of variation.

For miRNA-seq data, read count data for 187 tumor samples were extracted from Level 3 data archives on the TCGA Data Portal website (tcga.cancer.gov/dataportal). The set of isoform.quantification.txt files, which give read counts at base pair resolution, was processed to report total read counts for mature and star strands (corresponding to miRBase v13 MIMAT identifiers), and read counts for each sample were normalized to RPM, i.e. to reads per million reads aligned to miRBase mature or star strands. Strands corresponding to miRNAs that had been removed from v18 miRBase (miRNA.dead) were eliminated from the data matrix. Mature and star strands were ranked by RPM variance across the samples, and the most variant 25% (214 MIMATs) were retained.

For both RNA-seq and miRNA seq data, we generated unsupervised consensus clustering results with NMF v0.5.02 or v0.5.06[39] in R v2.12.0, with the default Brunet algorithm, and 50 and 200 iterations respectively for the rank survey and clustering runs. A preferred cluster result was selected by considering profiles of cophenetic score and average silhouette width[40]of the consensus membership matrix, for clustering solutions having between 3 and 15 clusters (data not shown). Silhouette results were generated from the NMF consensus membership matrix using the R 'cluster' package v1.14.1. Silhouette width profiles were generated by reordering samples to match the sample order in the NMF heatmap, and typical vs. atypical members were identified for each unsupervised group using a silhouette width threshold set to a fraction (e.g. 0.90) of the maximum width in that group.

Given NMF outputs for mRNA-Seq and miRNA-Seq data, we generated abundance heatmaps from the clustering input matrices as follows. We identified the top-ranked 20% of genes and 30% of microRNAs mature and star strands from the respective NMF metagene (i.e. W matrix) output files. Removing duplicate names resulted in 980 gene symbols and 106 miRNAs; we filtered the RNA-seq RPKM matrix and miRNA-seq RPM matrix to retain only records for these genes and miRNAs. We reordered columns in each matrix into the NMF output order for each data type. Finally, we used Cluster v3.0[41] (bonsai.hgc.jp/~mdehoon/software/cluster) to log-transform and median-center each row, then to reorder rows using hierarchical clustering with an centered correlation distance metric and complete linkage.

P-values for binomial and multinomial tests for association of covariates with cluster assignments, and of clustering assignments with each other across data types, were compiled en masse and adjusted using the R function p.adjust(), which implements Benjamini & Hochberg's sequential multiple comparisons adjustment for p-values. The significance of all 1383 comparisons was adjusted simultaneously, and results that remained significant at an FDR of 0.05 were retained for cross-datatype comparison and association. Univariate Kaplan Meier curves and p-values for overall survival were calculated with the R 'survival' package v2.36-12.

Relationships between sample order for different clustering solutions were visualized using Bezier curve graphics generated with a custom Mathematica v8 notebook (Wolfram Research, Champaign IL).

Asymptotic association p-values for covariate contingency tables were calculated using R's chi-square test.

A.6.8. Comparison of RNAseq and Microarray data

RNAseq gene-level RPKM data for 179 patients was log2 transformed and genes median centered. Data was filtered for genes that were present on at least 80% of samples (20,442 genes filtered to 15,848). ClaNC, a nearest centroid-based classification algorithm, was used to find signatures of each class[42]. 160 genes per class were selected based on a low overall cross validation rate for the NMF 7-class distinction. 178 of the samples were also run on Affymetrix microarrays (U133). Affymextrix data was normalized by MAS5 and the expression value was calculated by mapping the microarray probes to genes and taking the maximum value for each gene. Microarray data was log2 transformed and genes were median centered similar to the RNAseq data. 1073 of the 1120 genes were also identified in the microarray data. For visualization, the 178 samples and 1073 genes present on both platforms were used. In the RNAseq data, genes were hierarchically clustered, while maintaining the order of the subtype classes for samples. Microarray data were visualized by maintaining the same gene and sample order as in the RNAseq data (**Figure S14**).

A.6.9. Group-discriminatory genes and miRNAs

For each unsupervised sample group we identified discriminatory genes for mRNA-seq data, and discriminatory mature and star strands for miRNA-seq data, by generating a random forest classifier for samples in that group vs. all other samples[43], using R v2.15.1 and randomForest v4.6-6. Typically we used 50000 trees, and either mtry=100 or an optimized value of mtry, and a Gini variable importance. For each classifier, we profiled the estimated out-of-bag (OOB) estimated error as a function of the number of most-important genes or miRNA strands, and reported the smallest set of genes or miRNA strands that minimized or substantially minimized the OOB error. For miRNA-seq data, we note that only discriminatory mature and star strands that have substantial absolute abundances are likely to be biologically influential[44].

A.6.10. Associations among cytogenetic risk groups, genetic alterations, and mRNA/miRNA clusters

As a TCGA marker paper, the work reported here makes a large number of different genomic data types available to the community as a resource for ongoing work, and reports initial analyses. In analyzing the transcriptome sequencing data, we used unsupervised consensus clustering to identify groups of samples that had related abundance profiles from RNA-seq data (178 samples) and miRNA-seq data (187 samples). We then identified statistically significant associations between the groups and FAB subtypes, cytogenetic risk groups, and molecular alterations (mutations, fusions, other chimeric transcript) (**Figure 5**). We compared unsupervised groups from RNA-seq data to published microarray data. Finally, we identified discriminatory genes and miRs for each unsupervised group, taking all other samples as the second group. A gene or miR that helps a classifier distinguish or discriminate a group of samples from all other samples tends to have a relatively high or low abundance in samples in that group.

Work reported in the literature typically took different approaches than the above, for example: a) identifying genes/miRs that discriminate specific cytogenetic subtypes, driver gene mutations, or cytogenetic risk categories, e.g. Verhaak *et al.*[45]; identifying prognostic genes or miRNAs[46-49]; comparing the abundance levels of tumor and normal cells[50,51]; integrating gene and miRNA abundance to infer potential functional relations[52]; and comparing different types of leukemia, e.g. AML vs. ALL[53].

Given the above, comparing our results with published results is most meaningful for RNA-seq or miRNA-seq-based groups that have strong, statistically significant associations with a mutation in a known driver gene or with a cytogenetic subtype. There are two such cases in the work reported here: a) RNA-based group 3 and miRNA-based group 5 were strongly associated with FAB M3, and *PML*/*RARA* fusions; and b) miR-based group 3 was strongly associated with *NPM1*, *FLT3* and *DNMT3A* mutations. These mutations were distributed across four RNA-seq groups (1, 4, 5 and 7).

In our results (Gene expression analysis), we compare our results to the literature for *NPM1* mutations. For miRNA-based group 3 and NPM1 mutations, we have added text (see above) that indicates that, beyond miR-10a, miR-242, miR-196b, miR-130a and let-7b were also discriminatory (Fig S15j), consistent with (Bryant)[50,52,54].

In the work reported here, *NPM1* mutations were statistically associated with RNA-seq group 4, but were also present in groups 1, 5 and 7. *NPM1* mutations have been reported as associated with a discriminative HOX and TALE gene expression signature[55], and three HOX genes and *PBX3* have been associated with overall survival (Li Blood 2012).

Possibly because *NPM1* mutations were distributed across four RNA-seq groups, no HOX, PBX, MEIS or PREP gene[56] was discriminatory for any RNA-seq group.

Li *et al*.[46] reported a prognostic 24-gene signature (*ALS2CR8, ANGEL1, ARL6IP5, BSPRY, BTBD3, C1RL, CPT1A, DAPK1, ETFB, FGFR1, HEATR6, LAPTM4B, MAP7, NDFIP1, PBX3, PLA2G4A, PLOD3, PTP4A3, SLC25A12, SLC2A5, TMEM159, TRIM44, TRPS1,* and *VAV3*) from a meta-analysis of six independent AML datasets. The current work did not assess prognostic genes, and none of the 24 genes was discriminatory for an RNA-seq group, though two SLC25A family genes and an SLC2A family member were discriminatory for group 2, and this group was statistically associated with *TP53* mutations.

Of other overlaps with the literature that suggest that the data made available with this manuscript will be a valuable resource for ongoing work, we note two.

CD34 and *BAALC* were relatively abundant in RNA-seq group 6 (**Figure S16f**). Consistent with Rockova *et al.*[48] and Mendler *et al.*[57], this group was statistically associated with *RUNX1* mutations, was free of *NPM1* mutations (**Figure 5**), and had relatively unfavorable overall survival (**Figure S13**).

The mature strand of the prognostic miR-181a[58] was a relatively weak (11th) discriminator for miR groups 4 (weakly statistically associated with *PICALM-MLLT10* fusions) and 2. miR group 2 is statistically associated with *TP53* mutations, and Seoudi *et al.*[59] noted that *TP53* may regulate miR-181a.

A.7. Methylation array and analysis

A.7.1. Bisulfite Conversion

DNA samples (1 µg) were bisulfite converted using the Zymo Research EZ96 DNA methylation kit (Zymo Research, Irvine, CA, USA) as described by the manufacturer. We determined the completeness of bisulfite conversion and the amount of bisulfite-converted DNA for each sample using a panel of four MethyLight-based quality control (QC) reactions as described previously[60]. All samples passed these QC tests and subsequently were entered into the Illumina Infinium DNA methylation data production pipeline.

A.7.2. Illumina Infinium DNA methylation profiling

The Illumina Infinium 450k DNA methylation assay (HumanMethylation450) was performed for all 192 samples according to the experimental protocol outlined by the manufacturer (Illumina, San Diego, CA, USA). The assay interrogates the DNA methylation status of 482,421 CpG dinucleotides in the promoter and gene body regions of all human RefSeq genes, along with numerous intergenic regions. The DNA methylation score for each locus is presented as a beta value (beta = (M/(M+U)) in which M and U indicate the mean methylated and unmethylated signal intensities for each locus. Beta values range from zero to one, with scores of zero indicating low levels of DNA methylation and scores of one indicating high levels of DNA methylation. A detection *P*-value also accompanies each data point and compares the signal intensity difference between the analytical probes and a set of negative control probes on the array. Any data point with a corresponding *P*-value greater than 0.05 is deemed not to be statistically significantly different from background and is thus masked as “NA” in TCGA level 3 data packages. Genomic locations for each probe are available from Illumina (www.illumina.com).

A.7.3. TCGA Data Packages

DNA methylation data packages were generated using the ‘EGC.tools’ R package (version 1.3.0) after processing raw IDAT files (available as Level 1 data packages) for each sample with the ‘methylumi’ R package (version 2.3.22). Background correction and adjustment to equalize the red/green dye bias across samples was then performed, generating the processed signal intensities available as TCGA Level 2 data packages. Additionally, for downstream analysis, repeat regions (via RepeatMasker) and common SNPs (MAF > 0.01, per dbSNP build 135 via the UCSC snp135common track) expected to substantially interfere with probe specificity were identified, and measurements from the affected probes were masked as “NA”. Loci on chromosomes X and Y were also masked, after verifying the clinically annotated gender via probes on chromosome X, to avoid confounding by sex.

A.7.4. Unsupervised analysis

The observed/expected ratio for CpG dinucleotides at each locus was computed for a 3000-bp window, centered on the interrogated locus, for each probe on the HumanMethylation450 platform. The hg19 reference sequence for each window was extracted from the ‘BSgenome.Hsapiens.UCSC.hg19’ Bioconductor package[61], and the observed proportion of CG dinucleotides as a fraction of the total was computed. The expected proportion of CG dinucleotides as a fraction of the total is $\mathrm{Pr}C\mathrm{Pr}(G)2$, where Pr(C) is the fraction of cytosine bases in the window, Pr(G) is the fraction of guanine bases in the window, and CG is assumed to be equally likely to GC in calculating the expectation. The window size was chosen after Saxonov et al.[62], who distinguished two different classes of gene promoters in the human genome based on their CpG content via this metric.

Approximately 5% of the loci on the HumanMethylation450 platform (exclusive of SNPs, repeat regions, and sex chromosomes) have an observed-to-expected CpG ratio of 0.316 or lower. Approximately 5% of the loci on the platform have an observed-to-expected CpG ratio of 1.57 or higher. When replicate samples from a separate experiment were added, we observed stable and consistent clustering using these cutoffs to define the extremes of CpG density. We chose the 1000 most variable loci from the highest density 5% and the lowest density 5% of loci, and performed hierarchical clustering separately for each group of probes, using Ward's method for linkage and a Euclidean distance metric. Cluster assignments, for comparison with miRNA and mRNA expression cluster assignments, were generated by cutting the resulting dendrogram at progressively greater heights and assessing cluster stability across 1000 bootstrap iterations. For high-CpG clusters, which were relatively unstable, we cut the dendrogram to produce a 7-cluster solution. For low-CpG clusters, we found the 9-cluster solution to produce stable associations, and observed rough correspondence to the same patterns in when clustering at a subset of 4000 randomly designed probes included on the HumanMethylation450 assay by the manufacturer, suggesting that the low-CpG clusters are in fact more representative of changes to the genome as a whole among the patients in the study.

We performed Fisher's exact test to quantify associations between recurrent abnormalities and clustering assignments across microRNA, mRNA, high-CpG methylation, and low-CpG methylation results, adjusting p-values by the Benjamini-Hochberg procedure[63],retaining those with adjusted p-values less than 0.05 for annotation purposes (**Supplementary Tables 14-16**).

A.7.5. Supervised analyses

Marginal tests for significant differences in DNA methylation between mutant and wild-type patients were conducted separately for recurrent fusions and mutations, on logit-transformed and untransformed beta values (β = (M/(M+U), as previously). Additionally, we tabulated differences between AML patient blasts and CD34+ cells from healthy donors, as well as differences between AML patient blasts and committed myeloid cells (promyelocytes, neutrophils, and monocytes) from donors. Student's *t* test was applied row-wise across all loci remaining after masking (previously described). The resulting p-values were corrected using the Benjamini-Hochberg procedure. The subset of loci that exhibited statistically significant absolute differences in mean methylation greater than 10% were enumerated for each contrast. We then tabulated contiguous regions of greater than 1000 bases with significant (FDR < 0.1) changes in DNA methylation among those enumerated, to identify and annotate differentially methylated regions for each contrast.

A.8 Batch effects analysis

We used hierarchical clustering and Principal Components Analysis (PCA) to assess batch effects in the AML data sets. Three different data sets were analyzed: mRNA expression (RNA-seq Illumina GA), miRNA expression (RNA-seq Illumina GA), and DNA methylation (Illumina Infinium HumanMethylation450 microarray). All of the data sets were at TCGA level 3, since that is the level at which most of the analyses in the paper are based. We assessed batch effects with respect to the variable plate ID, which contained the identification number of the plate on which the samples were shipped to the processing centers.

For hierarchical clustering, we used the average linkage algorithm with 1 minus the Pearson correlation coefficient as the dissimilarity measure. We clustered the samples and then annotated them with colored bars shown at the bottom. Each color corresponded to a plate ID. For PCA, we plotted the first four principal components, but only plots of the first two components are shown in the Supplementary Figures. To make it easier to assess batch effects, we enhanced the traditional PCA plot with centroids. Points representing samples with the same plate ID were connected to the batch centroid by lines. The centroids were computed by taking the mean across all samples in the batch. That procedure produced a visual representation of the relationships among batch centroids in relation to the scatter within batches. The results for the three data sets follow:

A.8.1 mRNA (RNA-seq Illumina GA)

Figure S19 shows clustering and PCA plots for the RNA-seq platform. Genes with zero values were removed and the RPKM values were log2-transformed before generating the figures. None of the batches were distinct from the others, indicating that no serious batch effects were present.

A.8.2 miRNA Expression (miRNA-seq Illumina GA)

Figure S20 shows clustering and PCA plots for miRNA-seq data. Genes with zero values were removed and the read counts were log2-transformed before generating the figure. While hierarchical clustering did show a small cluster of samples belonging to plate 0740, the cluster was not distinct in PCA plots. We therefore did not perform a batch effects correction. None of the other batches were outliers.

A.8.3 DNA Methylation (Illumina Infinium HumanMethylation27 and HumanMethylation450 arrays)

Figure S21 shows clustering and PCA plots for the Infinium DNA methylation platform. None of the batches were distinct from the others, indicating that no serious batch effects were present.

A.8.4 Batch Effects Conclusions

Overall, the TCGA AML data sets contained no major batch effects by clustering or PCA plots. In the miRNA expression data, a small cluster of samples from the same plate ID was observed in the hierarchical clustering plot, but not in the PCA plots. This did not warrant a batch effects correction, since the correction algorithms run the risk of removing important biological variation (as well as technical variation). We concluded that technical batch effects in the data sets are reasonably small and unlikely to influence high-level analyses in an important way.

B. Supplementary Figures

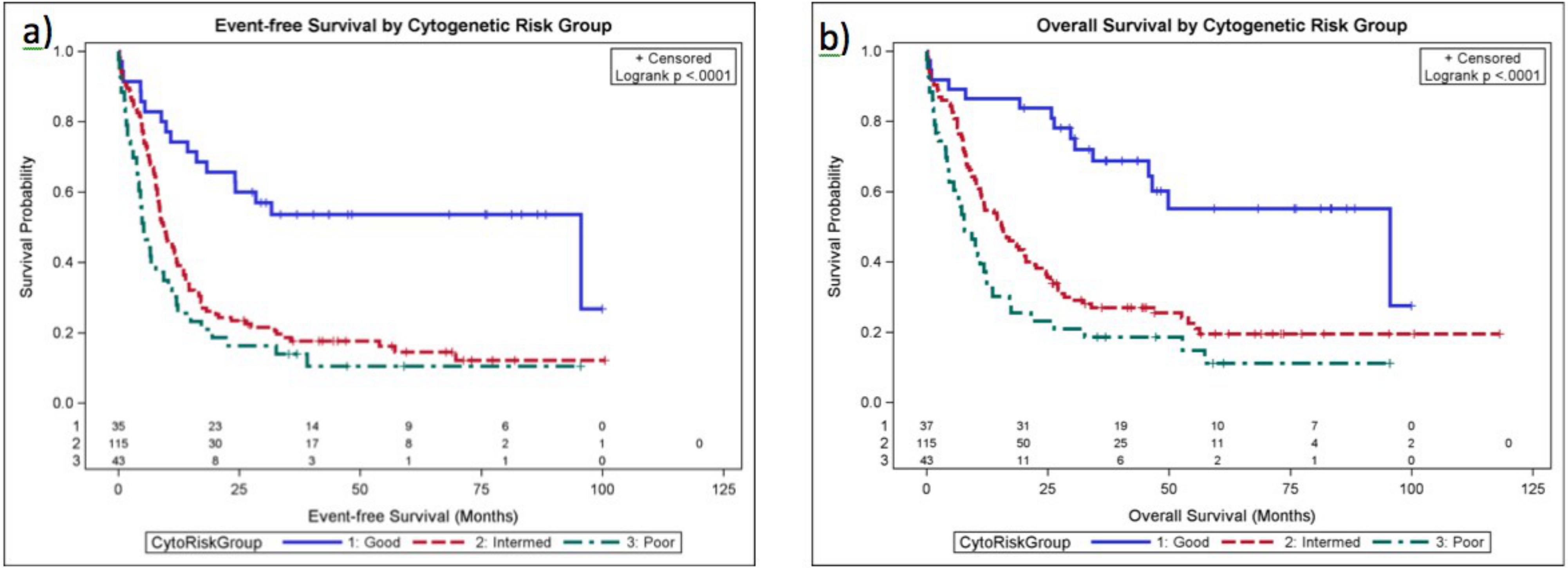


Figure S1. Survival by Cytogenetic Risk Group

Event-free (a) and Overall (b) Survival in the cohort, stratified by cytogenetic risk.

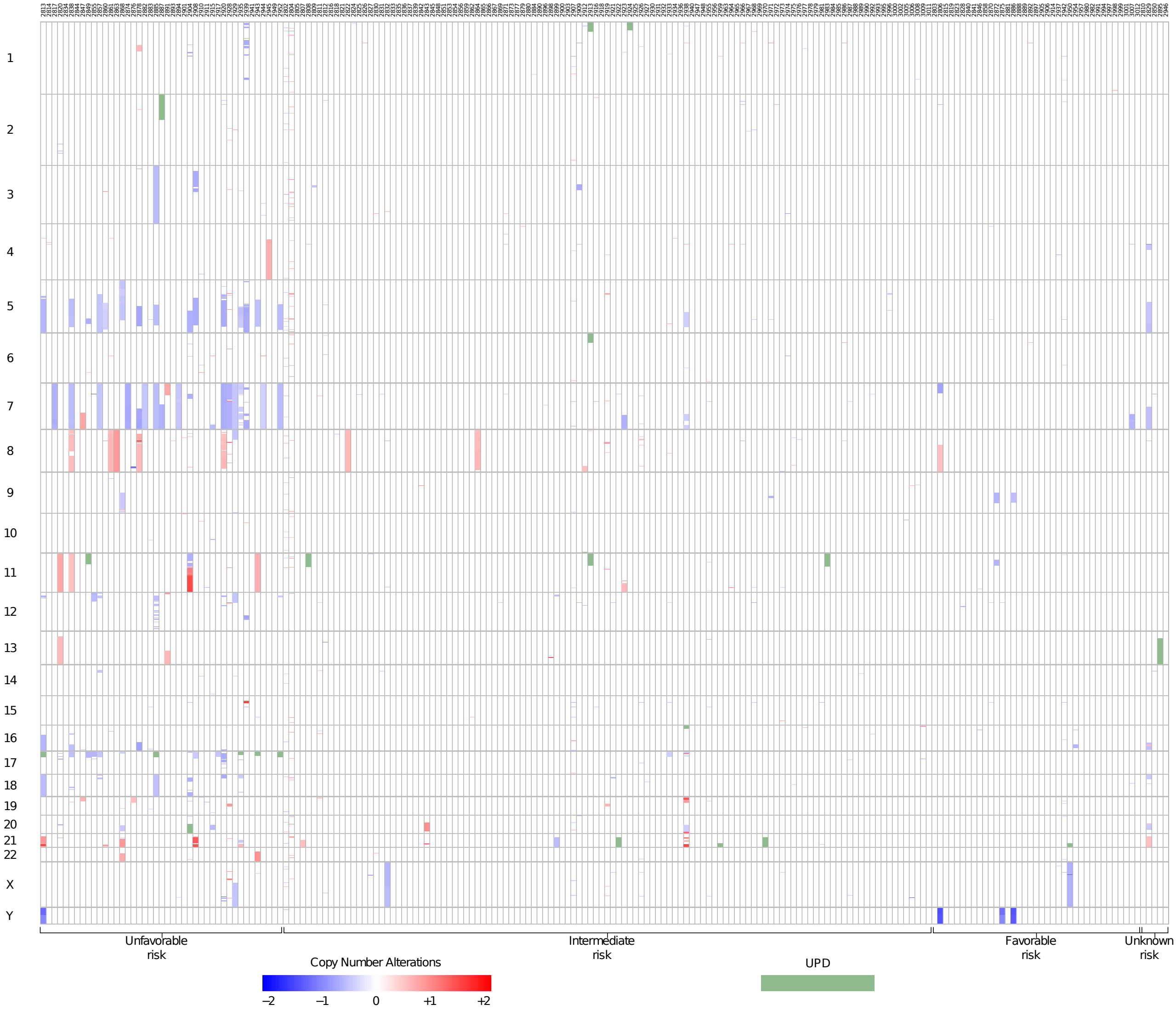


Figure S2. Copy number alterations and Uniparental Disomy events
Copy number alterations and Uniparental disomy events detected using SNP array data. Patients are grouped by cytogentic risk. Those with Intermediate or Favorable risk have almost no copy-number alterations, while unfavorable risk patients show recurrent loss of chromosomes 5 and 7, as well as recurrent UPD on chromosome 17p, which affects the TP53 locus.

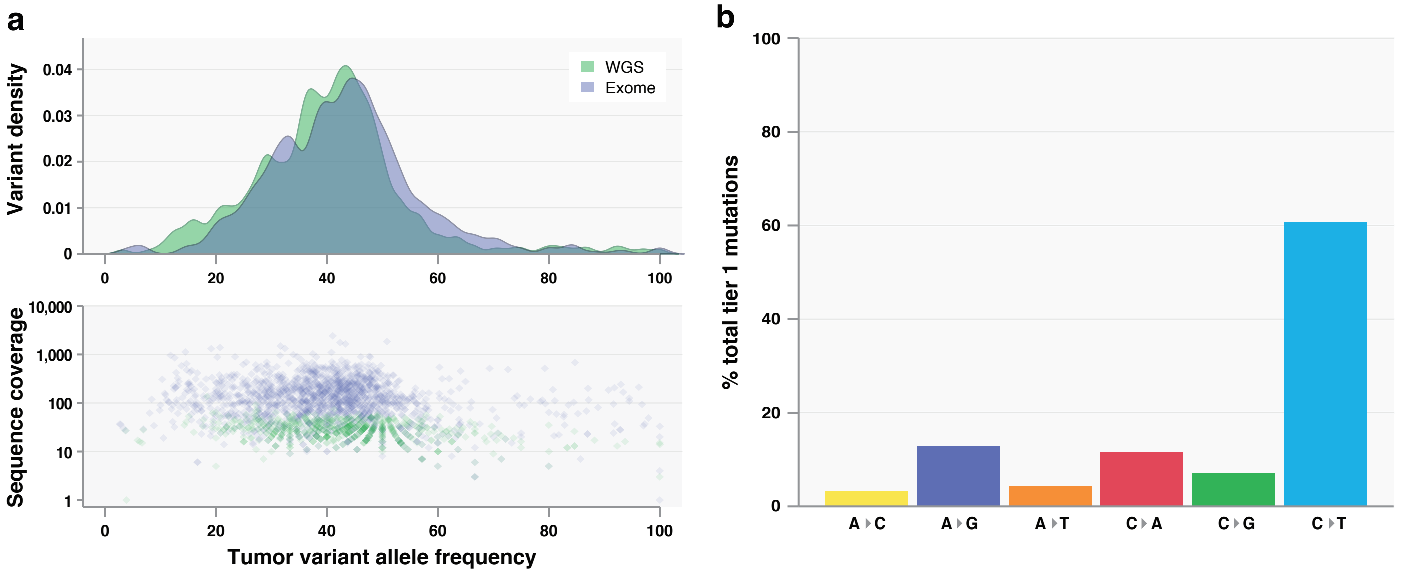


Figure S3: Variant frequency and mutation spectrum

a) Top panel: Variant density for all mutations validated with WGS (n=50, green) or exome sequencing (n=150, purple). Bottom panel: sequencing coverage for all variants. Note that few mutations with VAFs below 10% were validated. b) Mutational spectrum for all validated tier 1 mutations from all cases.

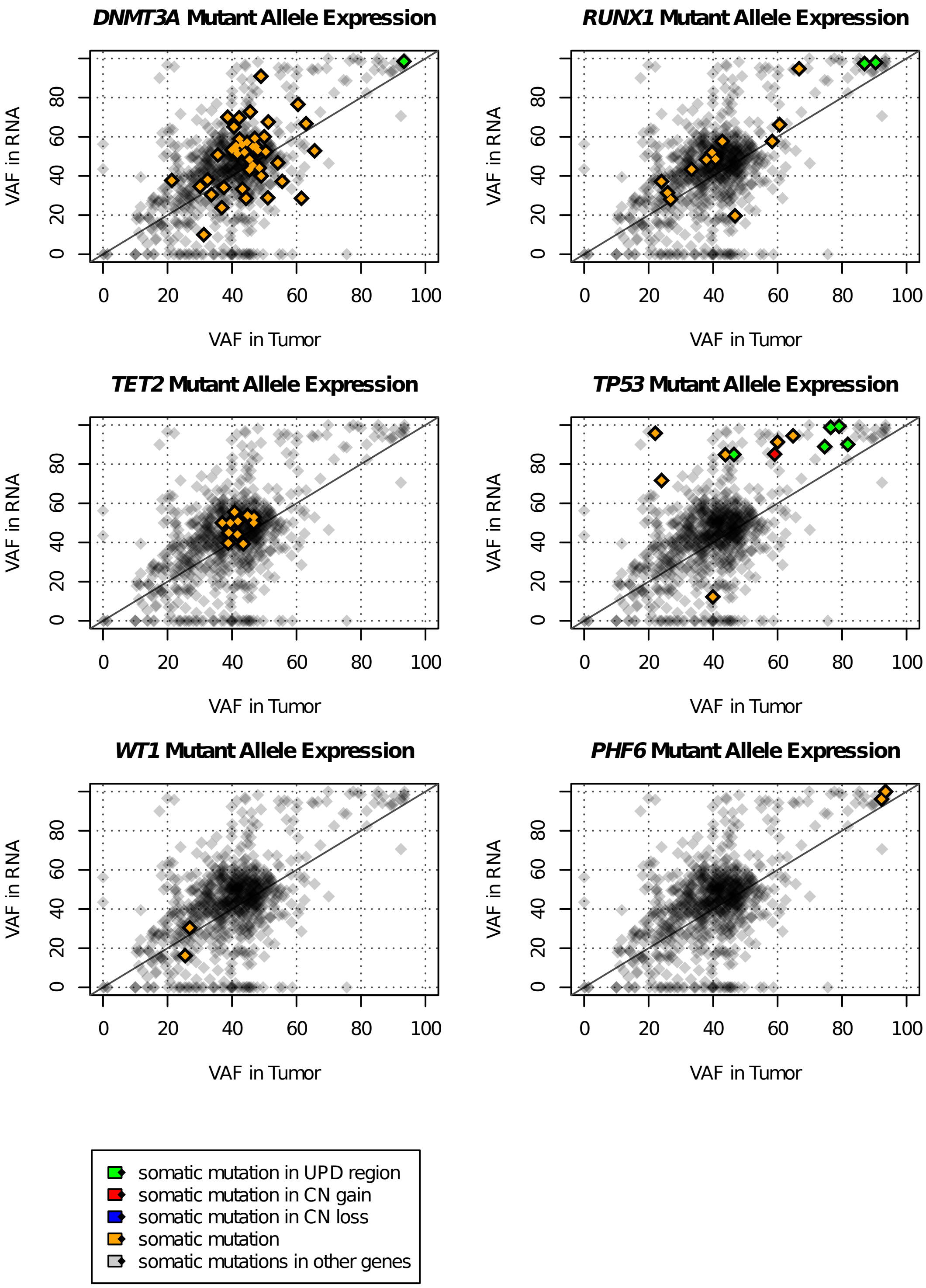


Figure S4. Allelic Expression Bias

Comparing variant allele frequencies between DNA and RNA reveals allelic bias in the expression of mutations in some putative tumor suppressor genes. Only SNVs are represented in the analysis. Copy number events and loss of heterozygosity explain some, but not all of this enrichment for expression of the mutant allele. Only mutations with at least 10x coverage in the RNA sequencing are shown. Only 2/12 *WT1* mutations are represented here, because most of the mutations were indels and did not yield accurate DNA readcounts; some of the SNVs also did not have adequate coverage by RNA-Seq. All six *PHF6* mutations occurred in male patients; since the *PHF6* gene is on the X chromosome, all mutations in this gene are therefore hemizygous. Only 2 of the 6 mutations had adequate readcounts for the analysis; the others may have been affected by nonsense-medicated decay.

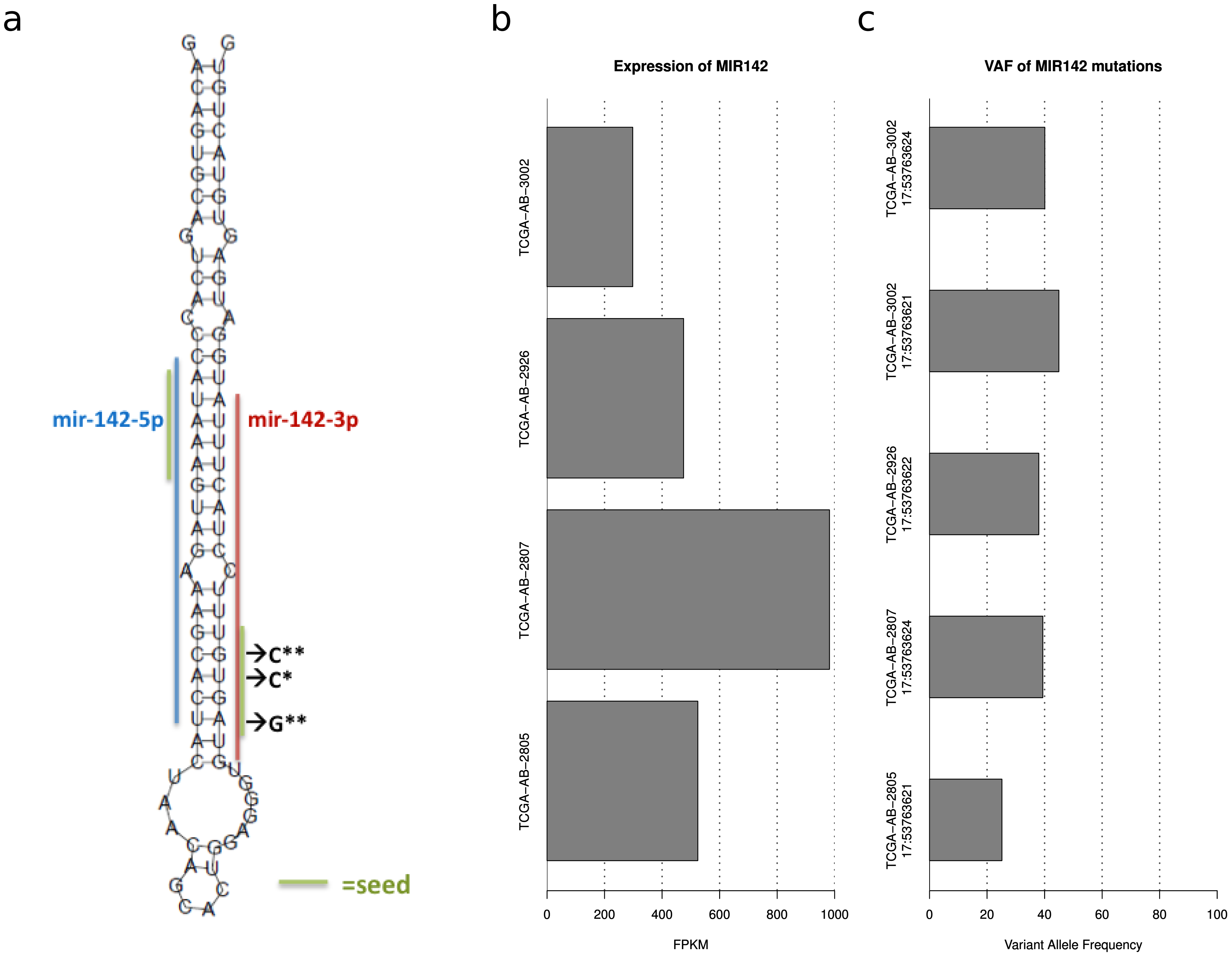


Figure S5. MIR142 mutations and expression
a) The secondary structure of MIR142, with the 5' product labelled in blue, the 3' product in red, and the seed regions in green. The locations of mutations are designated. Asterisks denote the number of mutations detected at each position. b) Expression levels (expressed as FKPM) of MIR142 derived from miRNA sequencing data. c) The variant allele frequencies of MIR142 mutations in the tumor data are shown for each affected case. The variant allele was expressed in all cases where mutations were detected.

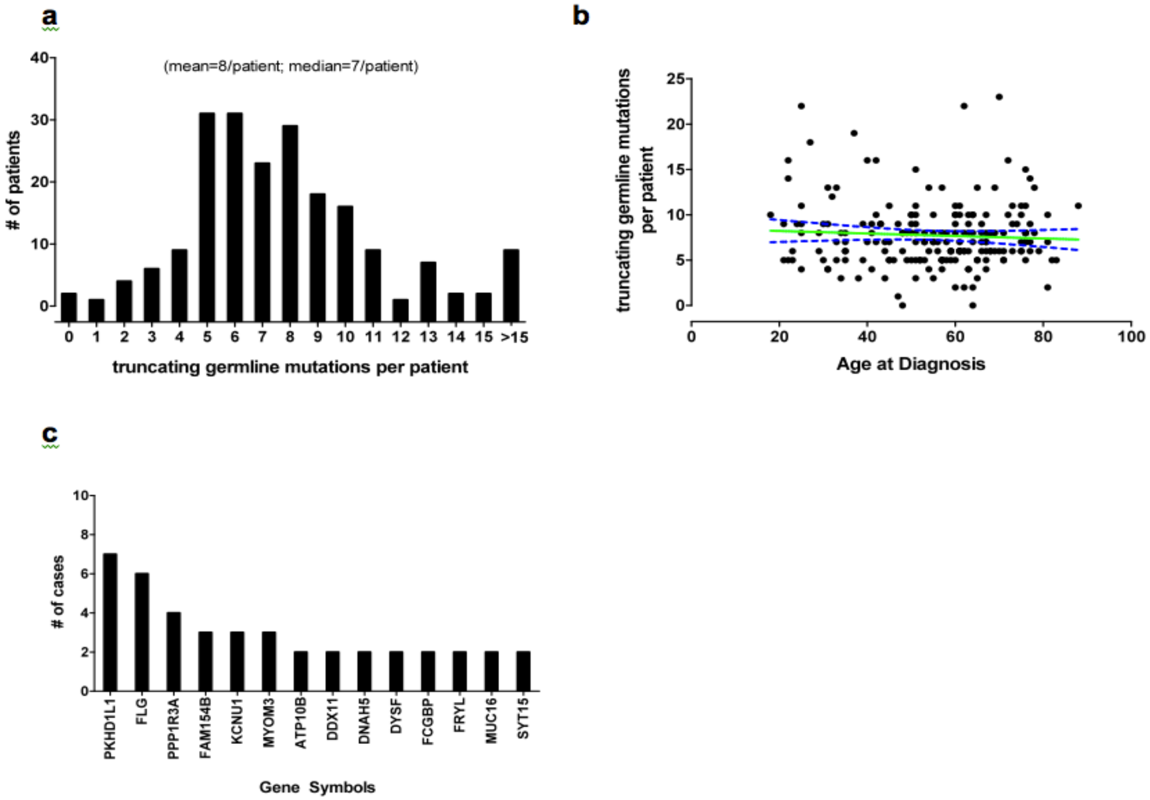


Figure S6. Germline truncating variants in AML patients

a) Distribution of truncating germline variants in all genes. b) No relationship between age at AML diagnosis and burden of germline truncating variants (linear regression in green; 95% confidence intervals in blue). c) Frequency of germline truncating variants in genes that have recurrent somatic mutations in AML

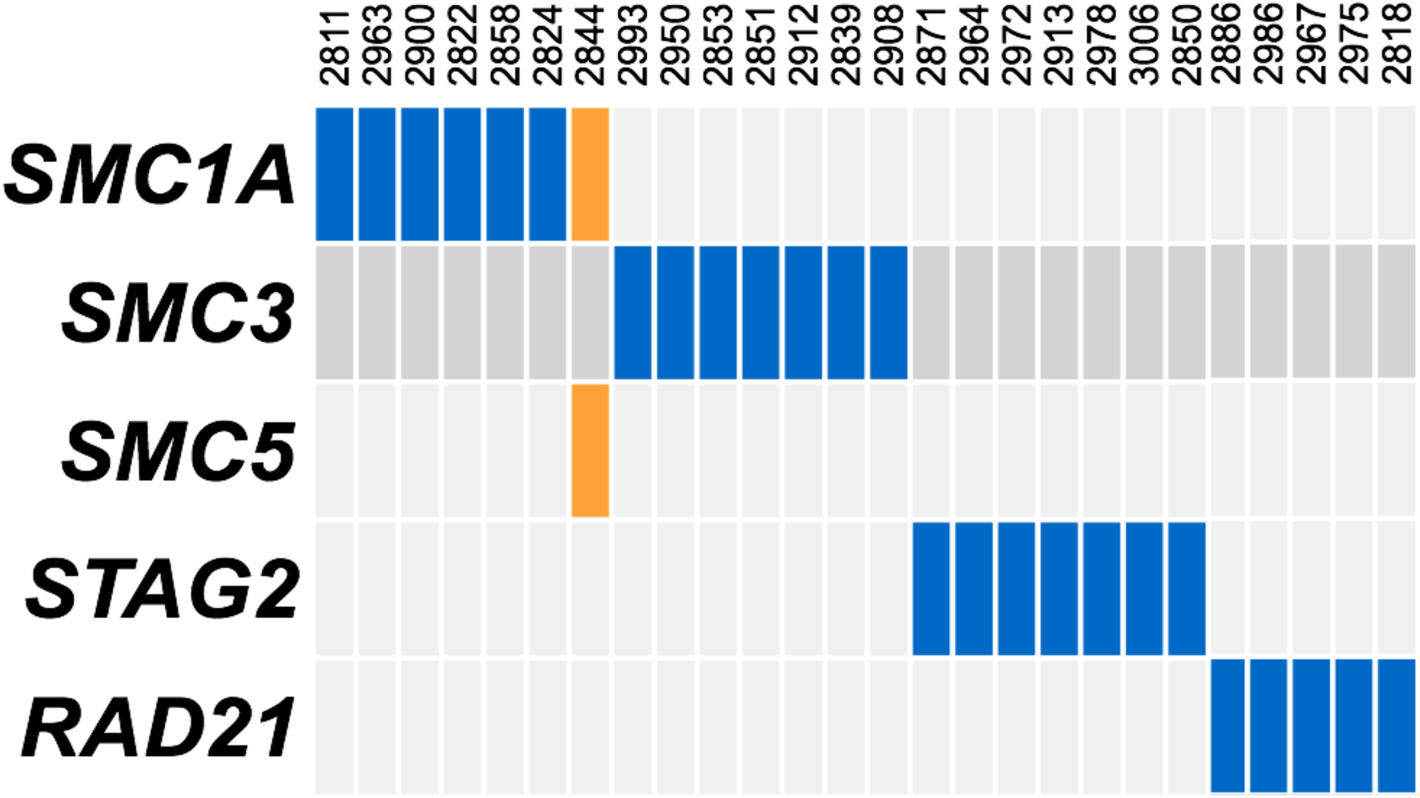


Figure S7. Mutual exclusivity in cohesin complex mutations
A mutation matrix of cohesin complex genes. 26 samples (13%) have at least one mutation in this set of genes. Blue and orange boxes indicate exclusive and co-occurring mutations in a sample, respectively.

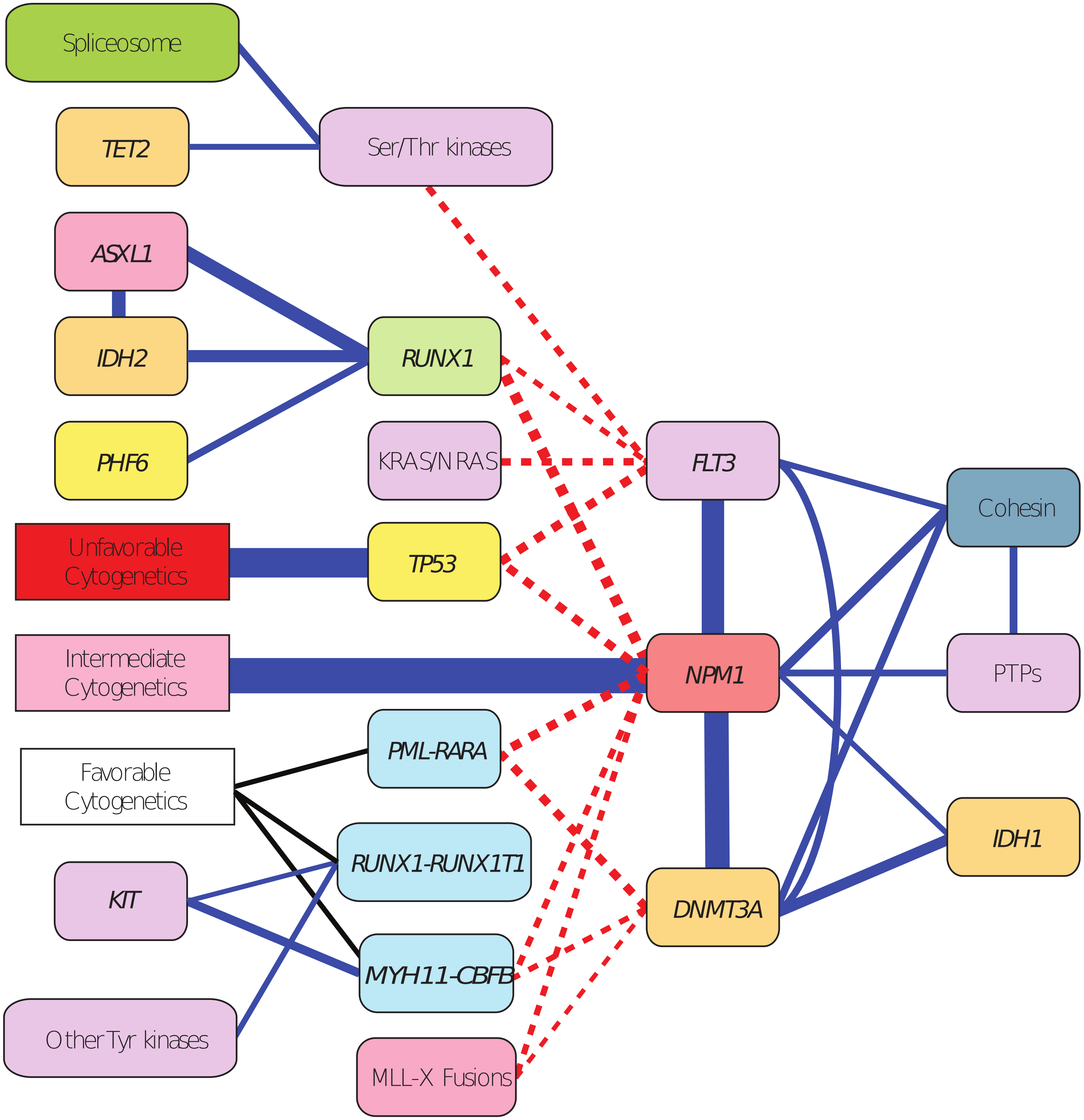


Figure S8. Co-occuring and mutually exclusive mutations in genes/groups
Nodes represent genes, gene groups, or cytogenetic risk groups. Blue edges connect nodes that co-occur in a significant number of samples. Red edges connect nodes that are mutually exclusive. Black edges indicate mutations that define favorable cytogenetics. The thickness of each edge corresponds to the strength of the association. Supplementary Table 19 gives the corresponding p-values (uncorrected) for each association.

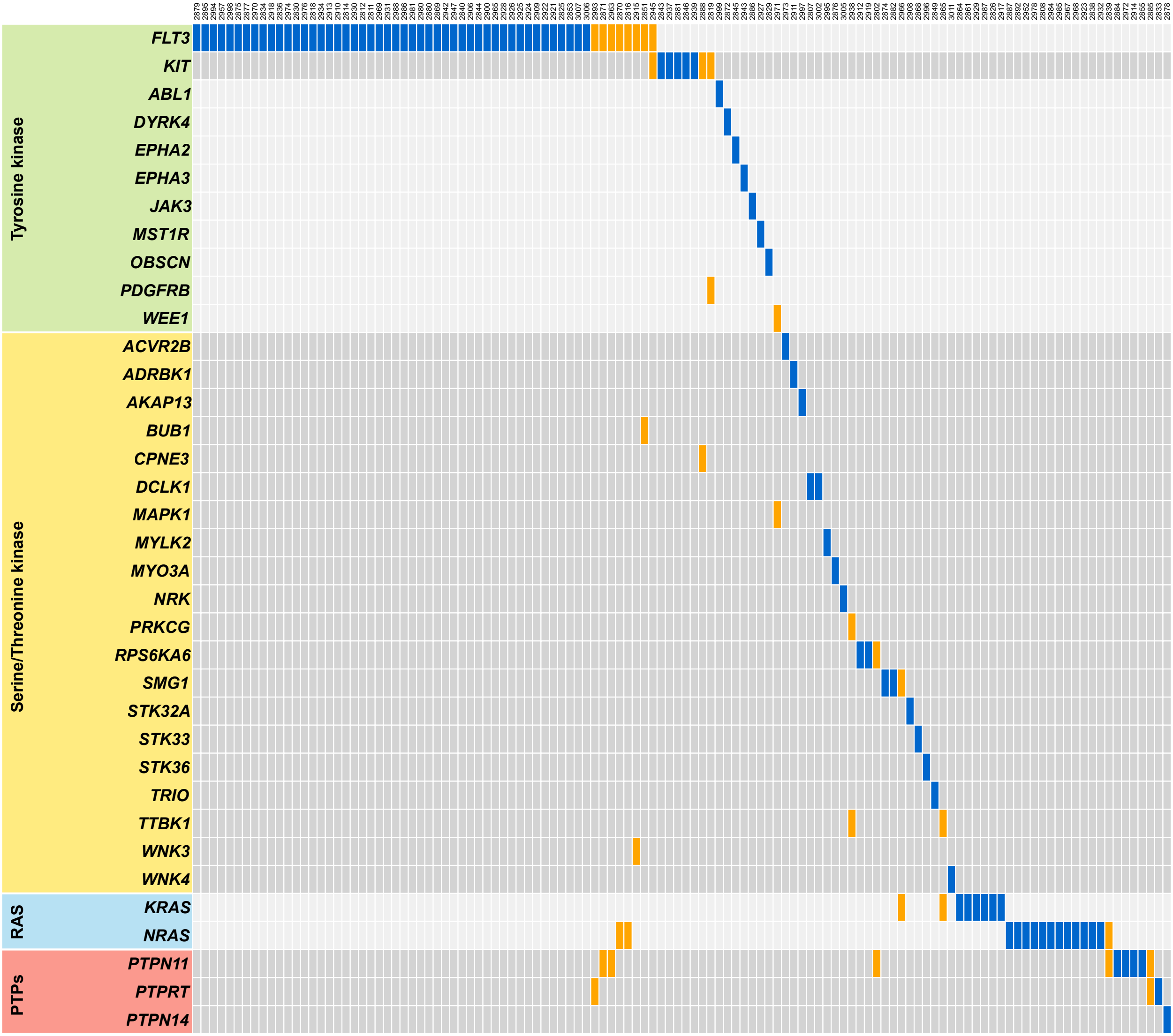


Figure S9. Mutations in kinases, RAS proteins, and phosphatases

A mutation matrix of protein kinases, phosphatase, and RAS proteins. The categories include tyrosine kinases (green), serine/threonine kinases (yellow), Ras proteins (blue), and protein tyrosine phosphatases (red). In total, 118/200 (59%) of samples have a mutation in these genes, with FLT3 mutations accounting for 56 of these samples. Blue boxes indicate mutations appear in only one of the listed samples, while orange boxes indicate mutations that co-occur within one of the listed samples.

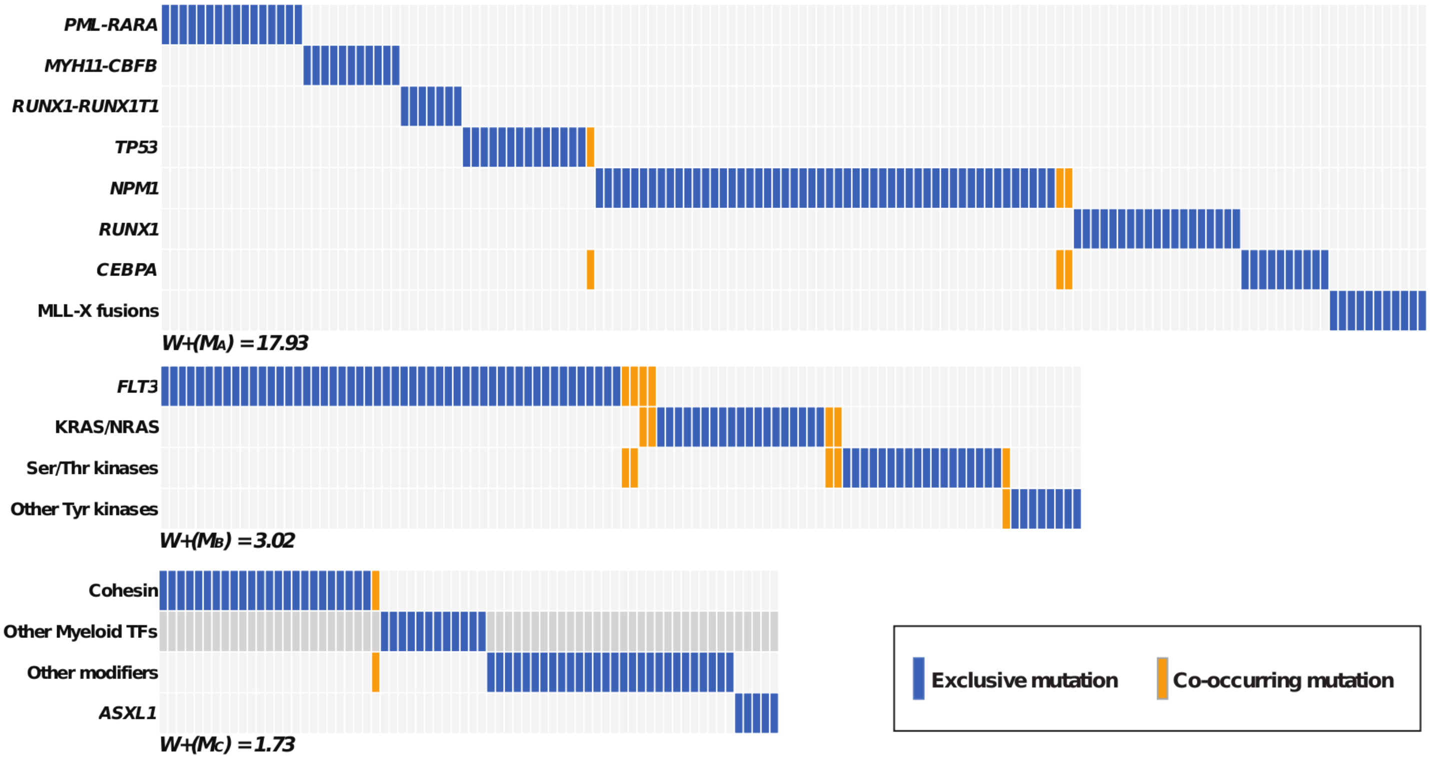


Figure S10. Mutual exclusivity between genes

Computational analysis using the Dendrix++ algorithm identifies the strongest mutually-exclusive sets of genes, with P < 0.001, P < 0.021, and P < 0.103 for groups A, B, and C, respectively. Blue boxes indicate mutations that are exclusive across all listed genes and samples; orange boxes indicate co-occurring mutations.

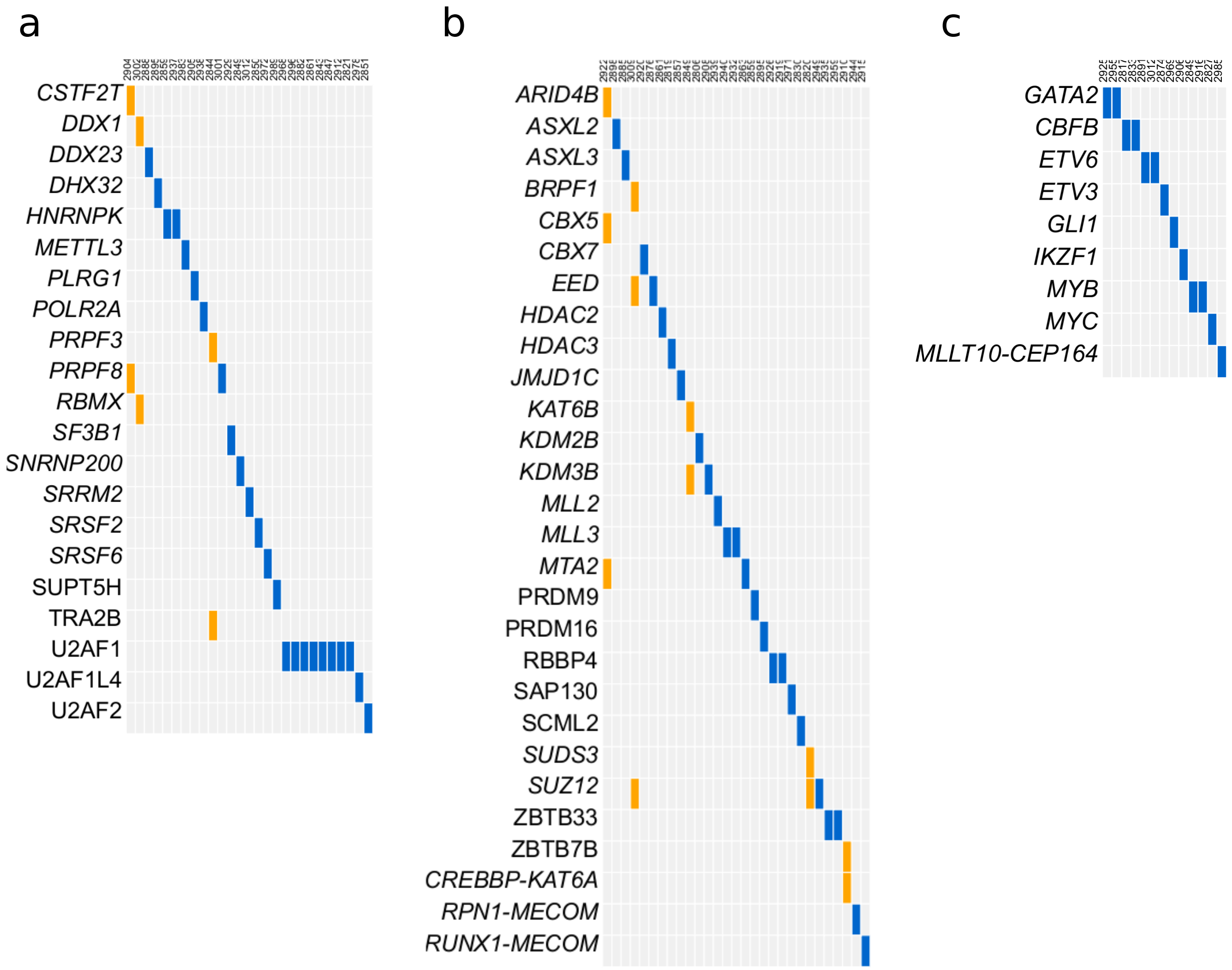


Figure S11. Mutation Matrices of spliceosome and chromatin modifier genes

a) A mutation matrix of spliceosome genes. The coverage of the gene set is 25 samples (12.5%). Blue and orange boxes indicate exclusive and co-occurring mutations in the sample, respectively. b) A mutation matrix of other chromatin modifiers. 29 samples (14.5%) have at least one mutation in this set of genes. c) A mutation matrix for myeloid transcription factors. Blue and orange boxes indicate exclusive and co-occurring mutations in a sample, respectively.

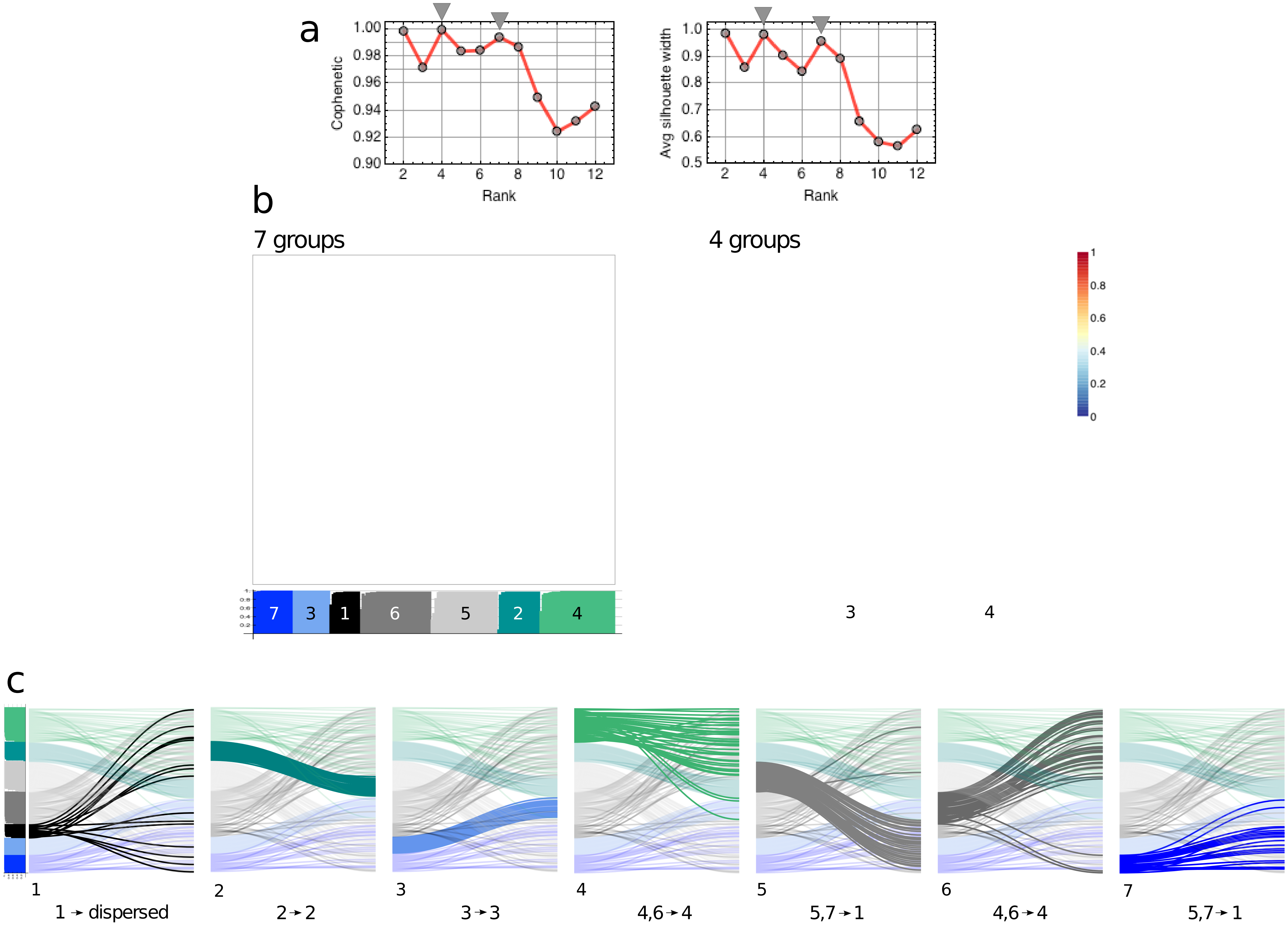
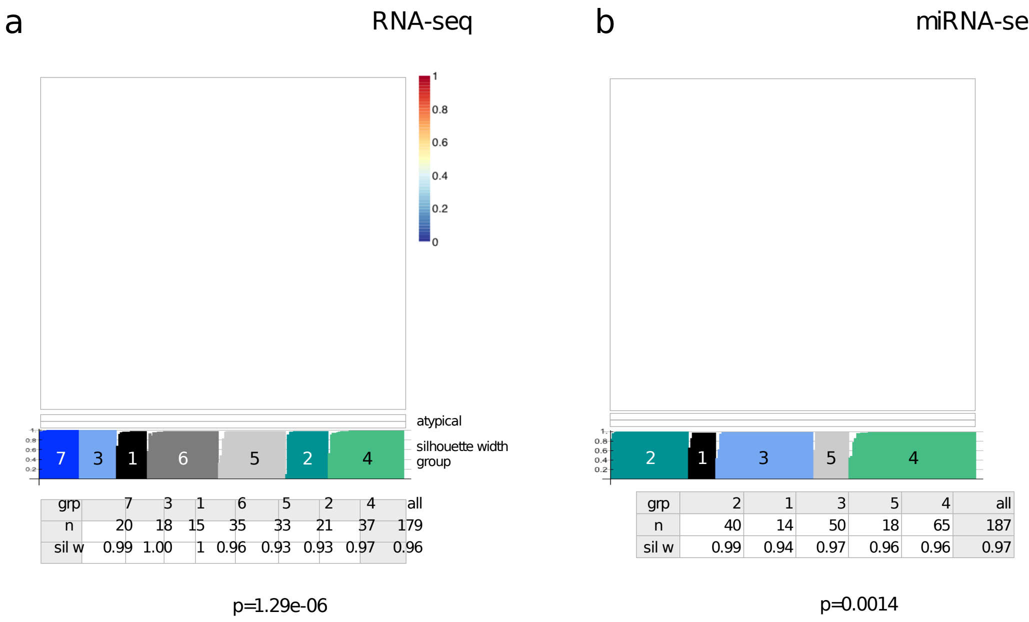


Figure S12. RNA-seq clustering Relationship between four- and seven-group clustering solutions for RNA-seq.
a) From the NMF clustering rank survey, profiles of cophenetic correlation coefficient and of average silhouette width of the consensus memberships matrix. Grey triangles show the four- and seven-group solutions suggested by these two metrics. b) Consensus membership heatmaps with silhouette width profiles. c) Connectivity for individual groups from the seven-group solution (left) into the four-group solution (right). Bezier curves link a sample's positions in the two clustering results, and are drawn using colours for the seven-group solution. Samples with relatively low silhouette widths can be considered as 'atypical' group members. Text under each graphic summarizes the dominant connectivity relationship(s).



grp	7	3	1	6	5	2	4	all
n	20	18	15	35	33	21	37	179
sil w	0.99	1.00	1	0.96	0.93	0.93	0.97	0.96

grp	2	1	3	5	4	all
n	40	14	50	18	65	187
sil w	0.99	0.94	0.97	0.96	0.96	0.97

Figure S13. Univariate Kaplan Meier results for overall survival for unsupervised groups. a) mRNA-seq, b) miRNA-seq. For each data type, top to bottom: 1) consensus membership heatmaps, with a scale bar showing consensus membership values; 2) atypical members of each group are those with relatively low silhouette widths (black, for width thresholds of f=0.95 and 0.90); 3) a silhouette width profile calculated from the consensus membership matrix; 4) summary tables showing the number of samples and the average silhouette width for each group, and for all samples; 5) univariate Kaplan-Meier plots with log-rank p-values.

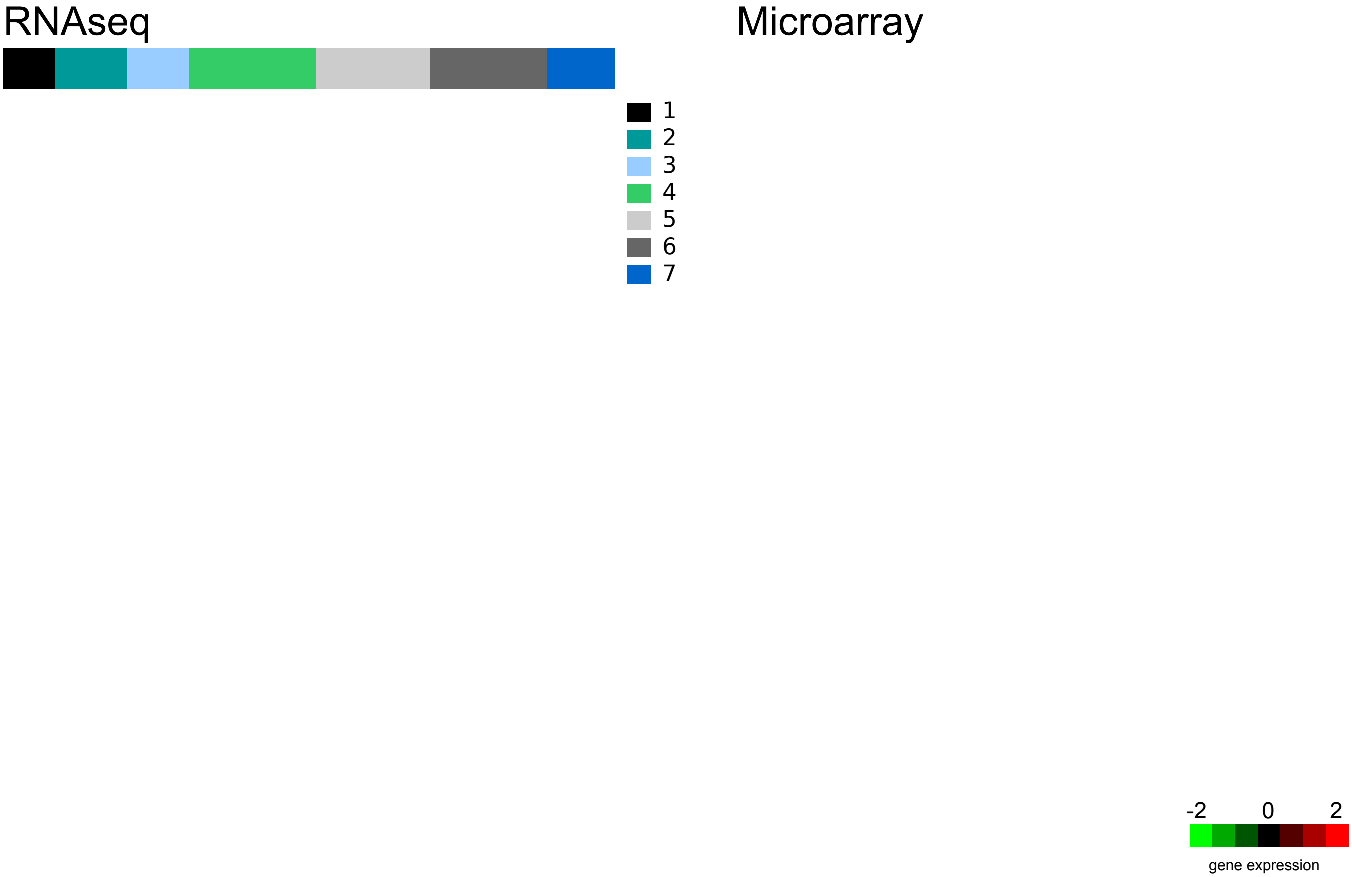
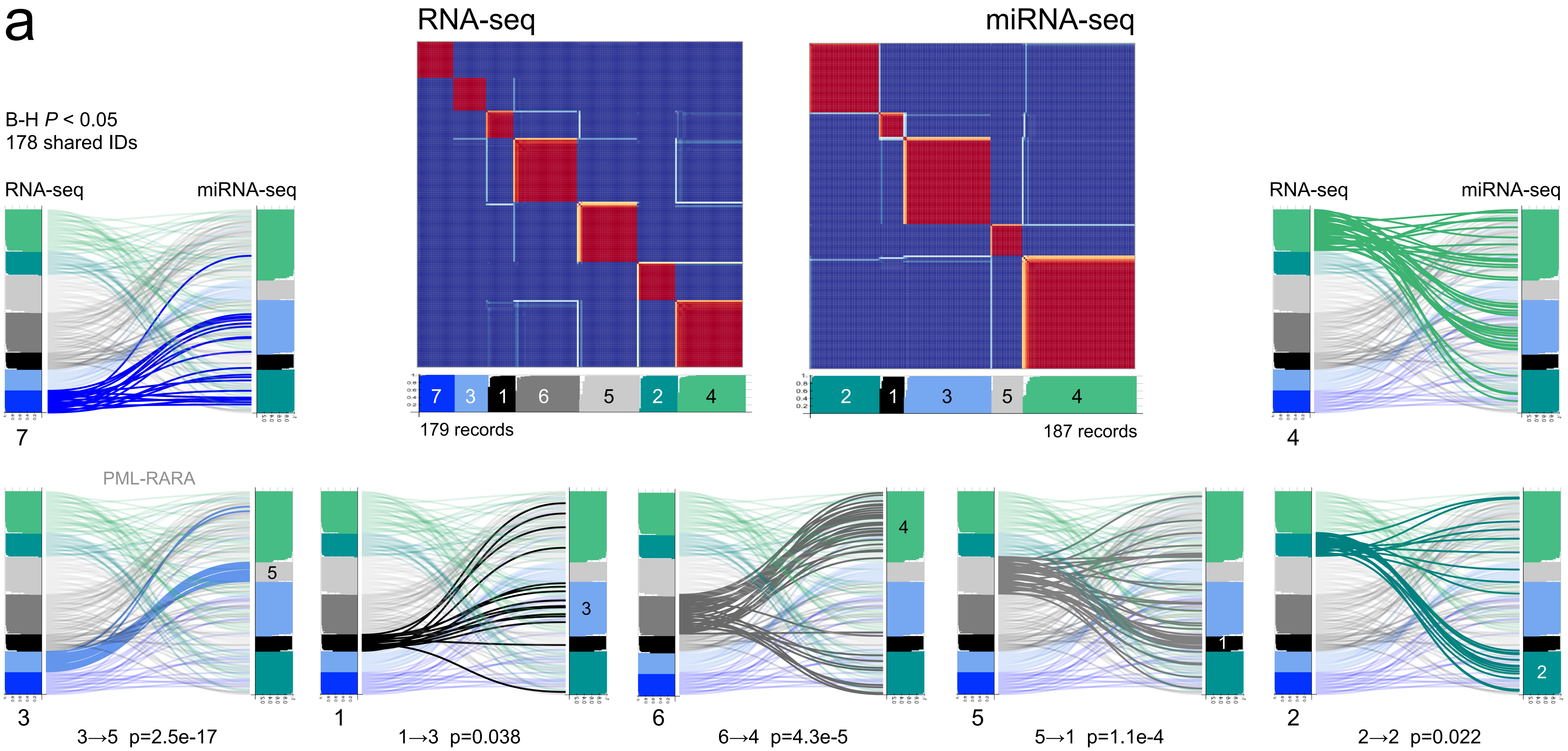
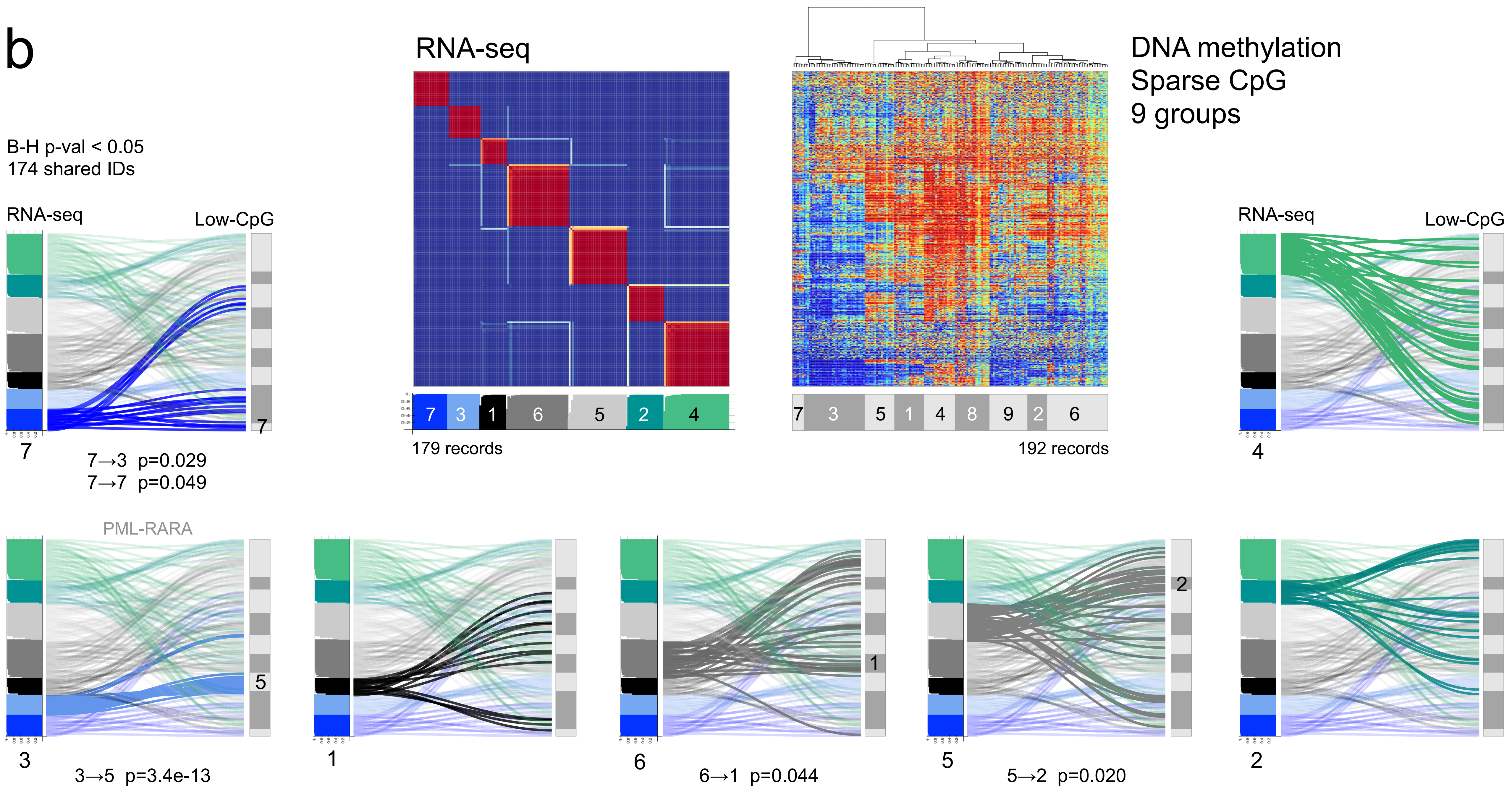


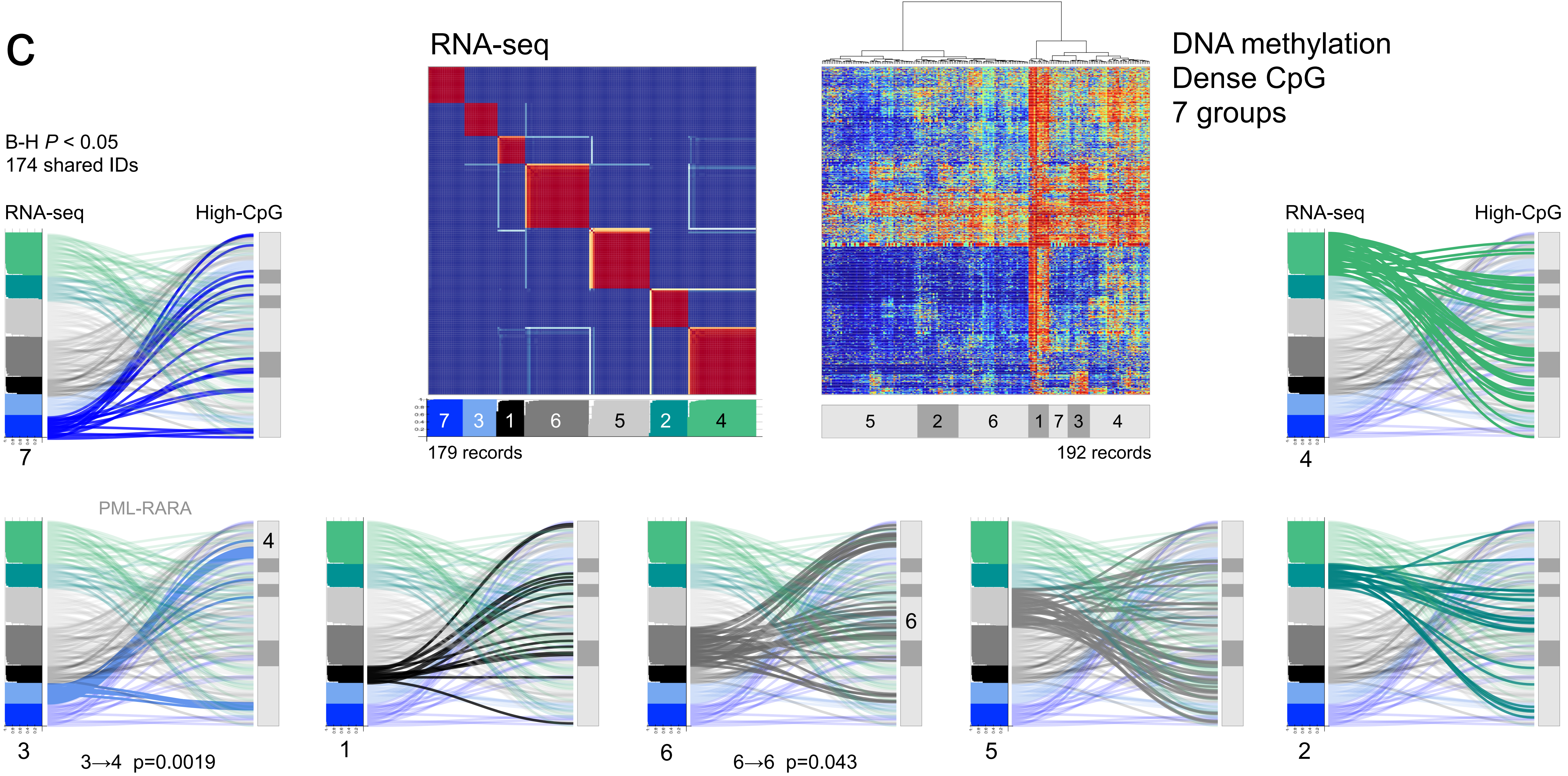
FIgure S14. Comparison between RNA-seq and microarray clustering
Expression-based clusters identified in RNAseq were used to compare expression in matched microarray data. Genes predictive of the seven RNAseq clusters were used to cluster the microarray data, identifying similar patterns in the microarray data. In the RNAseq data, genes were hierarchically clustered, while maintaining the order of the subtype classes for samples from NMF clustering. Microarray data were visualized by maintaining the same gene and sample order as in the RNAseq data

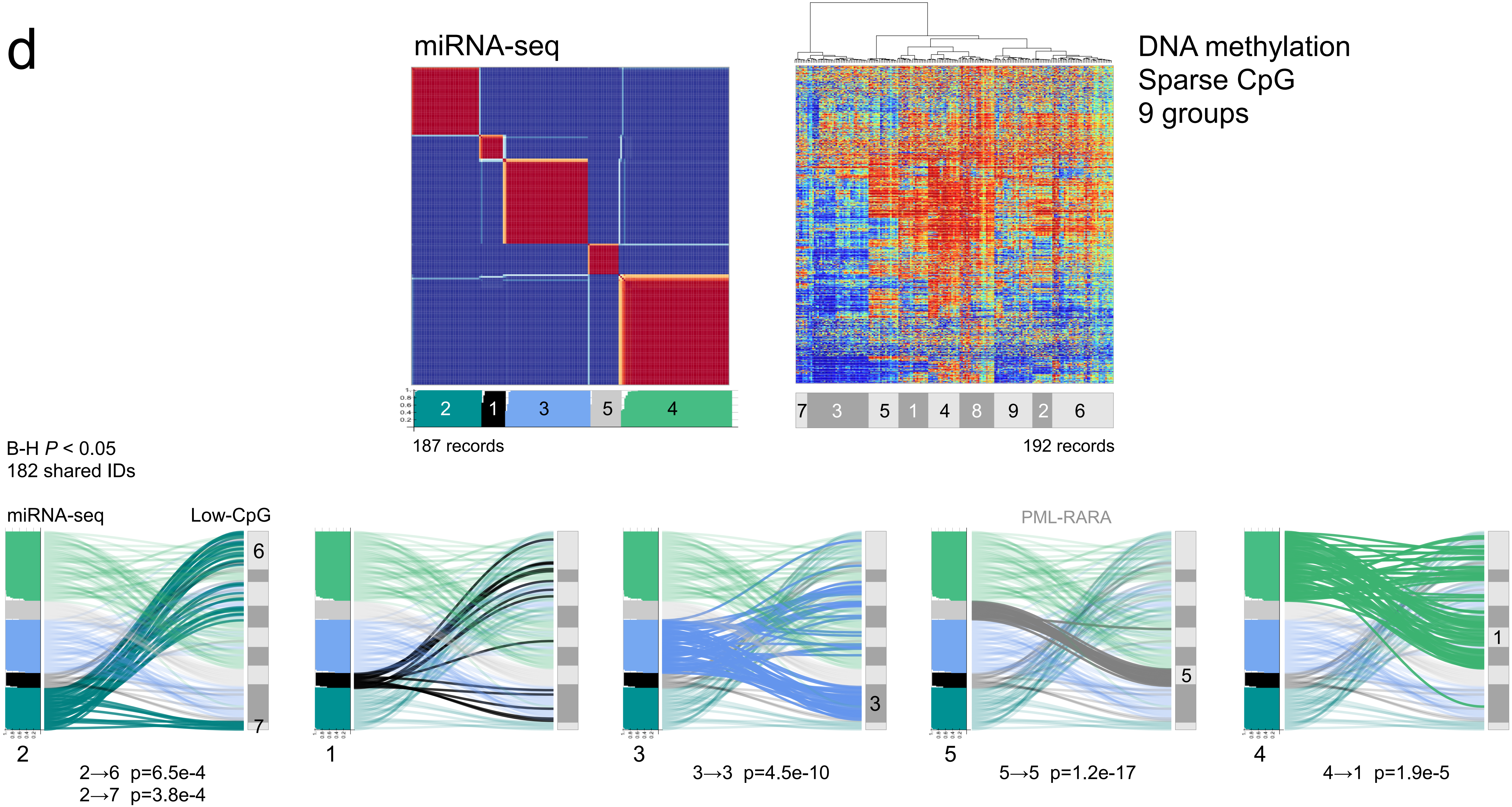
Figure S15. Relationships between unsupervised groups for RNA-seq, miRNA-seq and DNA methylation data. (next 5 pages)

Curved lines map the sample order from the left clustering result into the sample order in the right clustering result. Silhouette width profiles are as in Fig. 4. a) Seven RNA-seq groups vs. five miRNA-seq groups. b,c) Seven RNA-seq groups vs. b) nine DNA methylation groups for sparse-CpG regions, and vs. c) seven DNA methylation groups for dense- CpG regions. d,e) As b,c), but for the five miRNA-seq groups in a). Text under each graphic summarizes the statistically significant connectivity relationship(s).


a
RNA-seq
miRNA-seq
B-H P < 0.05
178 shared IDs
RNA-seq
miRNA-seq
7
7
3
1
6
5
2
4
179 records
2
1
3
5
4
187 records
RNA-seq
miRNA-seq
4
PML-RARA
5
3
3→5 p=2.5e-17
3
1
1→3 p=0.038
4
6
6→4 p=4.3e-5
1
5
5→1 p=1.1e-4
2
2
2→2 p=0.022


b
B-H p-val < 0.05
174 shared IDs
RNA-seq
Low-CpG
7
7
7→3 p=0.029
7→7 p=0.049
RNA-seq
7
3
1
6
5
2
4
179 records
DNA methylation
Sparse CpG
9 groups
7
3
5
1
4
8
9
2
6
192 records
RNA-seq
Low-CpG
4
PML-RARA
5
3
3→5 p=3.4e-13
1
1
6
6→1 p=0.044
2
5
5→2 p=0.020
2


C
RNA-seq
DNA methylation
Dense CpG
7 groups
B-H P < 0.05
174 shared IDs
RNA-seq
High-CpG
7
7
3
1
6
5
2
4
179 records
5
2
6
1
7
3
4
192 records
RNA-seq
High-CpG
4
PML-RARA
4
3
3→4 p=0.0019
1
6
6
6→6 p=0.043
5
2


d
miRNA-seq
2
1
3
5
4
187 records
DNA methylation
Sparse CpG
9 groups
7
3
5
1
4
8
9
2
6
192 records
B-H *P* < 0.05
182 shared IDs
miRNA-seq
Low-CpG
6
7
2
2→6 p=6.5e-4
2→7 p=3.8e-4
1
3
3
3→3 p=4.5e-10
PML-RARA
5
5
5→5 p=1.2e-17
1
4
4→1 p=1.9e-5

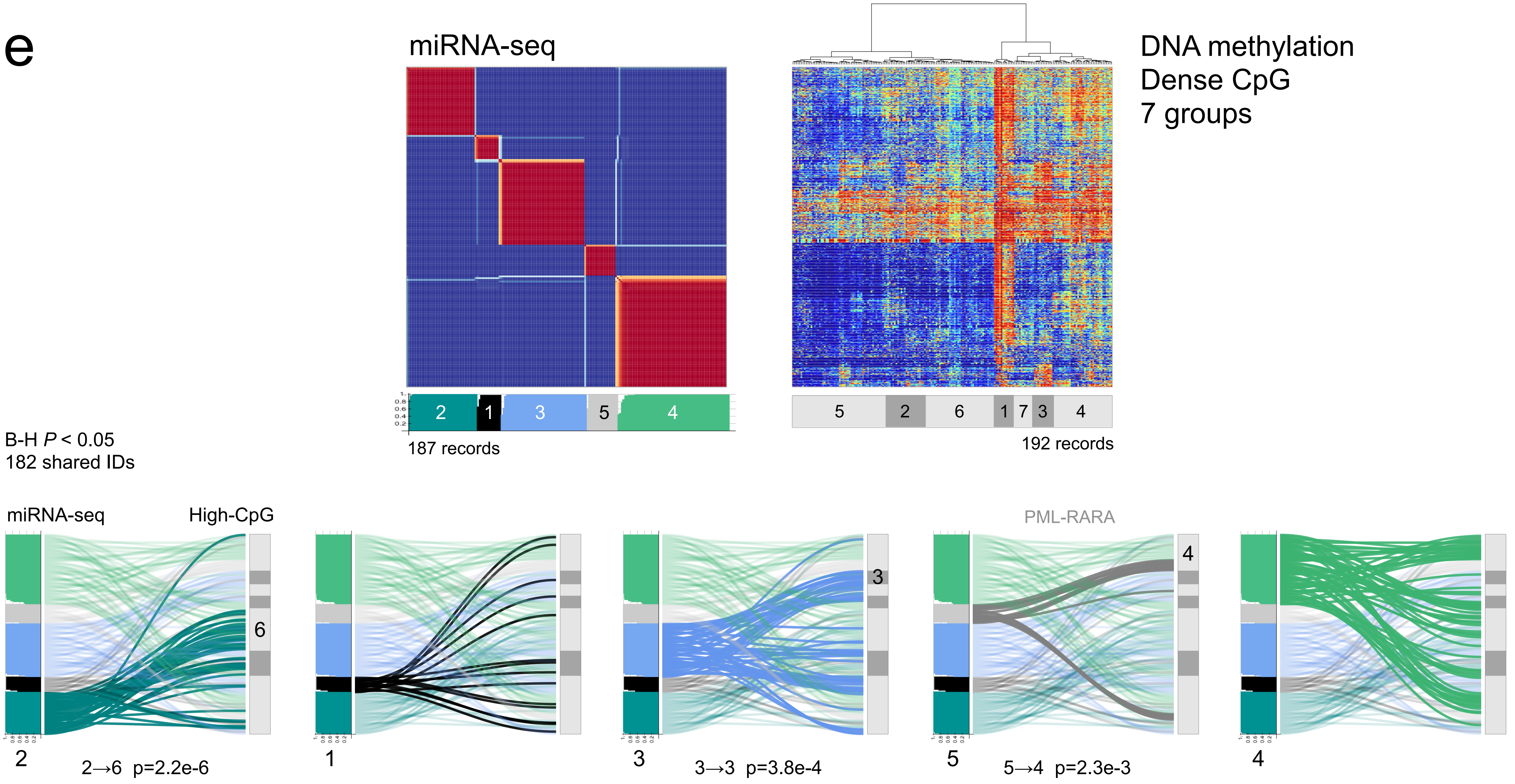
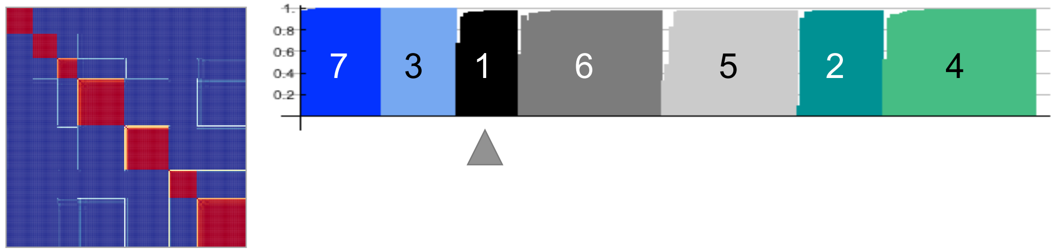

e
miRNA-seq
2
1
3
5
4
187 records
DNA methylation
Dense CpG
7 groups
5
2
6
1
7
3
4
192 records
B-H P < 0.05
182 shared IDs
miRNA-seq
High-CpG
6
2
2→6 p=2.2e-6
1
3
3
3→3 p=3.8e-4
PML-RARA
4
5
5→4 p=2.3e-3
4

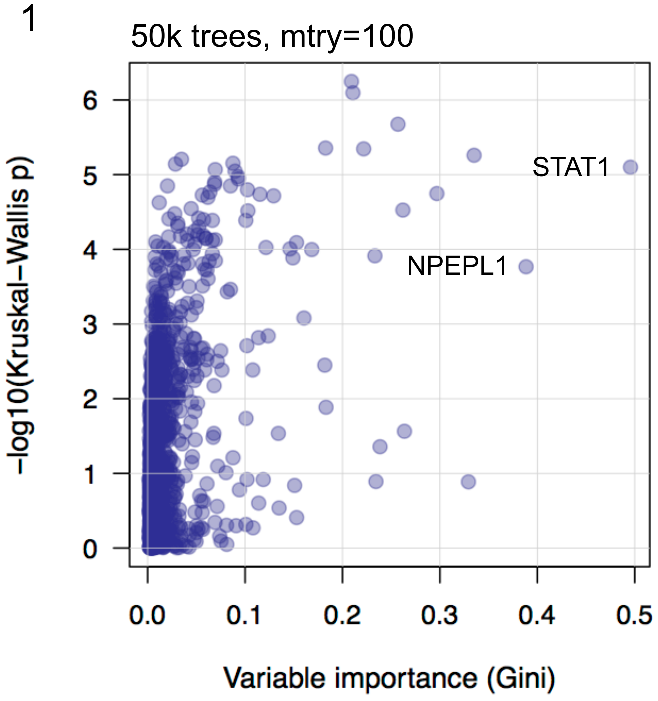
Figure S16. Discriminatory genes and miRNAs for unsupervised groups. (next 12 pages)

a-g) Seven RNA-seq groups. h-l) Five miRNA-seq groups. Discriminatory genes and miRNAs were identified by random forest classifiers, using RPKM for RNA-seq data and RPM for miRNA-seq mature/star strand data (see Supplementary Methods). A gene or miR that helps a classifier separate or discriminate a group of samples from all other samples tends to have a higher or lower abundance in samples in that group. See, for example, the box- whisker plots for MPO and CALR for RNA-seq group 3, in which each gene's RPKM distribution in group 3 is highlighted by a red rectangle. For each sample group, panels show (left to right, top to bottom): 1) The importances of genes or miRs in a classifier are correlated to Kruskal-Wallis P-values for genes or miRs being differentially abundant; 2) Profile of the estimated classifier error rate as a function of the number of most important 20 genes or 40 miRs; 3) Table of discriminatory genes or miRs, ranked by importance (Gini), with variables to the right of the minimum of the minimum error rate in the profile in (2) shown in gray; 4) Box-whisker plots of abundance across the groups for a subset of highly-ranked genes or miRs; 5) Estimated overall error rate, and table of the number of correctly and incorrectly classified samples (i.e. confusion matrix) for subset of genes or miRs to the left of the minimum error in (2).

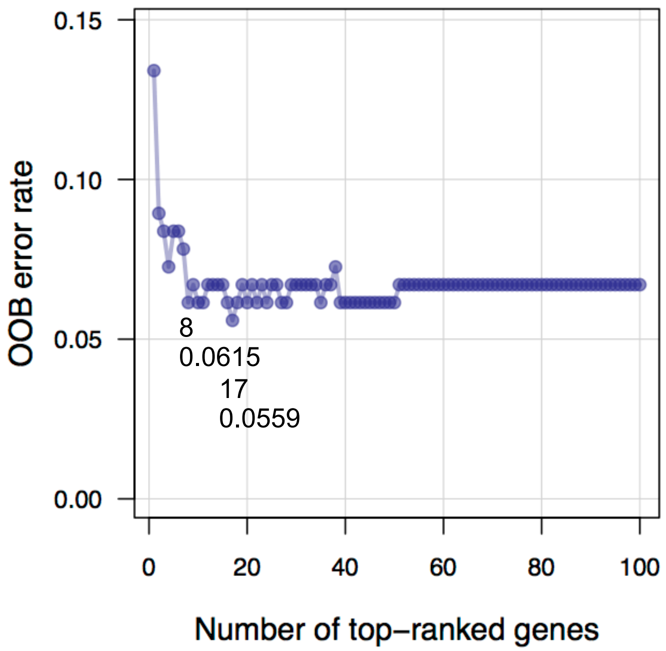
a) RNA-seq: group 1 of 7



1



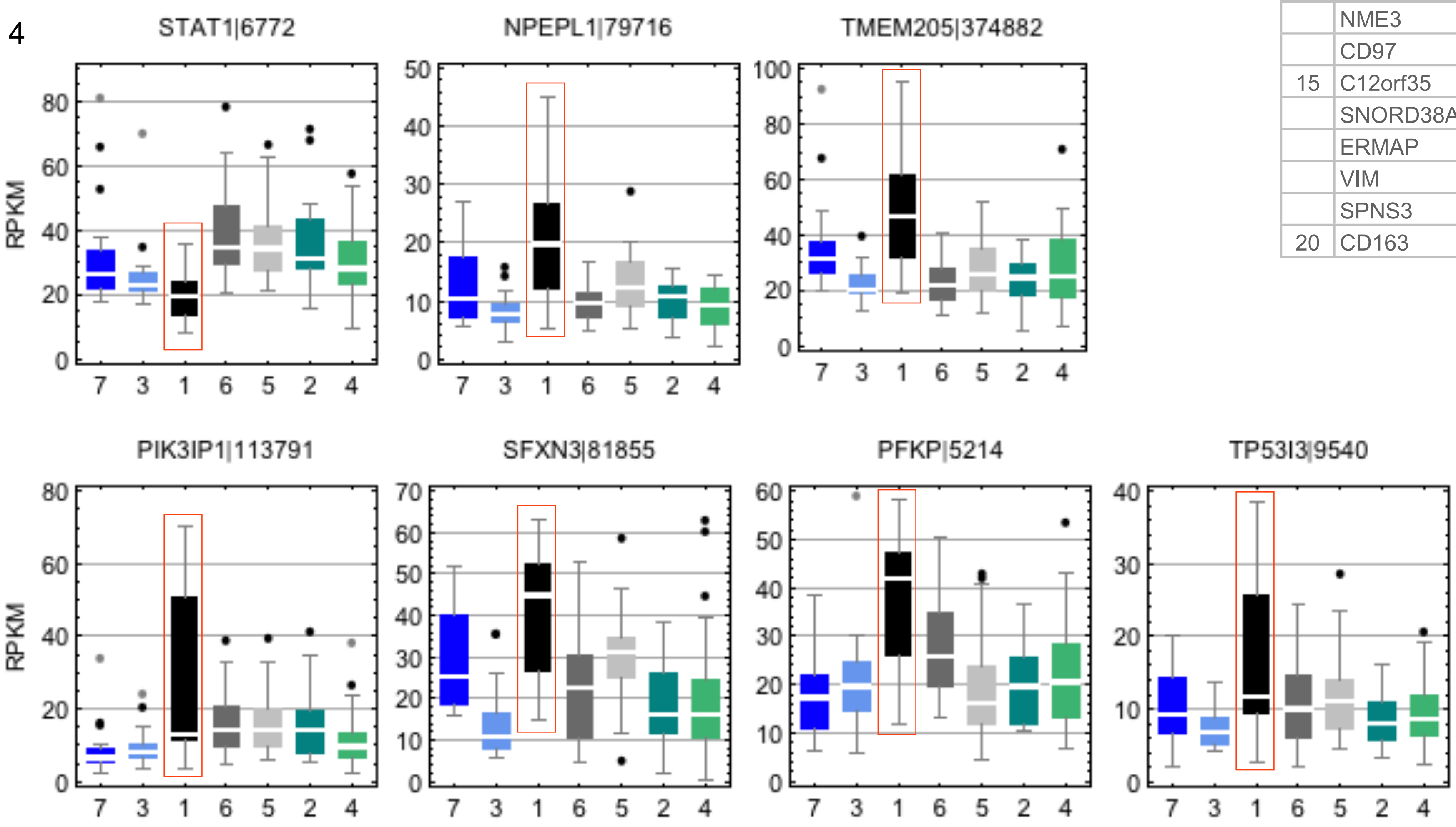
2



3

n	Symbol	Entrez	MD Acc	MD Gini
	STAT1	6772	4.58	0.50
	NPEPL1	79716	4.54	0.39
	TMEM205	374882	3.89	0.34
	PIK3IP1	113791	4.84	0.33
5	SFXN3	81855	2.33	0.30
	TP53I3	9540	3.40	0.26
	PFKP	5214	2.16	0.26
	TTC38	55020	1.93	0.26
	CCS	9973	2.22	0.24
10	SMAP2	64744	3.36	0.23
	FAM116B	414918	2.72	0.23
	ARHGDIA	396	2.82	0.22
	NME3	4832	2.41	0.21
	CD97	976	2.01	0.21
15	C12orf35	55196	3.02	0.18
	SNORD38A	94162	2.13	0.18
	ERMAP	114625	2.04	0.18
	VIM	7431	1.91	0.17
	SPNS3	201305	1.89	0.16
20	CD163	9332	2.46	0.15

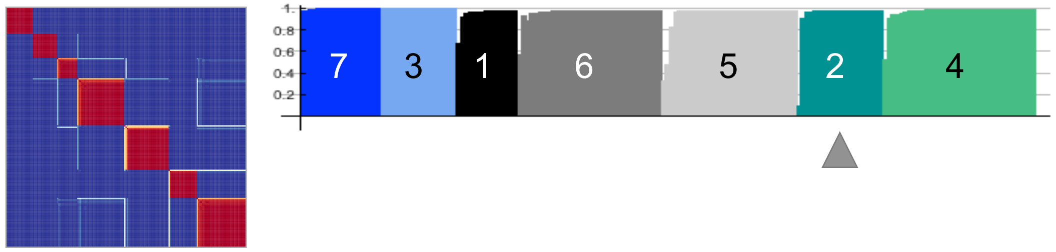
4



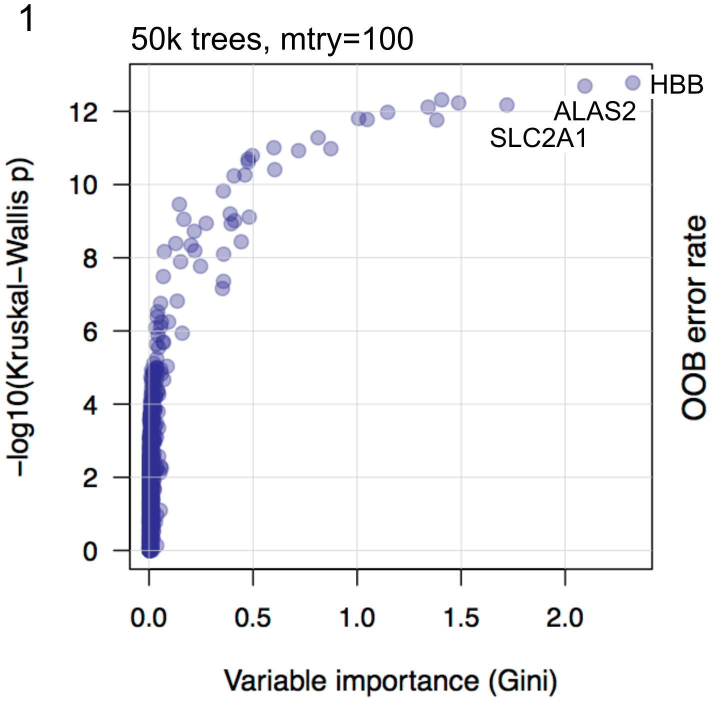
5

```
Random forest: classification
Number of trees: 50000
Vars tried at each split: 1
Top 17 genes
OOB est of  error rate: 5.59%
Confusion matrix:
    0 1 class.error
0 164 0   0.00
1  10 5   0.67

Top 8 genes
OOB est of  error rate: 6.15%
Confusion matrix:
    0 1 class.error
0 163 1 0.0061
1  10 5 0.67
```

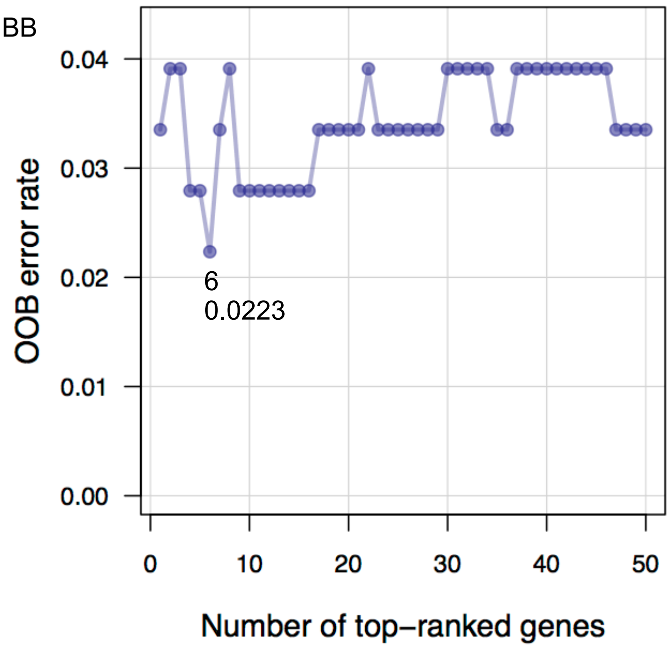
b) RNA-seq: group 2 of 7



1



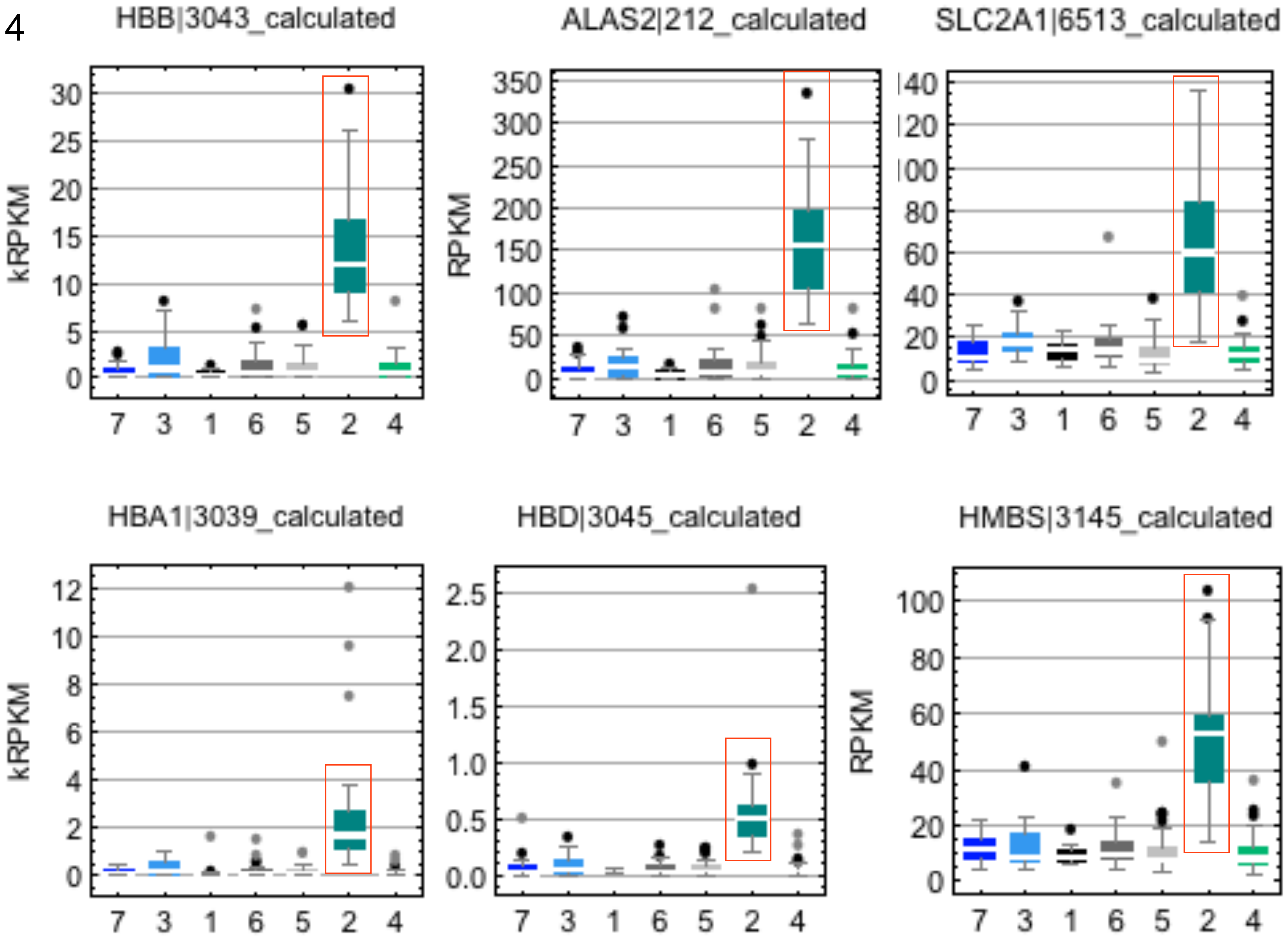
2



3

n	Symbol	Entrez	MD Acc	MD Gini
	HBB	3043	8.73	2.32
	ALAS2	212	8.21	2.10
	SLC2A1	6513	7.27	1.72
	HBA1	3039	7.08	1.49
	HBD	3045	6.68	1.41
6	HMBS	3145	6.61	1.38
	HBA2	3040	6.84	1.34
	SLC25A39	51629	6.24	1.15
	BCL2L1	598	5.74	1.05
10	SLC4A1	6521	5.48	1.01
	ERMAP	114625	5.08	0.87
	AHSP	51327	4.87	0.81
	RBM38	55544	4.58	0.72
	MARCH8	220972	4.25	0.60
15	BPGM	669	4.66	0.60
	KLF1	10661	4.09	0.50
	BLVRB	645	3.84	0.48
	HBM	3042	4.33	0.48
	SLC25A37	51312	3.81	0.48
20	RHAG	6005	3.28	0.46

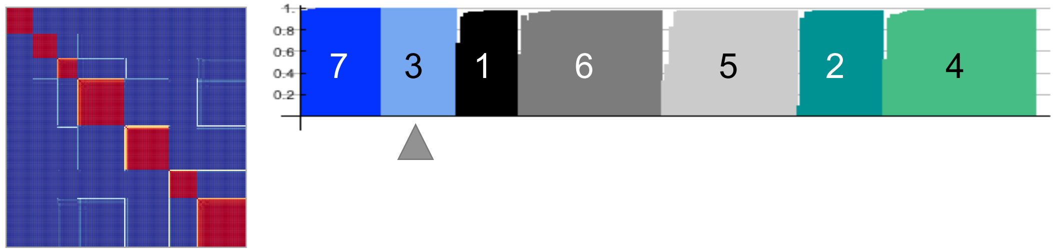
4



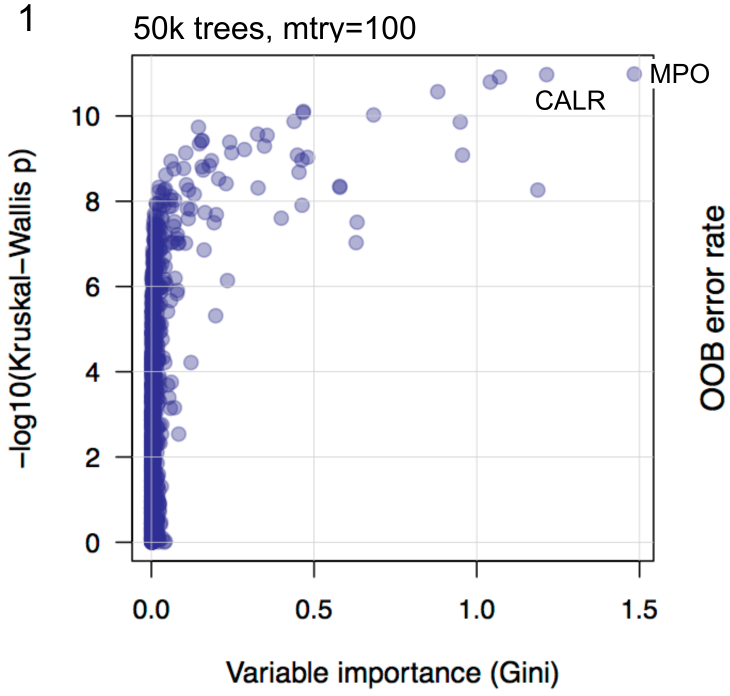
5

```
Random forest: classification
Number of trees: 50000
Vars tried at each split: 1
Top 6 genes
OOB est of  error rate: 2.23%
Confusion matrix:
    0   1 class.error
0 156   2   0.013
1   2 19   0.095
```

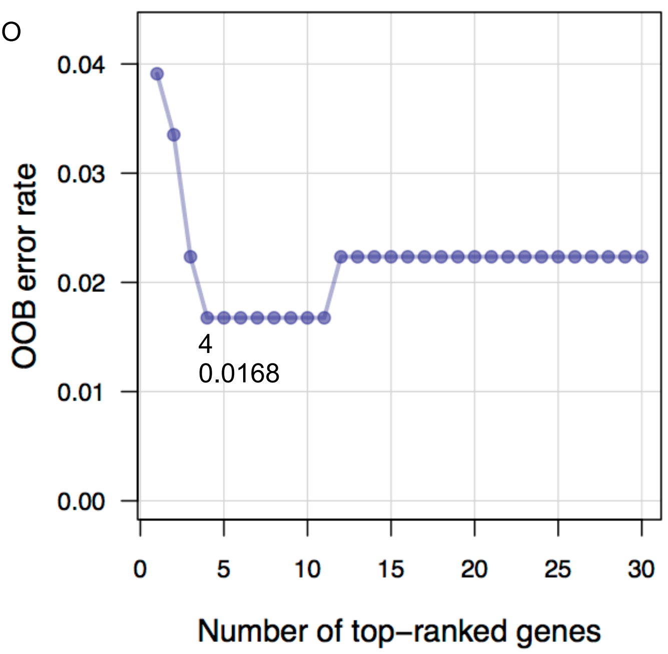
c) RNA-seq: group 3 of 7



1



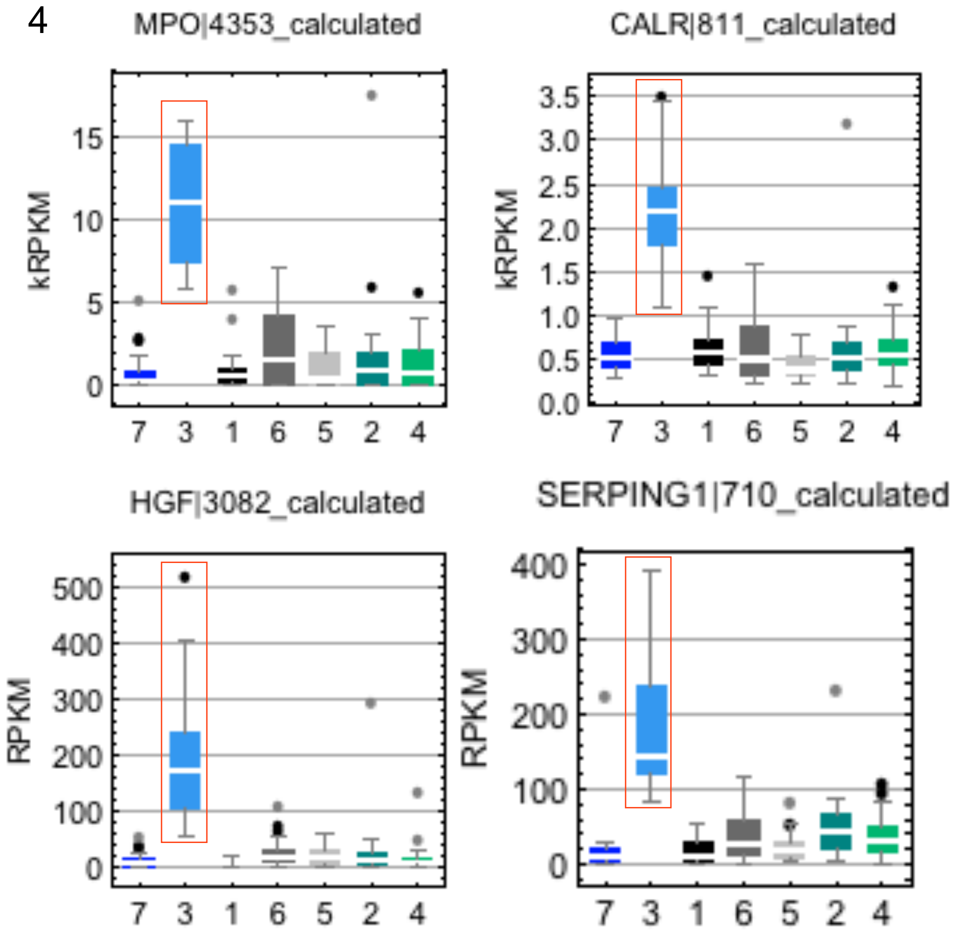
2



3

n	Symbol	Entrez	MD Acc	MD Gini
	MPO	4353	7.26	1.48
	CALR	811	6.24	1.21
	ARHGAP4	393	5.92	1.19
4	HGF	3082	5.60	1.07
	SERPING1	710	6.05	1.04
	CST7	8530	5.42	0.96
	SLC39A11	201266	5.29	0.95
	CLEC2B	9976	5.03	0.88
	FNDC3B	64778	4.64	0.68
10	PRDX4	10549	4.45	0.63
	NT5DC2	64943	4.12	0.63
	LGALS9	3965	4.07	0.58
	VNN1	8876	4.26	0.58
	NTNG2	84628	3.70	0.48
15	SOD2	6648	4.12	0.47
	LPO	4025	3.75	0.47
	TTYH3	80727	3.80	0.46
	PPIB	5479	3.71	0.46
	STAB1	23166	3.66	0.45
20	SKAP2	8935	3.88	0.45

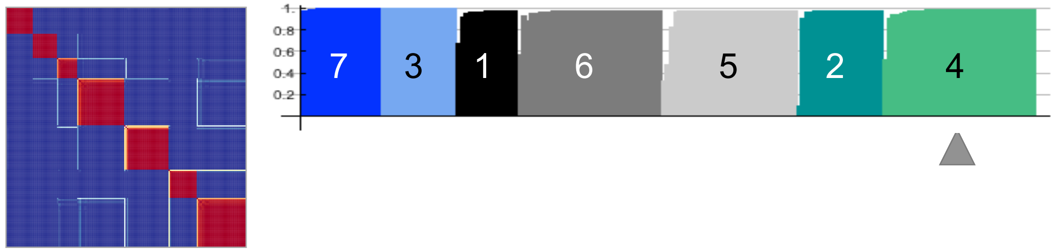
4



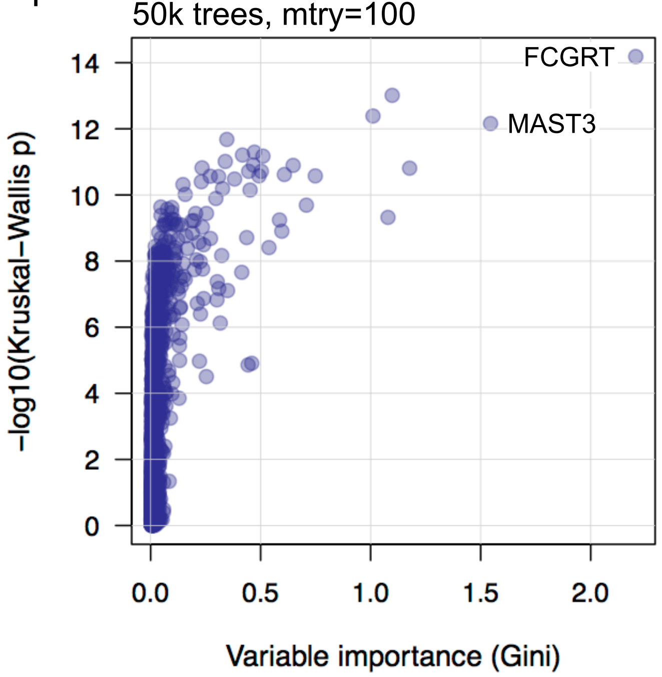
5

```
RF: classification
Number of trees: 50000
Vars tried each split: 1
Top 4 genes
OOB est error rate: 1.68%
Confusion matrix:
    0  1 class.error
0 160  1  0.0062
1   2 16  0.11
```

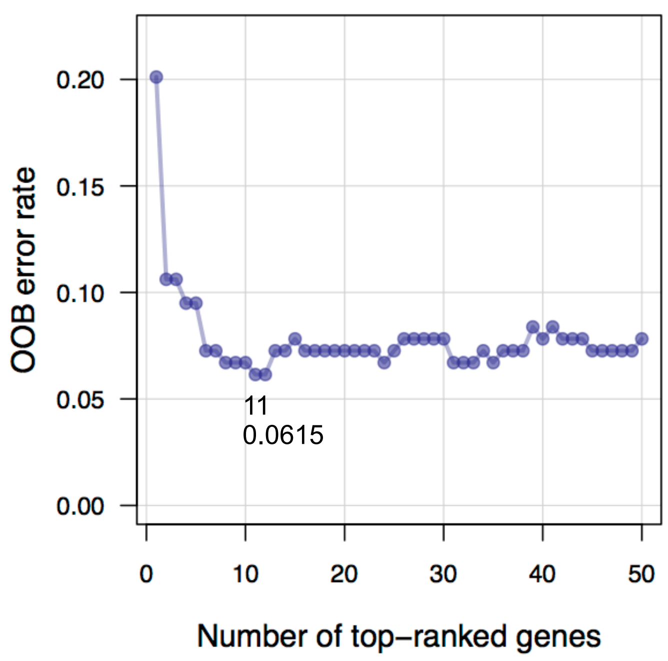
d) RNA-seq: group 4 of 7



1



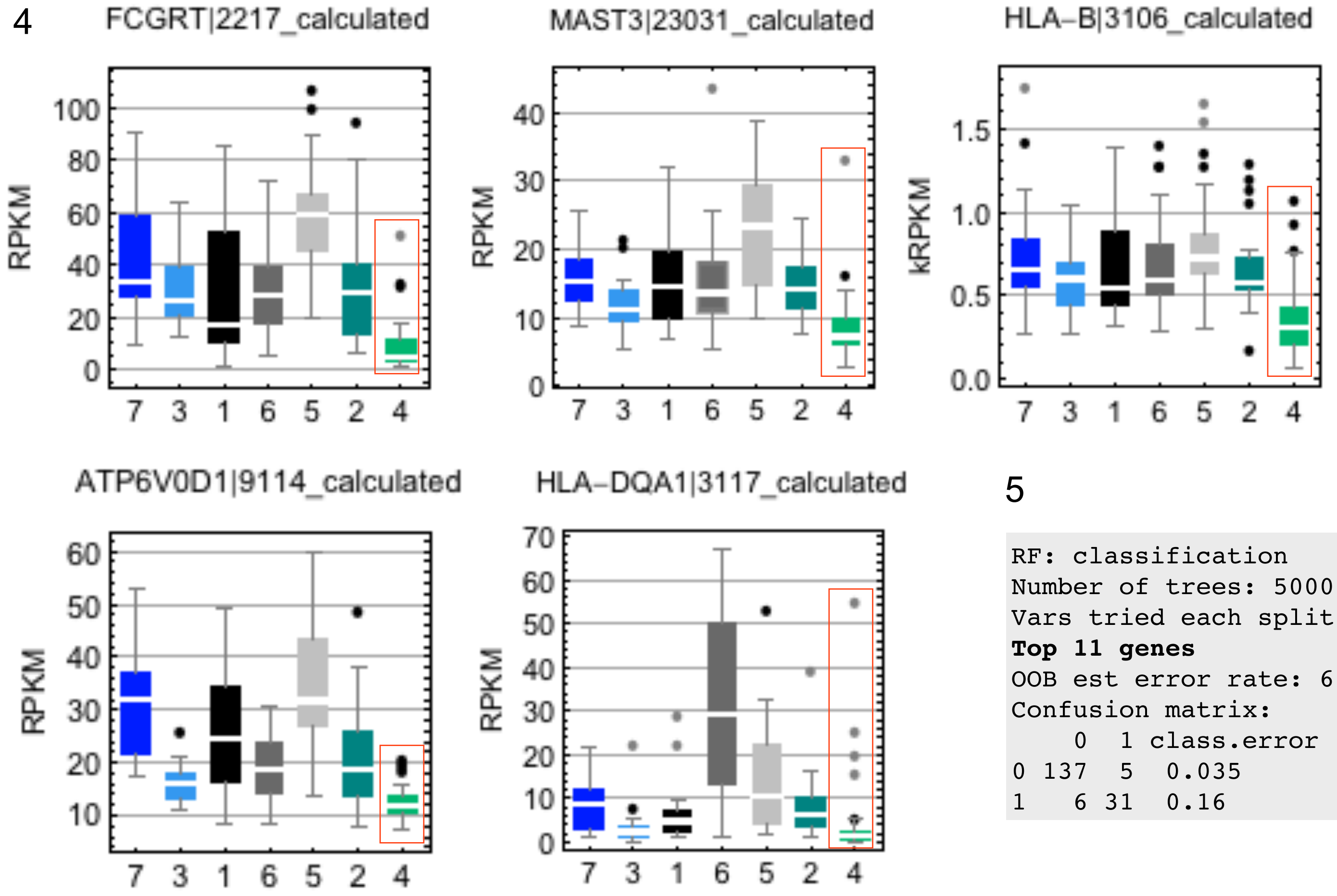
2



3

n	Gene symbol	Entrez	MD Acc	MD Gini
1	FCGRT	2217	7.77	2.20
	MAST3	23031	6.72	1.55
	HLA-B	3106	5.07	1.18
	CECR1	51816	5.08	1.10
5	HLA-DQA1	3117	6.77	1.08
	ATP6V0D1	9114	5.07	1.01
	HLA-DQB1	3119	5.38	0.75
	TSPAN14	81619	4.13	0.71
	GRN	2896	4.04	0.65
10	SNORA32	692063	4.24	0.61
11	SNORD5	692072	4.96	0.60
	NID1	4811	3.66	0.59
	CDKN2D	1032	3.79	0.54
	LGALS3	3958	3.50	0.51
15	CCDC104	112942	3.52	0.50
	RASSF4	83937	3.54	0.49
	RNH1	6050	3.44	0.47
	HVCN1	84329	3.50	0.47
	SEPP1	6414	3.67	0.46
20	IL10RA	3587	3.50	0.45

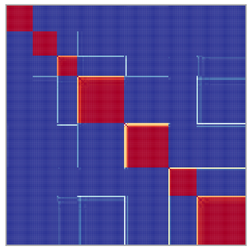
4

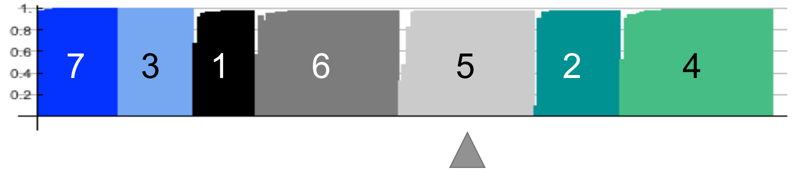


5

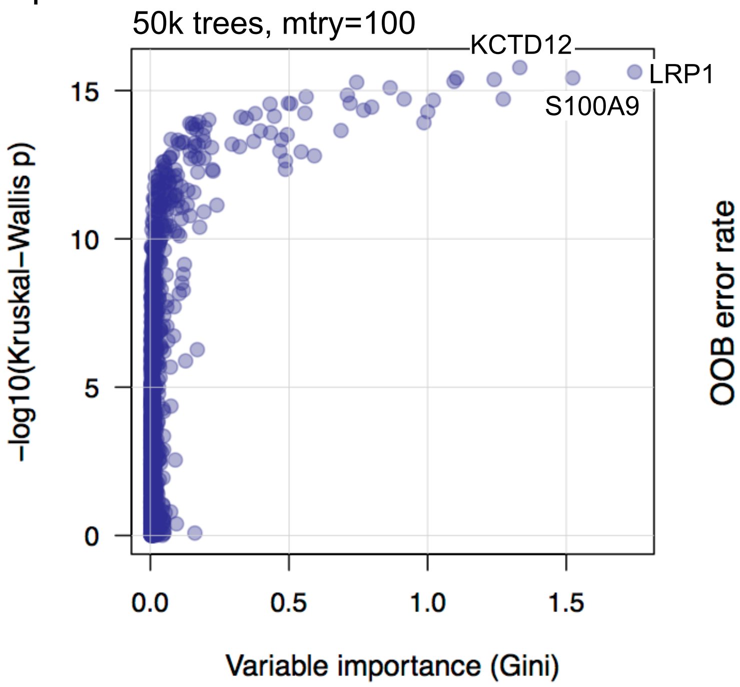
```
RF: classification
Number of trees: 50000
Vars tried each split: 1
Top 11 genes
OOB est error rate: 6.15%
Confusion matrix:
     0   1 class.error
0 137   5   0.035
1   6  31   0.16
```

e) RNA-seq: group 5 of 7

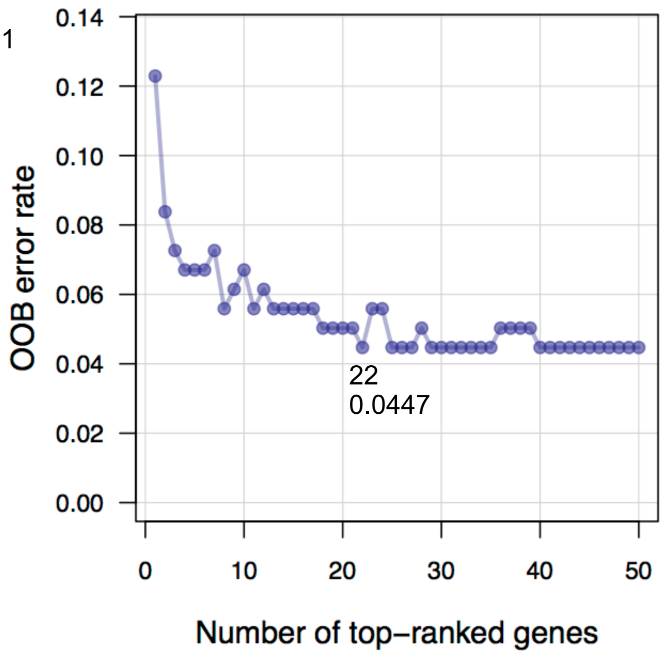




1



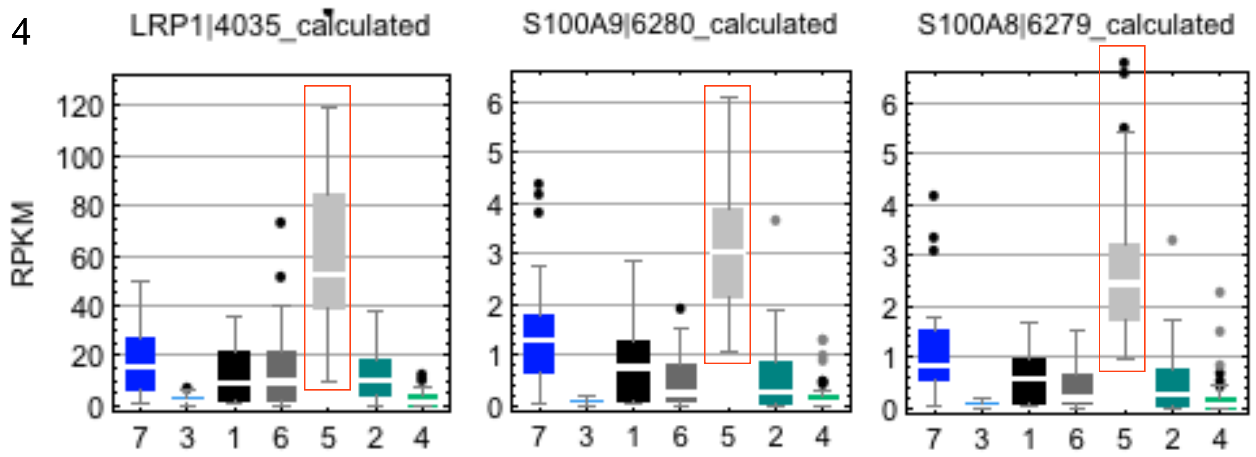
2

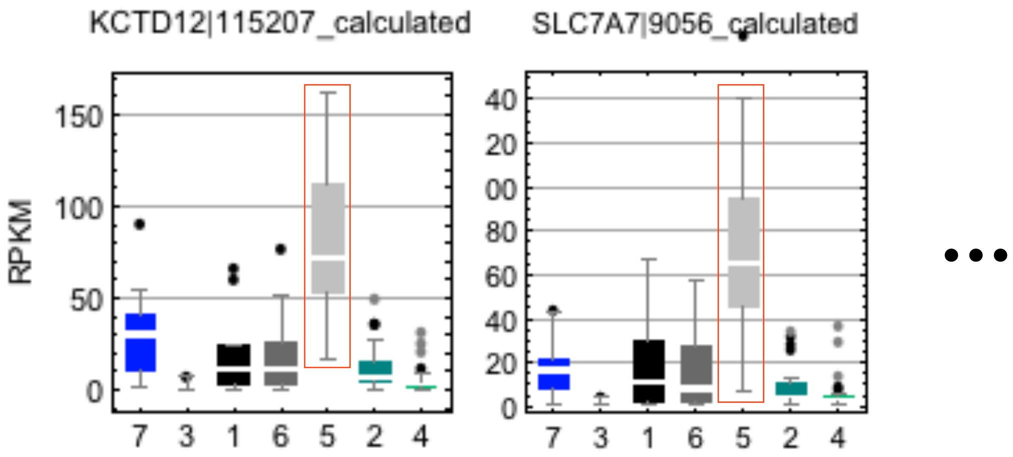


3

n	Symbol	Entrez	MD Acc	MD Gini
	LRP1	4035	6.54	1.75
	S100A9	6280	6.52	1.52
	KCTD12	115207	5.44	1.33
	SLC7A7	9056	5.47	1.27
5	S100A8	6279	6.13	1.24
	CTSS	1520	4.86	1.10
	IGSF6	10261	4.92	1.10
	P2RY13	53829	4.98	1.02
	VCAN	1462	5.20	1.00
10	TLR1	7096	5.27	0.99
	CLEC7A	64581	4.45	0.92
	MEGF9	1955	4.46	0.86
	PSAP	5660	4.11	0.80
	SIGLEC9	27180	4.21	0.77
15	CYBB	1536	4.20	0.74
	CD14	929	4.03	0.72
	OSBPL11	114885	3.99	0.71
	PLBD1	79887	5.66	0.69
	CPPED1	55313	3.91	0.59
20	TLR8	51311	3.58	0.56
	OGFRL1	79627	4.06	0.56
22	NAGA	4668	3.51	0.54

4

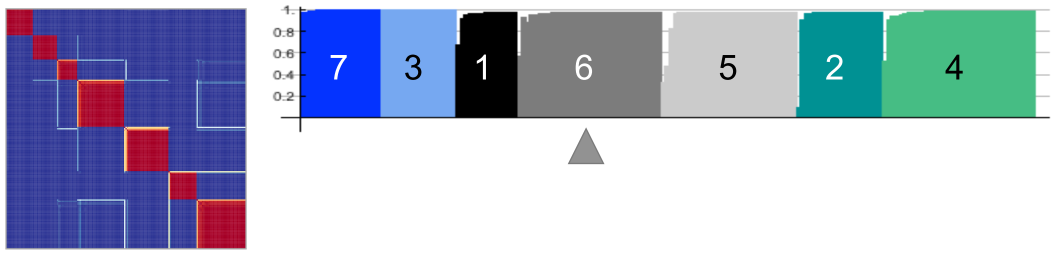




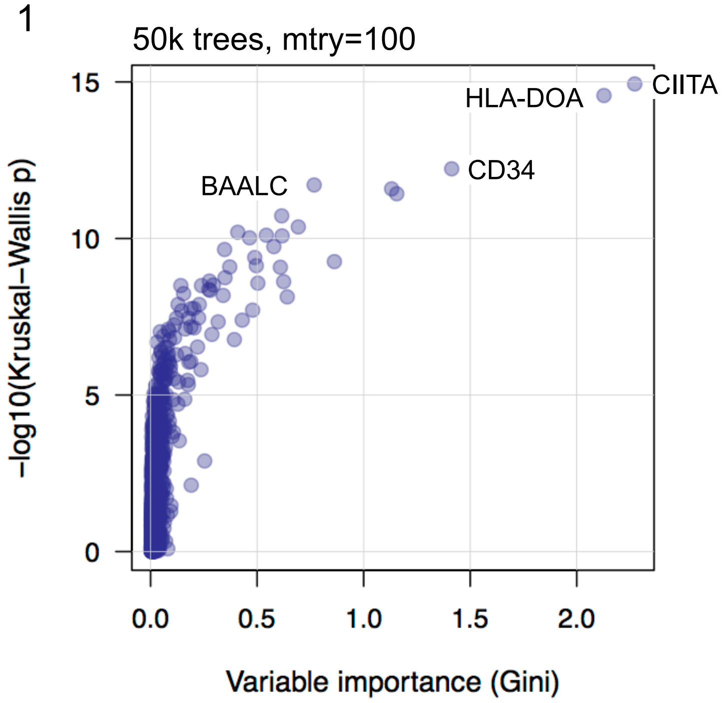
5

```
RF: classification
Number of trees: 50000
Vars tried each split: 1
Top 11 genes
OOB est error rate: 4.47%
Confusion matrix:
    0  1 class.error
0 145  1 0.0068
1   7 26 0.21
```

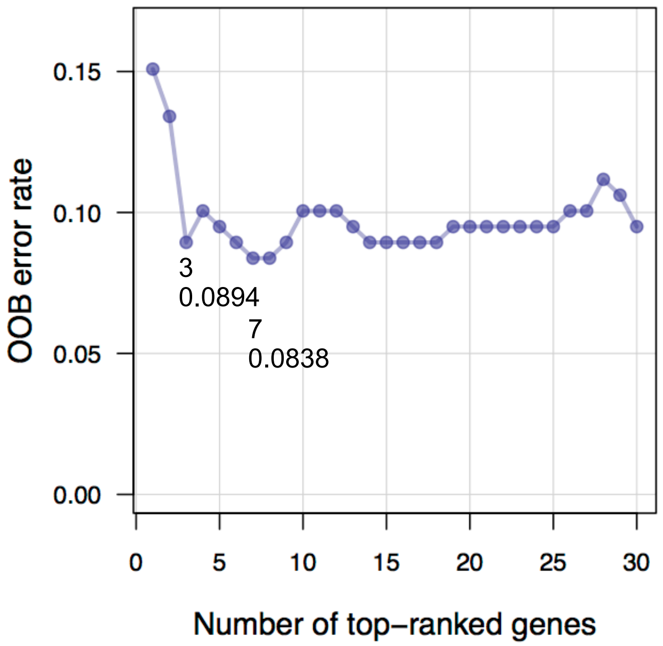
f) RNA-seq: group 6 of 7



1



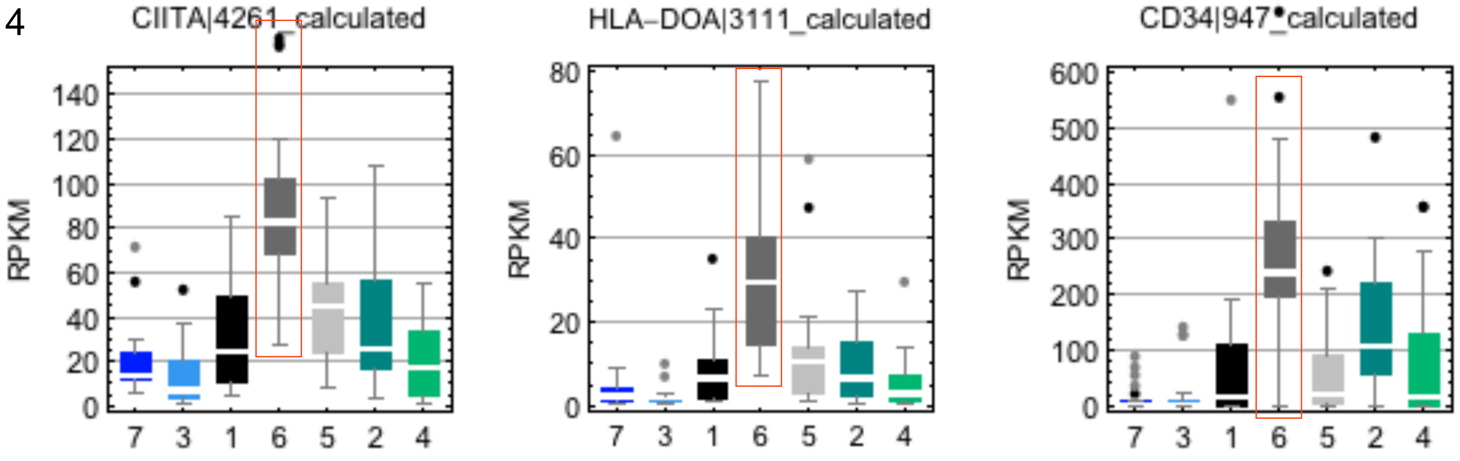
2

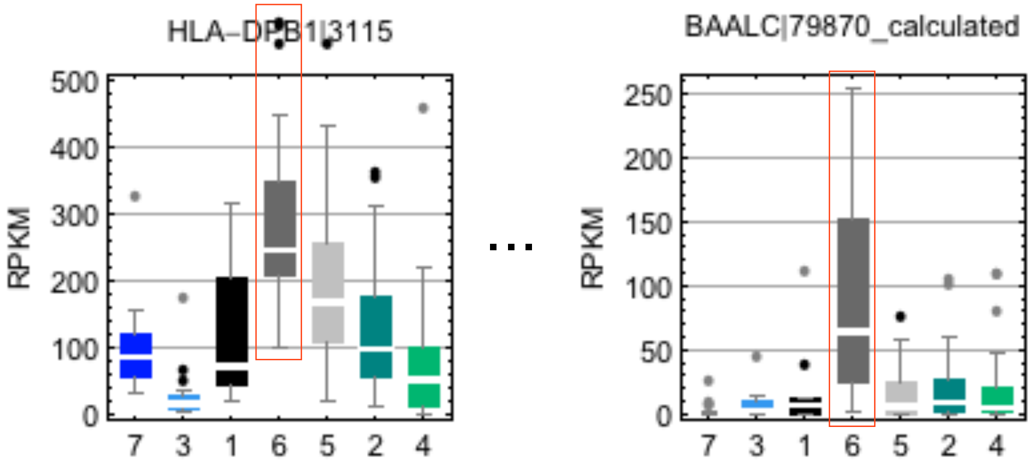


3

n	Symbol	Entrez	MD Acc	MD Gini
	CIITA	4261	8.27	2.27
	HLA-DOA	3111	7.84	2.13
3	CD34	947	6.21	1.41
	HLA-DPB1	3115	6.14	1.16
	HLA-DMA	3108	6.02	1.13
	HLA-DQA1	3117	5.12	0.86
7	BAALC	79870	4.34	0.77
	HLA-DPA1	3113	4.70	0.69
	PLEKHG2	64857	3.89	0.64
10	FLNB	2317	4.04	0.62
	NPR3	4883	3.76	0.62
	JUP	3728	4.30	0.62
	MOSC1	64757	3.99	0.61
	FAM69B	138311	3.75	0.58
15	C5orf23	79614	3.47	0.54
	UBA7	7318	3.96	0.50
	HLA-DRA	3122	4.03	0.50
	HLA-DMB	3109	4.07	0.49
	DNTT	1791	4.75	0.48
20	SLC37A1	54020	3.32	0.46

4

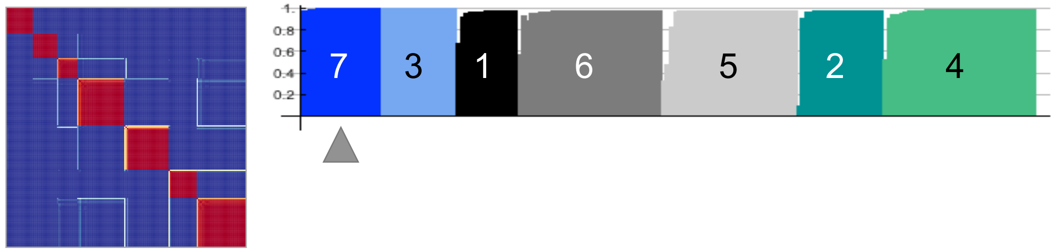




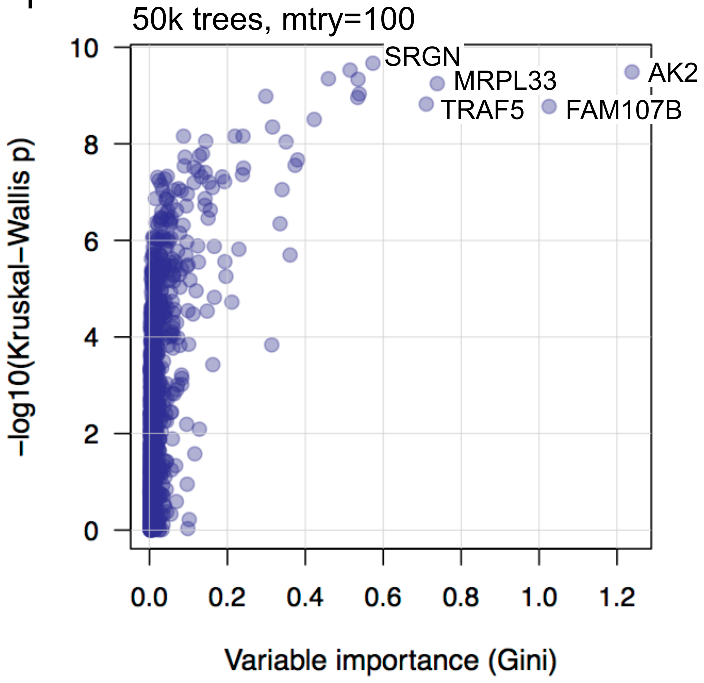
5

```
RF: classification
Number of trees: 50000
Vars tried each split: 1
Top 7 genes
OOB est err rate: 8.38%
Confusion matrix:
    0   1 class.error
0 139   5   0.035
1  10  25   0.29
```

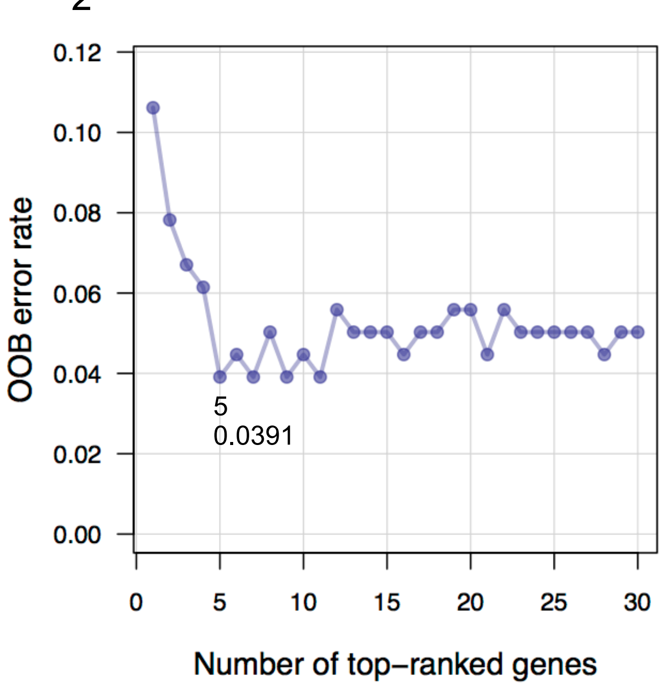
g) RNA-seq: group 7 of 7



1



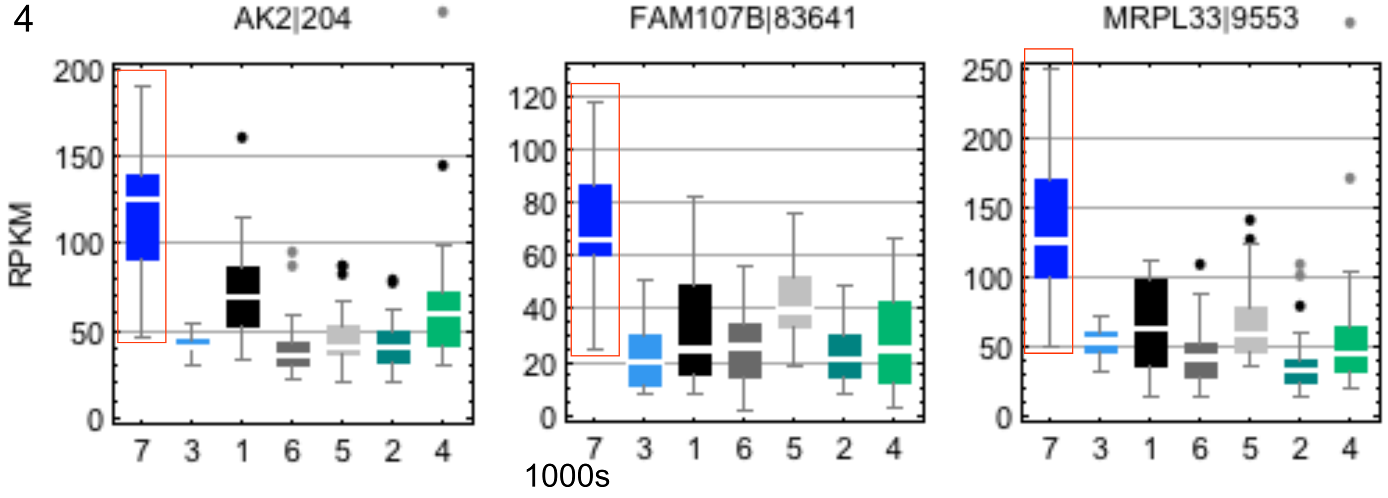
2

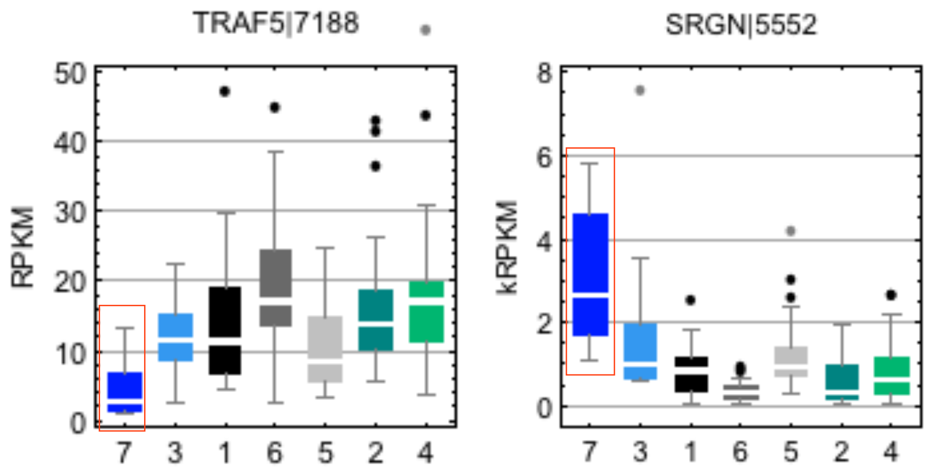


3

n	Symbol	Entrez	MD Acc	MD Gini
	AK2	204	6.71	1.24
	FAM107B	83641	5.89	1.03
	MRPL33	9553	4.98	0.74
	TRAF5	7188	5.37	0.71
5	SRGN	5552	4.36	0.57
	RNASE2	6036	4.20	0.54
	LYZ	4069	4.71	0.54
	PTPRCAP	5790	3.97	0.53
	JUP	3728	4.24	0.51
10	ATG3	64422	4.41	0.46
	TNFSF13B	10673	4.21	0.42
	CDC42EP3	10602	3.47	0.38
	SERPINB10	5273	3.91	0.37
	DNASE2	1777	3.76	0.36
15	FUT4	2526	3.28	0.35
	CCL23	6368	3.21	0.34
	TP53INP2	58476	3.20	0.34
	SERPINB8	5271	3.06	0.32
	DEFB1	1672	3.42	0.31
20	PGAM1	5223	2.77	0.30

4

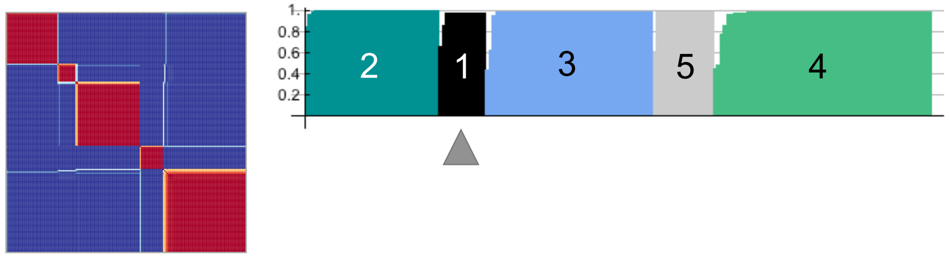




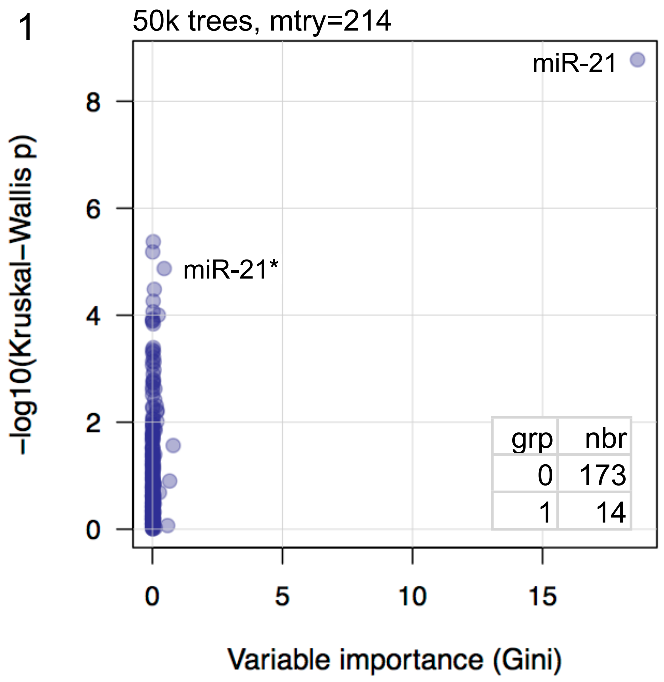
5

```
RF: classification
Number of trees: 50000
Vars tried each split: 1
Top 5 genes
OOB est err rate: 3.91%
Confusion matrix:
    0  1 class.error
0 157  2  0.013
1   5 15  0.25
```

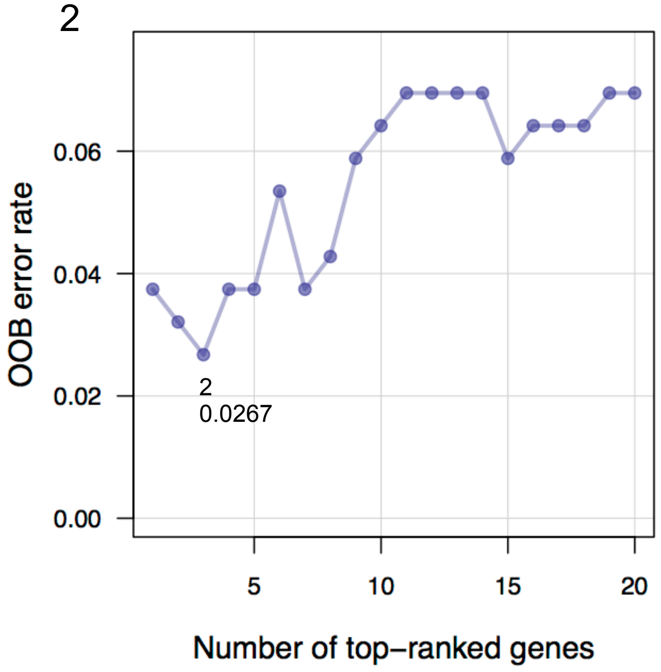
h) miRNA-seq: group 1 of 5



1



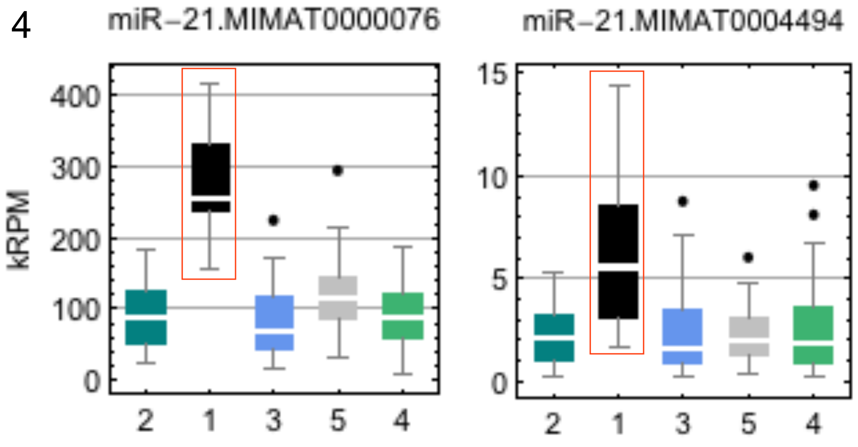
2



3

n	mir	MIMAT	MD Acc	MD Gini
1	miR-21	MIMAT0000076	151.12	13.61
	miR-21	MIMAT0004494	22.45	1.23
	miR-181a-2*	MIMAT0004558	21.44	0.97
	miR-92a	MIMAT0000092	15.17	0.72
5	miR-103	MIMAT0000101	23.38	0.60
	miR-22	MIMAT0000077	4.15	0.58
	miR-20a	MIMAT0004493	12.33	0.56
	miR-142	MIMAT0000433	1.55	0.54
	miR-148b	MIMAT0000759	-5.53	0.43
10	miR-589	MIMAT0004799	5.54	0.34
	miR-326	MIMAT0000756	6.72	0.28
	miR-125b	MIMAT0000423	0.23	0.25
	miR-32	MIMAT0000090	5.30	0.23
	miR-27a	MIMAT0000084	-3.57	0.22
15	miR-335	MIMAT0000765	-4.49	0.20
	miR-16-2	MIMAT0004518	4.98	0.13
	miR-19b	MIMAT0000074	2.32	0.12
	miR-17	MIMAT0000070	4.36	0.11
	miR-374b	MIMAT0004955	4.52	0.11
20	miR-146b	MIMAT0004766	9.36	0.10
	miR-125a	MIMAT0000443	4.39	0.10
	miR-625	MIMAT0004808	4.33	0.10
	miR-100	MIMAT0000098	3.17	0.10
	miR-30d	MIMAT0000245	-1.70	0.09
25	miR-154	MIMAT0000452	4.50	0.08
	miR-199b	MIMAT0004563	1.75	0.08
	miR-199a	MIMAT0000232	2.49	0.08
	miR-500	MIMAT0002871	-1.09	0.08
	miR-99b	MIMAT0000689	5.13	0.07
30	miR-183	MIMAT0000261	-0.86	0.07
	miR-181c	MIMAT0000258	5.86	0.07
	miR-25	MIMAT0000081	-0.08	0.07
	miR-27b	MIMAT0000419	-4.50	0.07
	miR-455	MIMAT0004784	4.11	0.07
35	miR-181d	MIMAT0002821	5.08	0.06
	miR-181a-1*	MIMAT0000270	6.24	0.06
	let-7g	MIMAT0004584	-3.87	0.06
	miR-181b	MIMAT0000257	6.76	0.05
	miR-92b	MIMAT0003218	5.90	0.05
40	let-7e	MIMAT0000066	3.10	0.05

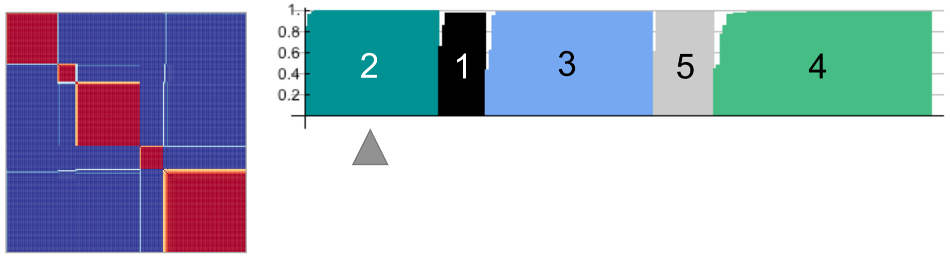
4



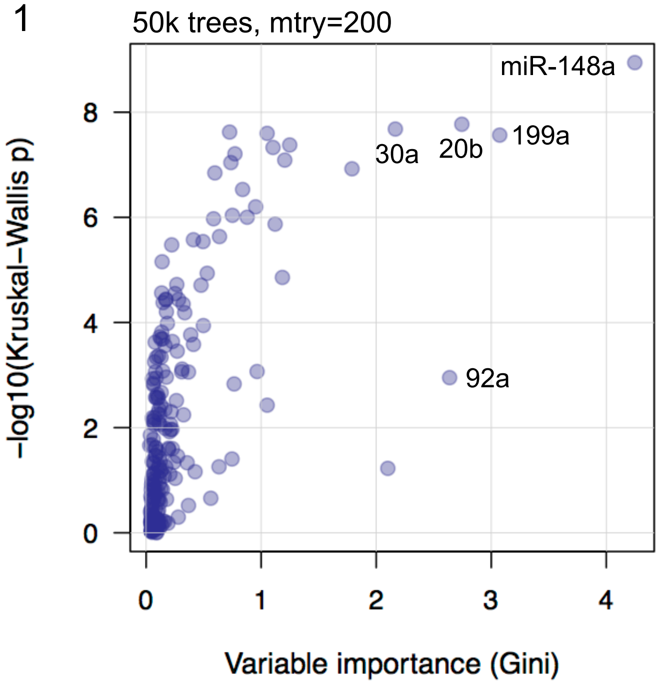
5

```
RF: classification
Number of trees: 50000
Vars tried each split: 1
Top 2 MIMATS
OOB est err: 2.67%
Confusion matrix:
    0  1 class.err
0 172  1 0.0058
1   4 10 0.29
```

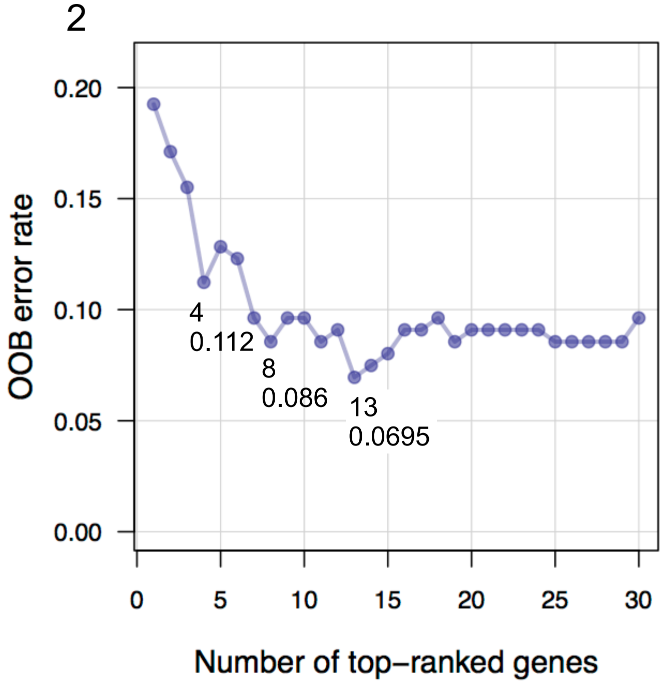
i) miRNA-seq: group 2 of 5



1



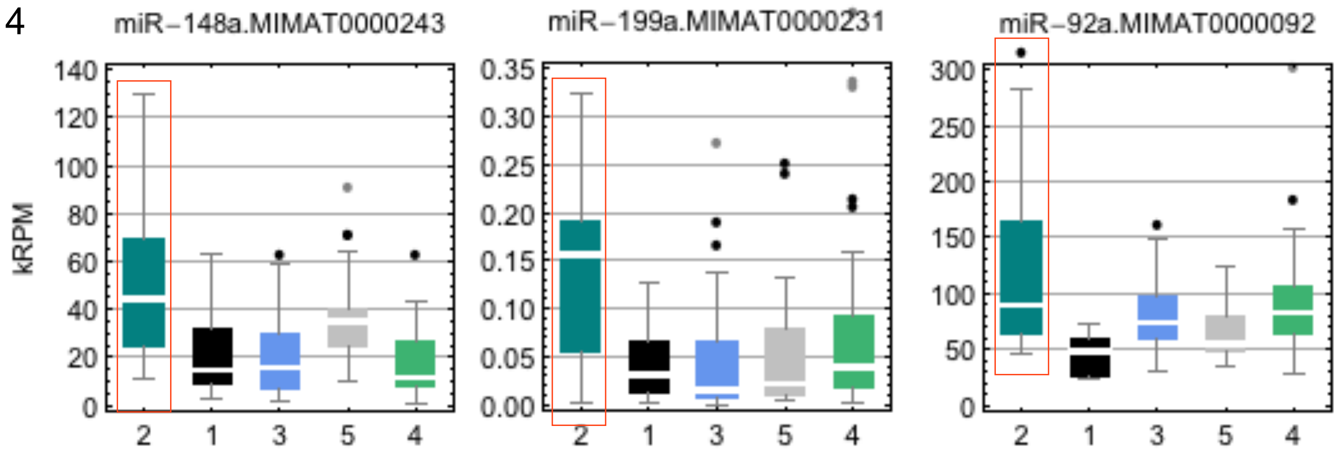
2



3

n	miR	MIMAT	MD Acc	MD Gini
1	miR-148a	MIMAT0000243	11.44	4.25
	miR-199a	MIMAT0000231	7.26	3.07
	miR-20b	MIMAT0001413	7.89	2.74
	miR-92a	MIMAT0000092	10.79	2.64
5	miR-30a	MIMAT0000088	8.74	2.17
	miR-9	MIMAT0000441	10.09	2.10
	miR-181b	MIMAT0000257	8.05	1.79
	miR-532	MIMAT0004780	4.93	1.25
	miR-30a	MIMAT0000087	6.28	1.21
10	miR-185	MIMAT0000455	5.93	1.18
	miR-181a	MIMAT0000256	5.91	1.12
	miR-181a-1*	MIMAT0000270	5.10	1.10
	miR-20a	MIMAT0004493	6.77	1.05
	miR-363	MIMAT0000707	4.58	1.05
15	miR-106a	MIMAT0000103	6.08	0.96
	miR-486	MIMAT0002177	3.53	0.95
	miR-192	MIMAT0000222	4.83	0.88
	miR-500	MIMAT0002871	3.79	0.84
	miR-182	MIMAT0000259	3.29	0.77
20	miR-425	MIMAT0001343	4.25	0.77
	miR-874	MIMAT0004911	4.11	0.75
	miR-20a	MIMAT0000075	5.54	0.75
	miR-183	MIMAT0000261	3.94	0.74
	miR-532	MIMAT0002888	3.07	0.73
25	miR-551b	MIMAT0003233	3.40	0.64
	miR-17	MIMAT0000071	4.85	0.63
	miR-151	MIMAT0000757	3.72	0.60
	miR-144	MIMAT0004600	2.85	0.59
	miR-10a	MIMAT0000253	4.63	0.56
30	miR-194	MIMAT0000460	3.98	0.53
	miR-339	MIMAT0000764	3.37	0.50
	miR-10b	MIMAT0000254	3.29	0.49
	miR-628	MIMAT0004809	4.18	0.48
	miR-92a-1	MIMAT0004507	3.09	0.43
35	miR-493	MIMAT0002813	3.13	0.41
	miR-455	MIMAT0004784	2.61	0.41
	miR-503	MIMAT0002874	3.28	0.39
	miR-191	MIMAT0000440	2.45	0.37
	miR-17	MIMAT0000070	3.45	0.37
40	miR-128	MIMAT0000424	2.66	0.36

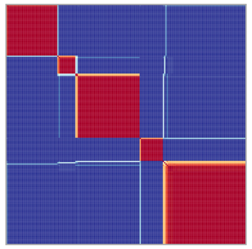
4

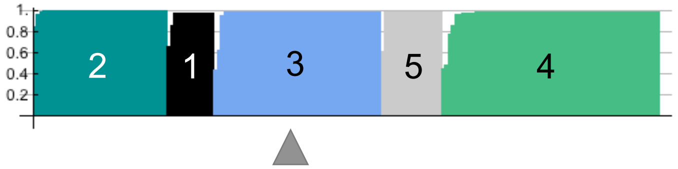


5

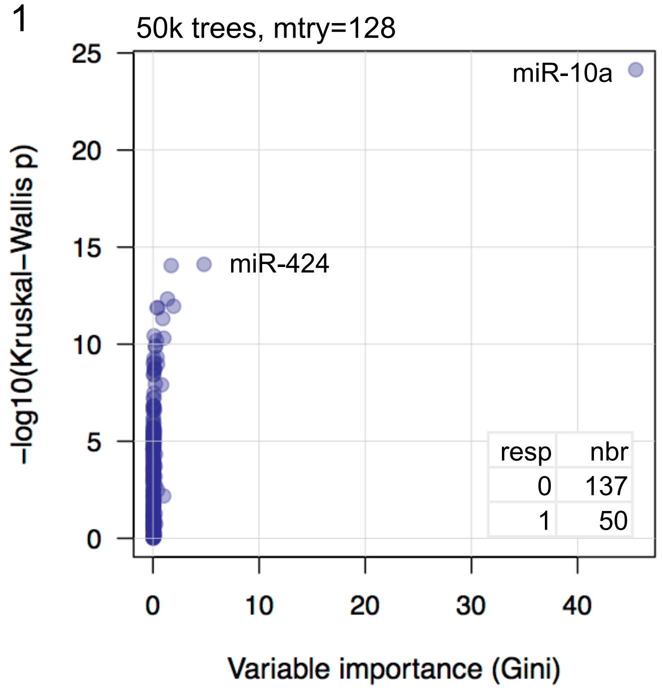
```
RF: classification
Number of trees: 50000
Vars tried each split: 1
Top 13 MIMATS
OOB est err: 6.95%
Confusion matrix:
    0  1 class.err
0 144  3  0.020
1  10 30  0.25
```

j) miRNA-seq: group 3 of 5

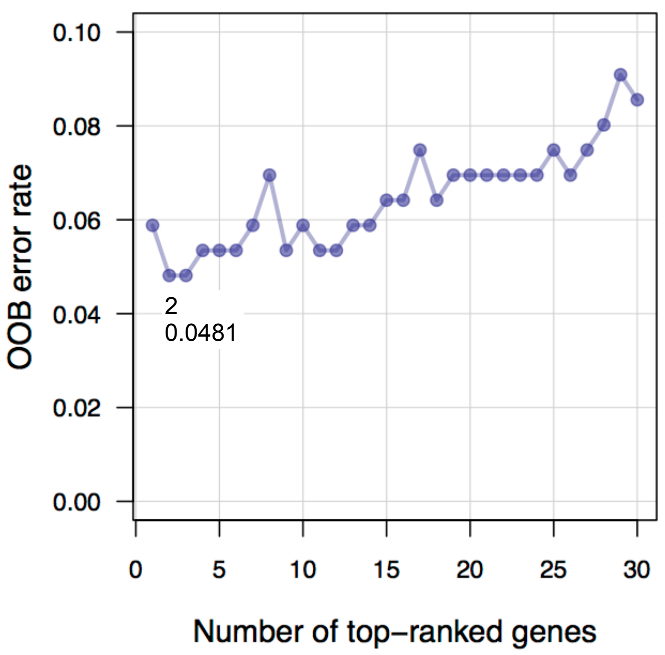




1



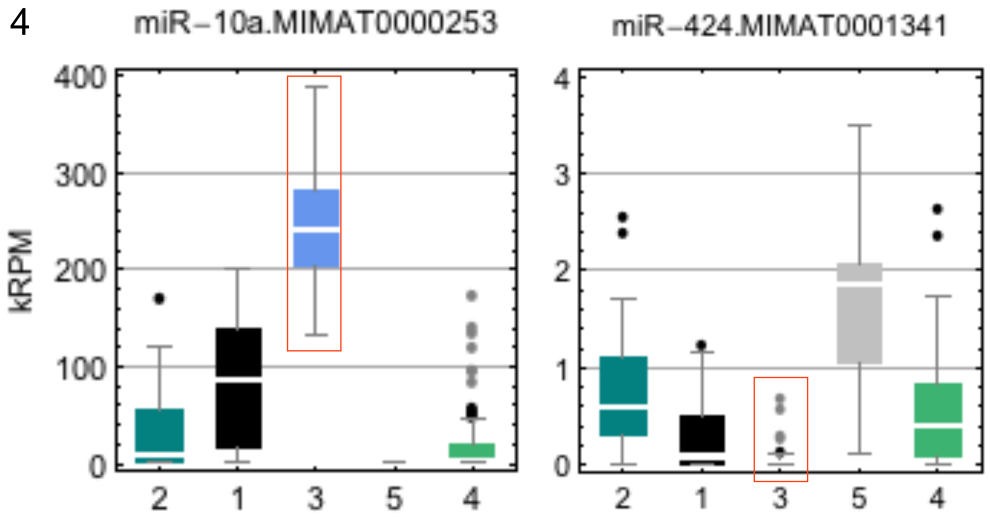
2



3

n	mir	MIMAT	MD Acc	MD Gini
1	miR-10a	MIMAT0000253	25.4223	45.4778
	miR-424	MIMAT0001341	6.5771	4.8106
	miR-193a	MIMAT0004614	7.3993	1.9423
	miR-542	MIMAT0003389	3.8524	1.7167
5	miR-196b	MIMAT0001080	5.8071	1.3714
	miR-331	MIMAT0000760	4.081	1.0211
	miR-21	MIMAT0000076	7.5922	1.0071
	let-7b	MIMAT0000063	2.919	0.922
	miR-181a-2*	MIMAT0004558	2.737	0.815
10	miR-503	MIMAT0002874	2.4028	0.4676
	miR-181c	MIMAT0004559	1.8267	0.4336
	miR-23b	MIMAT0000418	4.5886	0.4302
	let-7b	MIMAT0004482	1.6873	0.3893
	miR-103	MIMAT0000101	3.2674	0.3828
15	miR-107	MIMAT0000104	1.8236	0.335
	miR-100	MIMAT0000098	-1.0579	0.2594
	miR-32	MIMAT0000090	2.7866	0.2406
	miR-199b	MIMAT0004563	0.5418	0.2395
	let-7d	MIMAT0000065	3.4617	0.2311
20	let-7g	MIMAT0004584	-2.3083	0.2274
	miR-199a	MIMAT0000232	0.5265	0.2247
	miR-7-1	MIMAT0004553	-0.0043	0.2088
	miR-374b	MIMAT0004955	-1.0926	0.2053
	miR-142	MIMAT0000434	0.9631	0.1912
25	let-7a	MIMAT0004481	2.0138	0.1665
	miR-335	MIMAT0004703	1.2061	0.1586
	miR-106b	MIMAT0000680	-0.8357	0.144
	miR-625	MIMAT0004808	1.1086	0.1348
	miR-577	MIMAT0003242	1.9945	0.1348
30	miR-130a	MIMAT0000425	2.628	0.1327
	miR-30a	MIMAT0000087	2.7265	0.1317
	miR-133a	MIMAT0000427	1.4797	0.1308
	miR-664	MIMAT0005949	0.3344	0.1302
	miR-139	MIMAT0000250	4.0262	0.1299
35	miR-379	MIMAT0000733	2.3074	0.1287
	miR-148b	MIMAT0000759	0.8831	0.1239
	miR-30d	MIMAT0000245	1.8049	0.1191
	miR-145	MIMAT0000437	3.2158	0.1182
	miR-29-2	MIMAT0004515	-1.8803	0.1112
40	miR-98	MIMAT0000096	1.7731	0.1111

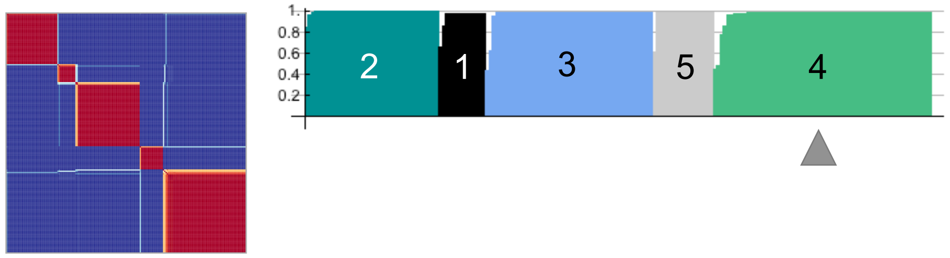
4



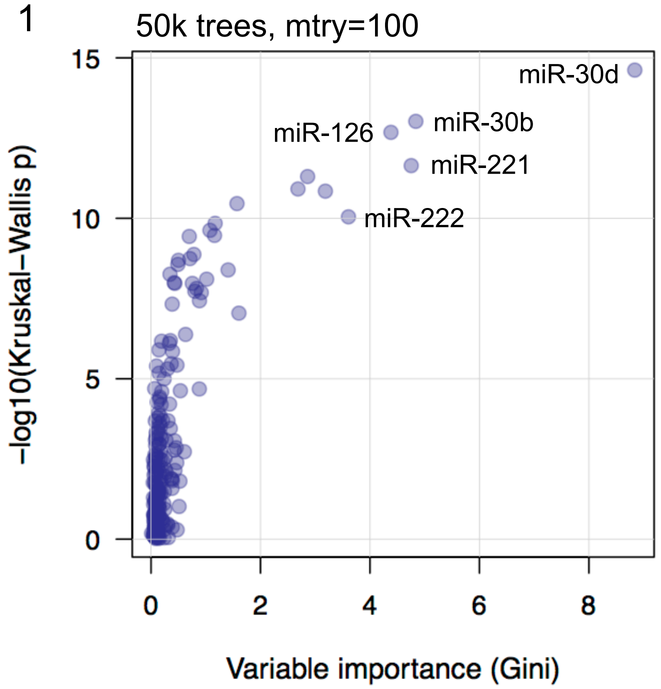
5

```
RF: classification
Number of trees: 50000
Vars tried each split: 1
Top 2 MIMATS
OOB est error rate: 4.81%
Confusion matrix:
    0  1 class.err
0 132  5   0.037
1   4 46   0.080
```

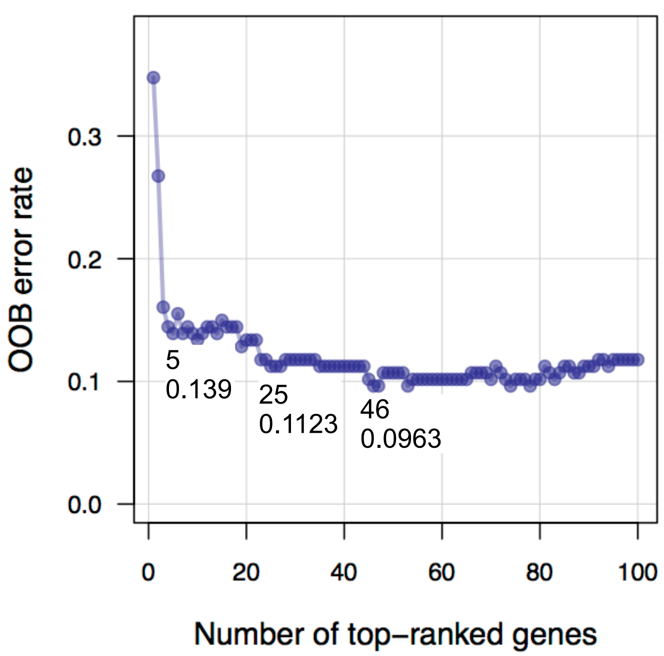
k) miRNA-seq: group 4 of 5



1



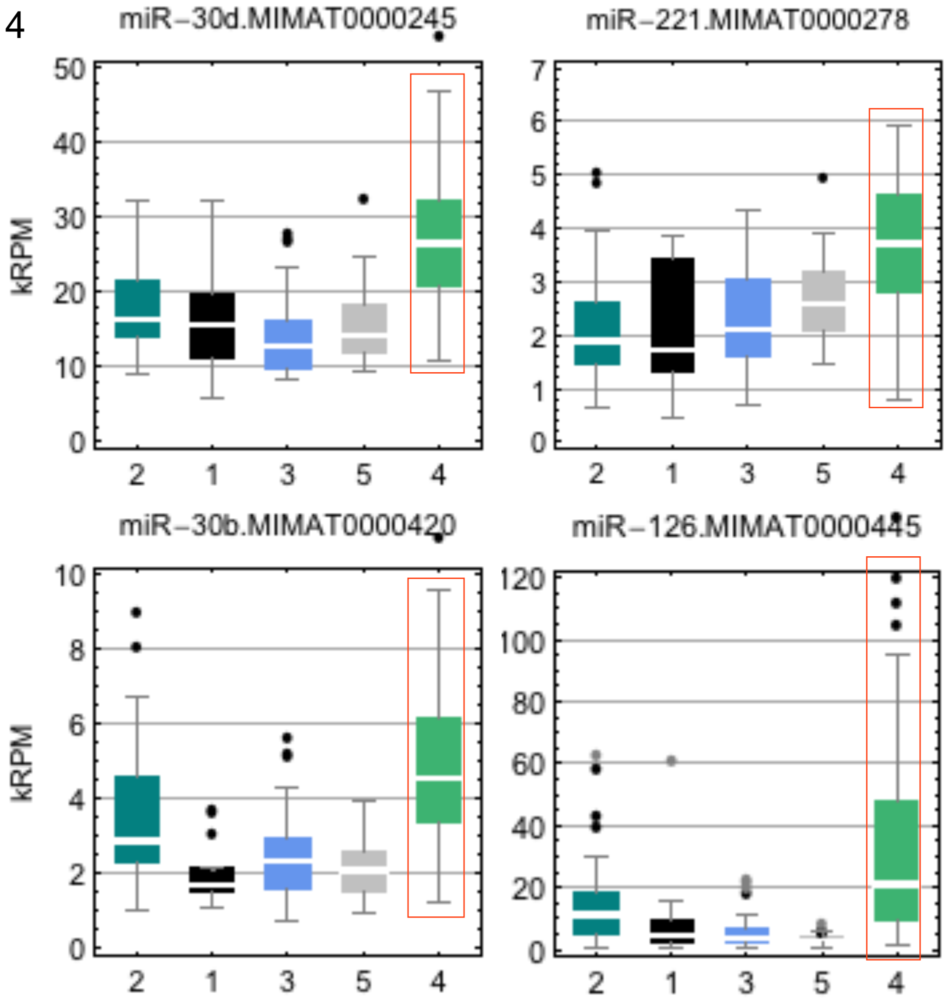
2



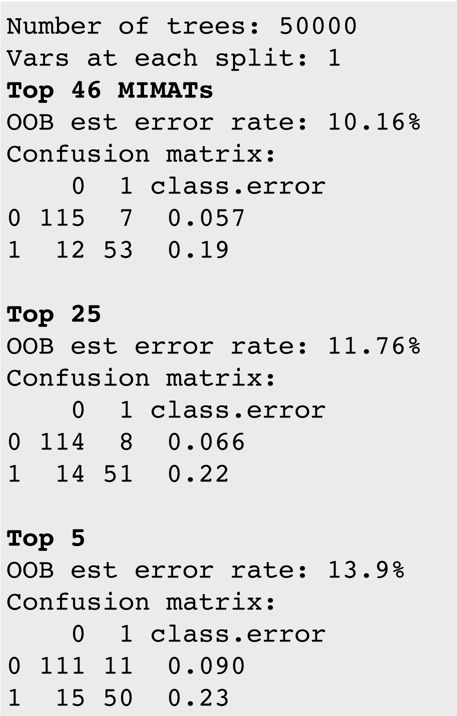
3

n	miRNA	MIMAT	MD Acc	MD Gini
	miR-30d	MIMAT0000245	12.58	8.84
	miR-30b	MIMAT0000420	9.83	4.84
	miR-221	MIMAT0000278	9.27	4.75
	miR-126	MIMAT0000445	11.18	4.38
5	miR-222	MIMAT0000279	7.76	3.61
	miR-130a	MIMAT0000425	8.85	3.19
	miR-181c	MIMAT0004559	7.97	2.86
	miR-126	MIMAT0000444	10.26	2.68
	miR-128	MIMAT0000424	8.24	1.61
10	miR-181c	MIMAT0000258	6.67	1.57
	miR-30e	MIMAT0000692	4.84	1.41
	miR-181d	MIMAT0002821	5.80	1.17
	miR-181a-2*	MIMAT0004558	4.86	1.16
	miR-181a	MIMAT0000256	5.69	1.08
	miR-505	MIMAT0002876	6.17	1.02
	miR-335	MIMAT0004703	6.14	0.92
	miR-10a	MIMAT0000253	6.06	0.89
	miR-148a	MIMAT0000243	5.70	0.88
	let-7b	MIMAT0000063	4.91	0.83
20	miR-766	MIMAT0003888	5.17	0.80
	miR-181b	MIMAT0000257	4.54	0.78
	miR-374b	MIMAT0004955	4.88	0.76
	miR-181a-1*	MIMAT0000270	4.63	0.71
	miR-331	MIMAT0000760	3.40	0.70
25	miR-142	MIMAT0000434	2.98	0.63
	miR-337	MIMAT0000754	4.20	0.61
	let-7a	MIMAT0000062	2.92	0.54
	miR-539	MIMAT0003163	4.50	0.53
	miR-376c	MIMAT0000720	3.75	0.51
30	miR-9	MIMAT0000441	3.94	0.50
	miR-146a	MIMAT0000449	4.13	0.49
	miR-100	MIMAT0000098	3.77	0.48
	miR-98	MIMAT0000096	4.55	0.47
	miR-625	MIMAT0004808	3.53	0.47
35	miR-99b	MIMAT0000689	4.21	0.46
	miR-654	MIMAT0004814	3.49	0.44
	miR-335	MIMAT0000765	2.70	0.44
	miR-584	MIMAT0003249	3.93	0.43
	miR-532	MIMAT0002888	1.58	0.43
40	miR-361	MIMAT0000703	4.12	0.42
	miR-93	MIMAT0000093	2.90	0.39
	miR-409	MIMAT0001639	3.61	0.39
	let-7a-2	MIMAT0010195	3.40	0.39
	let-7d	MIMAT0000065	3.44	0.39
45	miR-20b	MIMAT0001413	3.35	0.38
	miR-363	MIMAT0000707	3.52	0.38

4

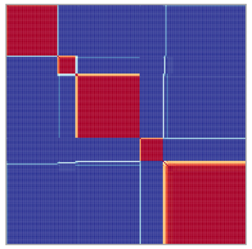


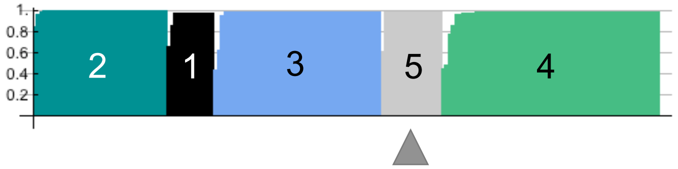
5



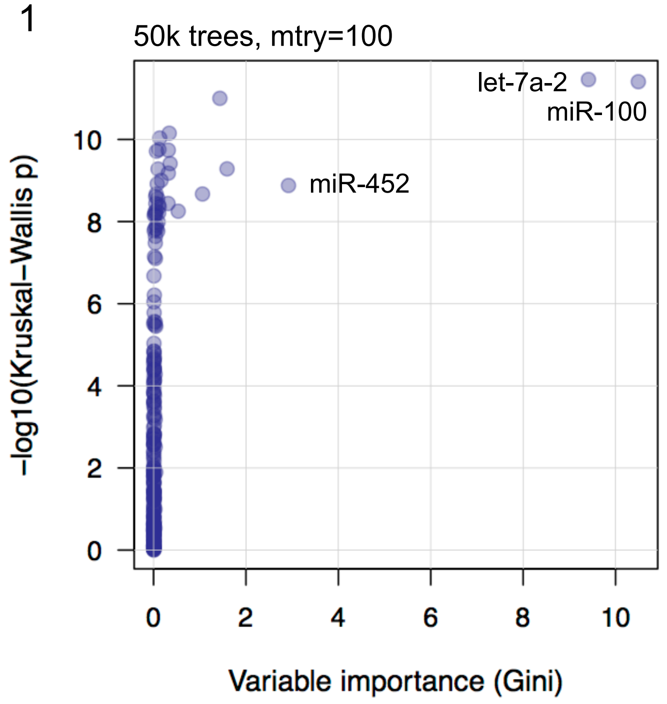
```
Number of trees: 50000
Vars at each split: 1
Top 46 MIMATs
OOB est error rate: 10.16%
Confusion matrix:
    0   1 class.error
0 115   7   0.057
1  12  53   0.19

Top 25
OOB est error rate: 11.76%
Confusion matrix:
    0   1 class.error
0 114   8   0.066
1  14  51   0.22

Top 5
OOB est error rate: 13.9%
Confusion matrix:
    0   1 class.error
0 111 11   0.090
1  15  50   0.23
```

I) miRNA-seq: group 5 of 5

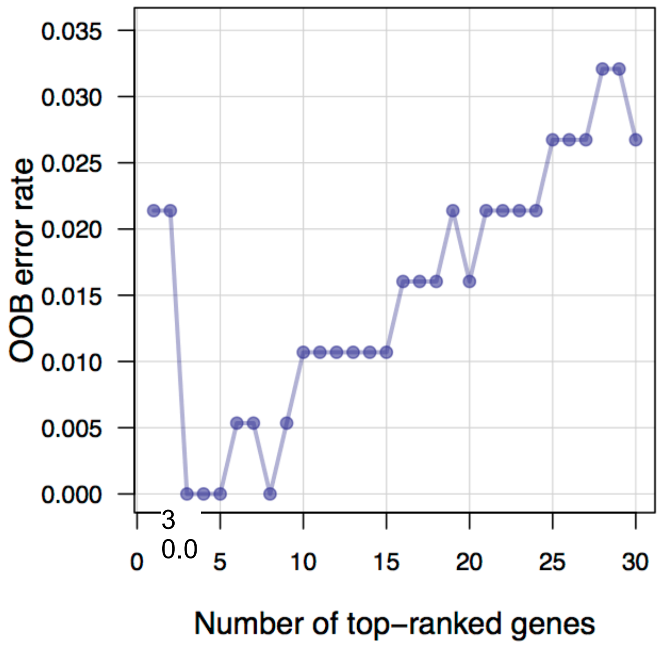




1



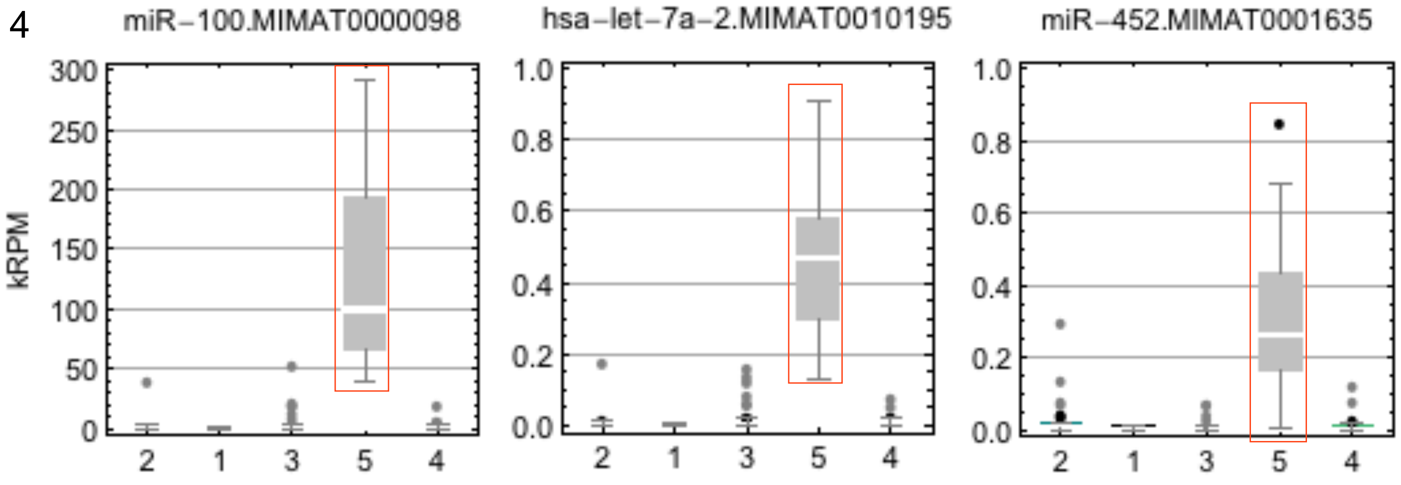
2



3

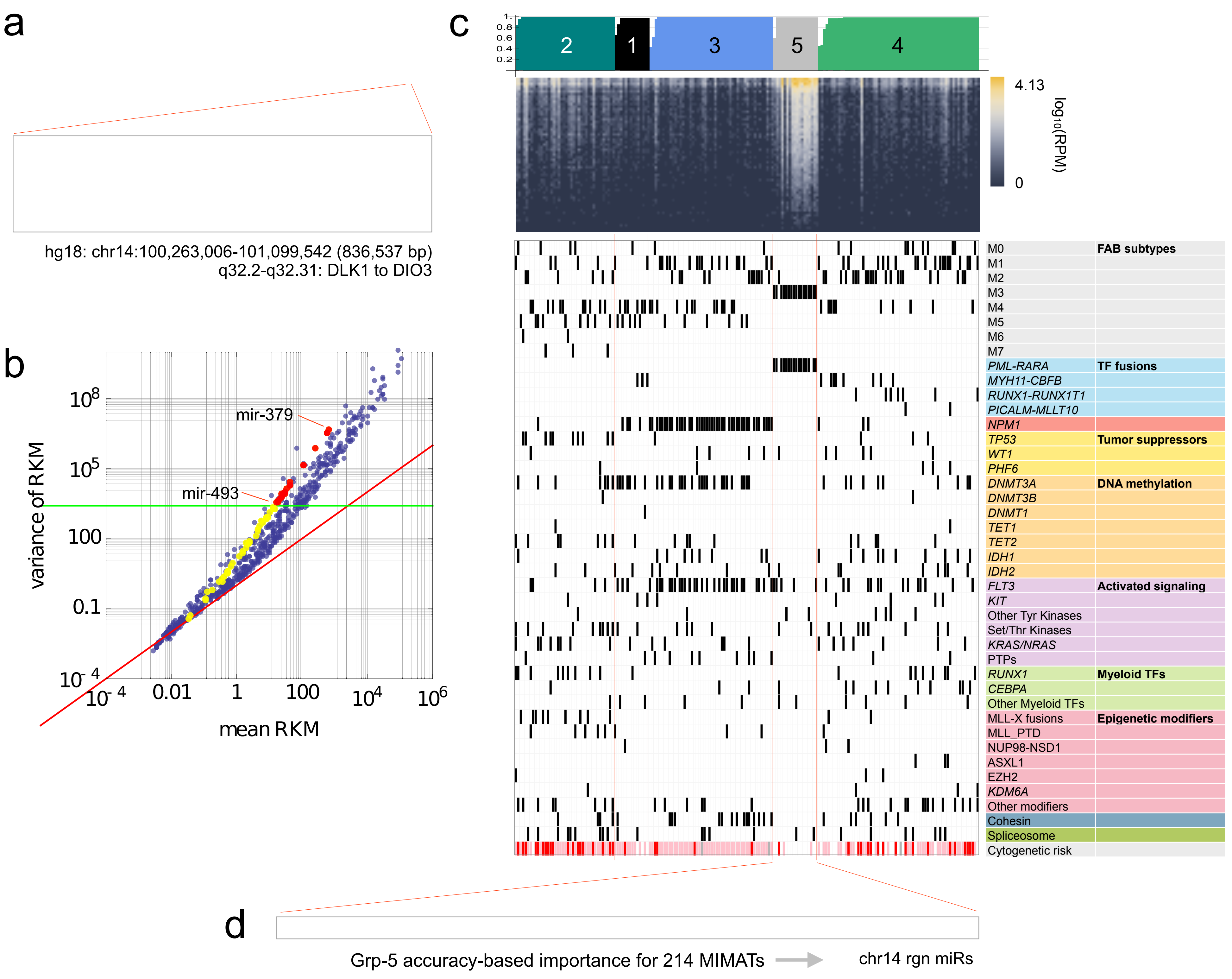
n	mir	MIMAT	MD Acc	MD Gini
1	miR-100	MIMAT0000098	15.59	10.49
	let-7a-2	MIMAT0010195	15.32	9.41
	miR-452	MIMAT0001635	8.00	2.92
	miR-224	MIMAT0000281	6.09	1.59
5	miR-125b	MIMAT0000423	7.16	1.44
	miR-193b	MIMAT0002819	6.09	1.06
	miR-889	MIMAT0004921	4.00	0.53
	miR-381	MIMAT0000736	3.77	0.36
	miR-376c	MIMAT0000720	3.63	0.34
10	miR-127	MIMAT0000446	3.58	0.32
	miR-136	MIMAT0004606	3.58	0.32
	miR-487b	MIMAT0003180	3.48	0.31
	miR-379	MIMAT0000733	2.77	0.16
	miR-493	MIMAT0003161	3.21	0.13
15	miR-431	MIMAT0004757	2.08	0.12
	miR-136	MIMAT0000448	2.60	0.12
	miR-654	MIMAT0004814	2.41	0.11
	miR-369	MIMAT0000721	2.35	0.11
	miR-10a	MIMAT0000253	2.83	0.10
20	miR-196b	MIMAT0001080	3.11	0.10
	miR-382	MIMAT0000737	2.46	0.09
	miR-370	MIMAT0000722	2.34	0.09
	miR-337	MIMAT0000754	2.16	0.08
	miR-409	MIMAT0001639	2.02	0.06
25	miR-154	MIMAT0000452	2.05	0.06
	miR-495	MIMAT0002817	1.84	0.06
	miR-223	MIMAT0000280	1.55	0.06
	miR-493	MIMAT0002813	1.58	0.06
	miR-339	MIMAT0000764	1.75	0.05
30	miR-181b	MIMAT0000257	2.66	0.05
	miR-363	MIMAT0000707	1.96	0.05
	miR-16-2	MIMAT0004518	2.17	0.05
	miR-127	MIMAT0004604	1.99	0.05
	miR-181a	MIMAT0000256	2.47	0.04
35	miR-199b	MIMAT0000263	1.99	0.04
	miR-539	MIMAT0003163	1.83	0.04
	miR-574	MIMAT0003239	1.73	0.04
	miR-92a-1	MIMAT0004507	2.12	0.04
	miR-320b	MIMAT0005792	1.77	0.04
40	miR-23a	MIMAT0000078	1.56	0.04

4



5

```
RF: classification
Number of trees: 50000
Vars tried each split: 1
Top 3 MIMATS
OOB est error rate: 0%
    0  1 class.err
0 169  0         0
1   0 18         0
```



e

mir.MIMAT	RPM mn (desc)	RPM variance	Var Imp rank
mir-379.MIMAT0000733	656.118	4.42E+06	14
mir-127.MIMAT0000446	580.587	3.46E+06	12
mir-134.MIMAT0000447	576.029	3.50E+06	34
mir-127.MIMAT0004604	251.965	795835	29
mir-381.MIMAT0000736	113.962	137485	8
mir-654.MIMAT0004814	108.838	137051	16
mir-136.MIMAT0000448	44.196	25635.7	13
mir-337.MIMAT0000754	41.9572	19855.5	17
mir-409.MIMAT0001639	40.9521	19509.9	19
mir-493.MIMAT0002813	34.747	14266.7	22
mir-495.MIMAT0002817	31.3997	10716	24
mir-376c.MIMAT0000720	29.3294	9902.31	9
mir-382.MIMAT0000737	28.5922	8644.23	20
mir-369.MIMAT0000721	26.645	7703.15	15
mir-136.MIMAT0004606	26.1565	8399.74	10
mir-431.MIMAT0004757	22.9447	7766.7	27
mir-539.MIMAT0003163	22.9294	5607.44	28
mir-154.MIMAT0000452	22.8352	6301.88	26
mir-487b.MIMAT0003180	21.154	5050.46	11
mir-323.MIMAT0000755	18.9954	4338.84	33
mir-889.MIMAT0004921	18.7535	3766.91	7
mir-493.MIMAT0003161	17.2745	3598.54	18
mir-485.MIMAT0002176	14.2023	1928.75	
mir-410.MIMAT0002171	13.6656	2175.81	
mir-409.MIMAT0001638	12.4815	1851.69	
mir-494.MIMAT0002816	10.2659	1304.31	
mir-432.MIMAT0002814	8.64855	804.562	
mir-411.MIMAT0003329	8.49853	799.679	
mir-758.MIMAT0003879	8.34291	774.264	
mir-369.MIMAT0001621	7.84822	693.054	
mir-655.MIMAT0003331	6.49073	469.361	
mir-431.MIMAT0001625	5.84699	536.751	
mir-299.MIMAT0002890	5.17319	330.057	
mir-654.MIMAT0003330	4.46793	208.065	
mir-377.MIMAT0004689	4.29314	253.909	
mir-496.MIMAT0002818	3.7839	120.147	
mir-411.MIMAT0004813	2.49763	81.214	
mir-377.MIMAT0000730	2.34512	65.2719	
mir-154.MIMAT0000453	2.31341	55.9208	
mir-433.MIMAT0001627	2.02848	66.7281	
mir-543.MIMAT0004954	2.00455	42.1723	
mir-656.MIMAT0003332	1.50216	25.44	
mir-376b.MIMAT0002172	1.46191	27.4015	
mir-379.MIMAT0004690	1.21407	17.0127	
mir-380.MIMAT0000735	0.734576	8.35261	
mir-380.MIMAT0000734	0.633931	6.99461	
mir-299.MIMAT0000687	0.539148	3.97514	
mir-487a.MIMAT0002178	0.512895	4.14629	
mir-541.MIMAT0004920	0.506782	3.35512	
mir-337.MIMAT0004695	0.505363	3.59005	
mir-485.MIMAT0002175	0.410875	2.34445	
mir-770.MIMAT0003948	0.366411	1.51216	
mir-668.MIMAT0003881	0.354173	1.99106	
mir-1247.MIMAT0005899	0.277022	1.36718	
mir-432.MIMAT0002815	0.177604	0.63608	
mir-323.MIMAT0004696	0.124711	0.56057	
mir-665.MIMAT0004952	0.118311	0.2936	
mir-1197.MIMAT0005955	0.111728	0.2361	
mir-541.MIMAT0004919	0.03918	0.0537	
mir-412.MIMAT0002170	0.03446	0.0389	
mir-544.MIMAT0003164	0.00485	0.00440	
mir-300.MIMAT0004903	0	0	

Figure S17. MiRNAs in the 14q DLK1 to DIO3 locus.
a) The 0.84 Mb genomic region between DKL1 and DIO3 on chromosome 14 contains clusters of snoRNAs (blue) and miRNAs (red). b) The overdispersed (Robinson and Smyth 2007) mean-variance relationship for normalized abundance (RPM) for the 706 miRBase v13 mature and star strands ('MIMATs') that had a nonzero mean RPM across the 187 samples. The slope of the red line is 1.0. The miRNAs in the chr14 region (yellow and red points, and listed in (e) in order of decreasing mean RPM) lie close to a straight line, and those with a mean above ~10 RPM have atypically high variances. The input to NMF clustering and to the classifier was the abundance matrix for the 214 MIMATs whose RPM variance was above the 75th variance percentile (horizontal green line); only 22 of the MIMATs in the chromosome 14 region passed this threshold (black text in (e)). c) Most of the MIMATs in the chr14 region were relatively abundant in most M3 (group 5) samples. d,e) The horizontal bar schematically shows variable importance for a random forest classifier for group 5, increasing from L to R. Of 214 MIMATs ranked by classifier accuracy for group 5 samples, all 22 from the chromosome 14 region that were input to NMF and the classifier were ranked in the top 34, consistent with the relatively high variances in (a) and the abundance heatmap in (e).

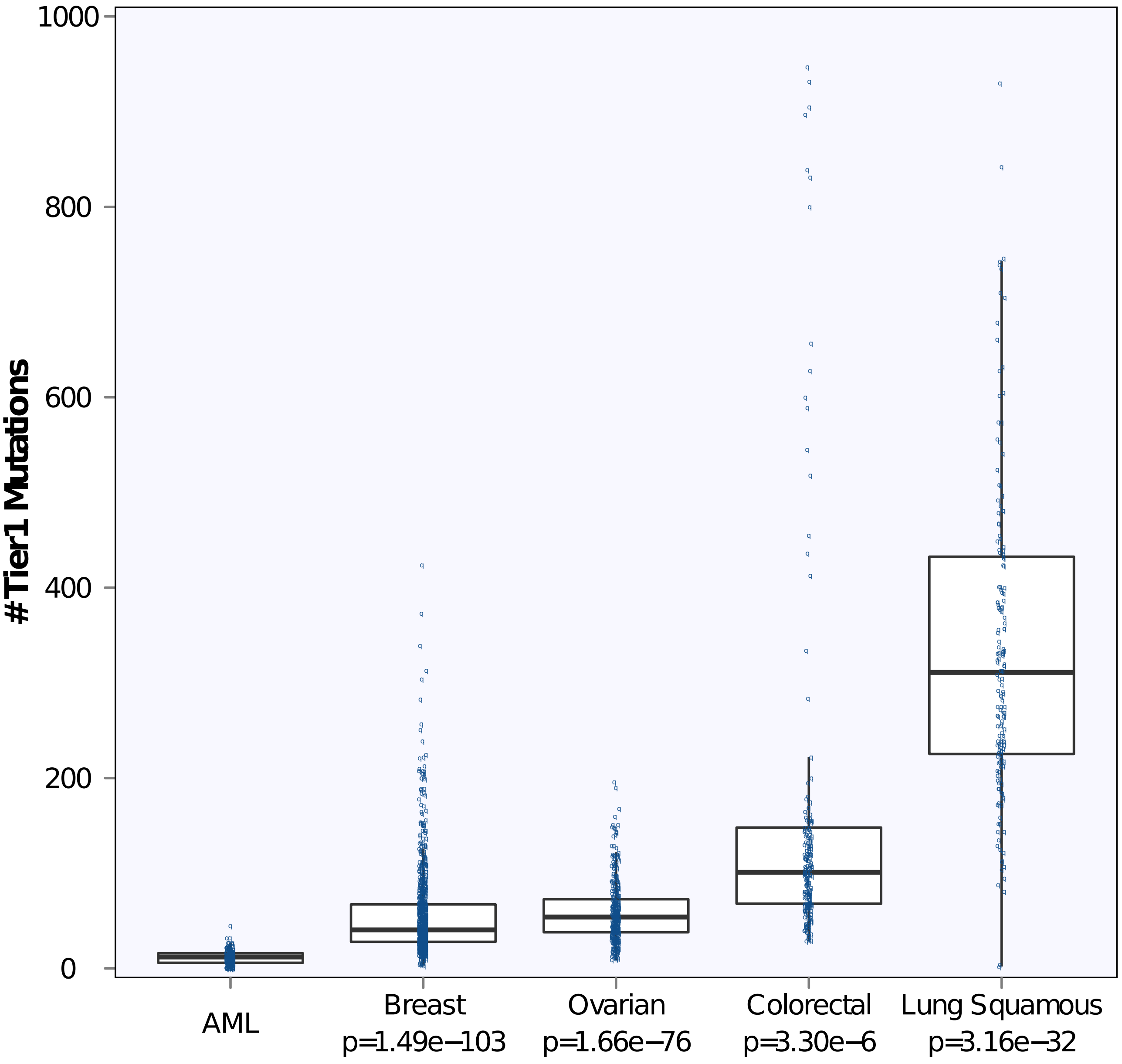


Figure S18. Numbers of Tier 1 (coding) mutations in different tumor types analyzed by TCGA.
Comparison of tier 1 mutation numbers across 5 published tumor types from The Cancer Genome Atlas. The mutational burden in AML is significantly lower than each of the other types (p values refer to comparisons of each tumor type to AML). All tumors analyzed were included in the calculations, but 4 lung and 21 colorectal cases were omitted from the plot, since they had greater than 1000 mutations (maximum of 12,411). The box plots identify the median values along with the 25th and 75th percentile.

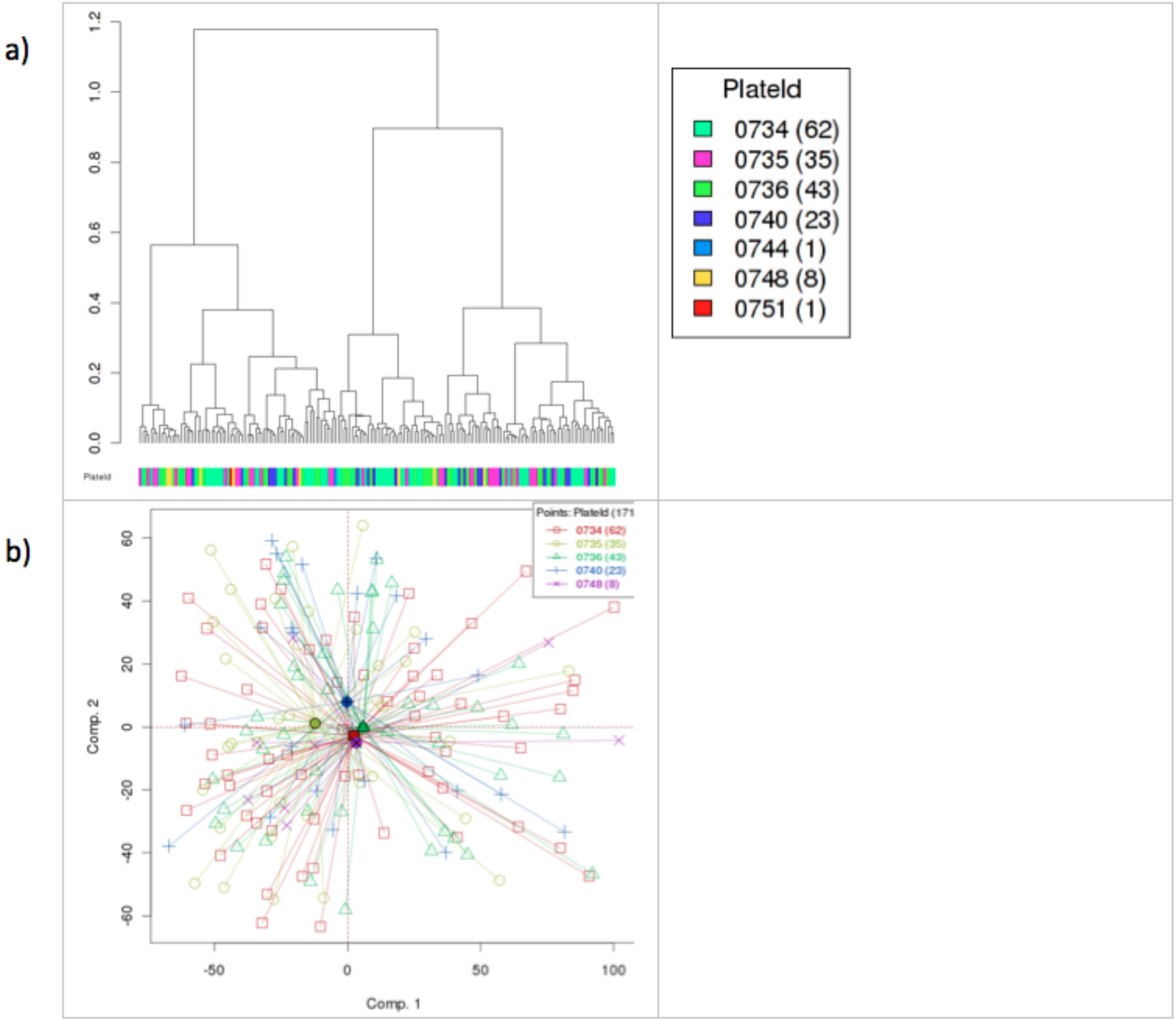


Figure S19. Batch Effects in mRNA-seq
a) Hierarchical clustering for mRNA expression from RNA-seq data b) PCA: First two principal components for RNA-seq, with samples connected by centroids according to plate ID.

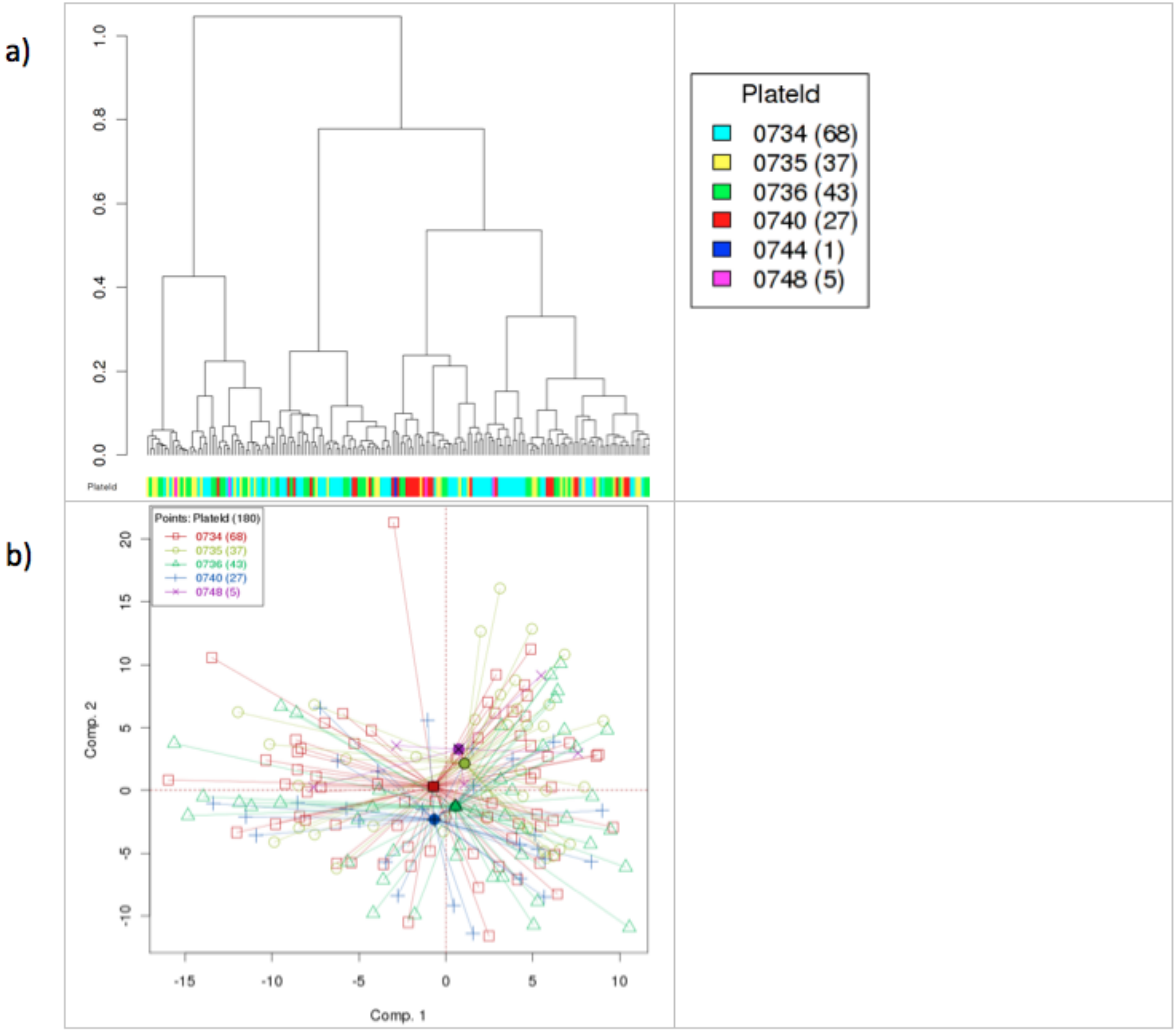


Figure S20. Batch effects in miRNA-seq

a) Hierarchical clustering of samples for miRNA expression from miRNA-seq data. b) PCA: First two principal components for miRNA expression from miRNA-seq data, with samples connected by centroids according to plate ID.

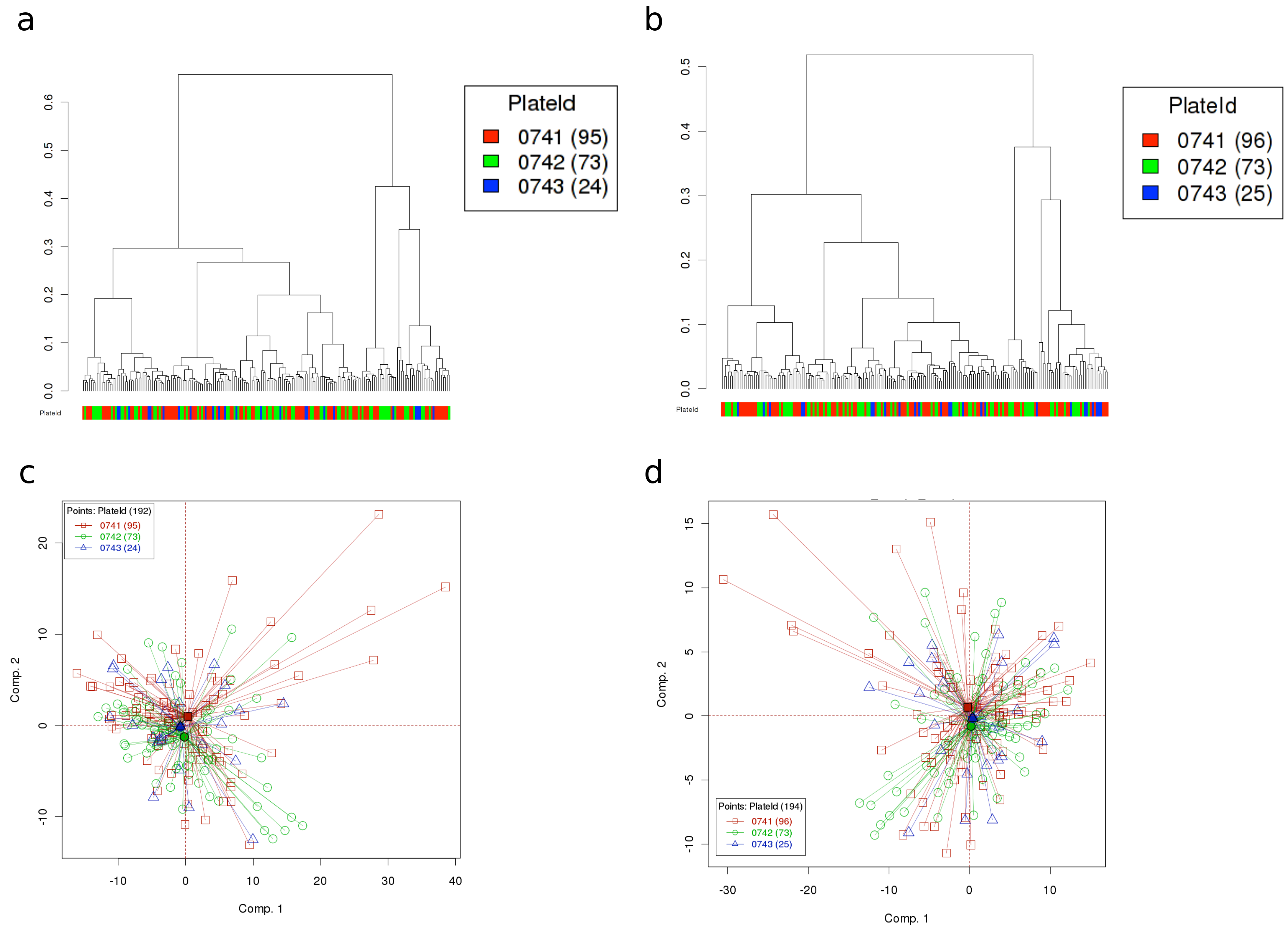


Figure S21. Batch effects in methylation arrays
a) Hierarchical clustering plot for DNA methylation b) HM27 data Hierarchical clustering plot for DNA methylation HM450 data c) PCA for DNA methylation HM27, with samples connected by centroids according to plate ID. d) PCA for DNA methylation HM450, with samples connected by centroids according to plate ID.

C. Supplementary Tables

Note: Some large tables are hosted on the TCGA DCC site:
https://tcga-data.nci.nih.gov/docs/publications/laml_2012/

Table S1: Clinical matrix
See separate xlsx:
https://tcga-data.nci.nih.gov/docs/publications/aml_2012/SupplementalTable01.xlsx

Per-sample table containing information on sex, race, cytogenetics, WBC, Fusions, SVs, risk groups, and a listing of all mutated genes across the cohort. Copy number amplifications are noted by a red background color, deletions are blue, and UPD events are green. Percent AML in skin (Column M) was determined by assaying the frequency of somatic mutations in the normal sample. The number of subclones (Column N) was determined as described in Supplementary Methods section A.3.11.

Table S2: Multivariate Survival statistics

a) Event Free Survival

Covariate	**Parameter Estimate**	**Standard Error**	**chi-square p-value**	**Hazard Ratio**	**95% CI**
Basic Model (stratifying by Age < 60 vs Age >= 60)					
WBC > 16	0.53	0.17	0.002	1.70	1.23 - 2.37
Cytogenetic Classification=Good	-0.75	0.27	0.006	0.47	0.28 - 0.80
Cytogenetic Classification=Poor	0.41	0.19	0.03	1.51	1.04 - 2.21
Terms added separately to the base model					
TP-53	0.86	0.35	0.01	2.37	1.20 - 4.71
DNMT3A	0.23	0.19	0.22	1.26	0.87 - 1.84
FLT3	0.39	0.19	0.04	1.47	1.10 - 2.15
PML-RARA	-0.14	0.49	0.77	0.87	0.33 - 2.25
MYH11-CBFB	-0.21	0.53	0.69	0.81	0.29 - 2.27
RUNX1-RUNX1T1	0.34	0.58	0.55	1.41	0.46 - 4.37
NUP98-NSD1	0.54	0.80	0.37	1.72	0.53 - 5.58
Final Model					
WBC > 16	0.57	0.18	0.002	1.76	1.23 - 2.52
Cytogenetic Classification=Good	-0.75	0.27	0.006	0.47	0.28 - 0.81
Cytogenetic Classification=Poor	0.30	0.22	0.18	1.35	0.87 - 2.09
TP-53	0.89	0.35	0.01	2.43	1.23 - 4.83
FLT3	0.41	0.19	0.04	1.50	1.03 - 2.20

b) Overall Survival

Covariate	**Parameter Estimate**	**Standard Error**	**chi-square p-value**	**Hazard Ratio**	**95% CI**
Basic Model (stratifying by Age < 60 vs Age >= 60)					
WBC > 16	0.35	0.17	0.04	1.43	1.02 - 2.00
Cytogenetic Classification=Good	-0.75	0.30	0.001	0.47	0.30 - 0.83
Cytogenetic Classification=Poor	0.46	0.20	0.02	1.58	1.07 - 2.33
Terms added separately to the base model					
TP-53	0.96	0.35	0.01	2.61	1.30 - 5.23
DNMT3A	0.34	0.20	0.08	1.41	0.96 - 2.08
FLT3	0.34	0.20	0.09	1.41	0.95 - 2.10
PML-RARA	-0.37	0.53	0.48	0.69	0.24 - 1.94
MYH11-CBFB	-0.18	0.56	0.76	0.84	0.28 - 2.53
RUNX1-RUNX1T1	0.65	0.59	0.27	1.92	0.61 - 6.09
NUP98-NSD1	0.81	0.61	0.18	2.26	0.69 - 7.41
Other Modifiers	0.16	0.24	0.52	1.17	0.73 - 1.87
Final Model					
WBC > 16	0.49	0.18	0.01	1.63	1.13 - 2.33
Cytogenetic Classification=Good	-0.76	0.29	0.01	0.47	0.27 - 0.83
Cytogenetic Classification=Poor	0.24	0.22	0.27	1.28	0.83 - 1.98
TP53	0.96	0.35	0.01	2.61	1.30 - 5.23

Table S3: Sequencing Coverage

Average genomic coverage for each tumor and normal sample

Sample	Mean Depth of Coverage	Sample	Mean Depth of Coverage
TCGA-AB-2802-03B-01W-0728-08	77.21	TCGA-AB-2904-03A-01W-0732-08	152.77
TCGA-AB-2802-11B-01W-0728-08	92.93	TCGA-AB-2904-11A-01W-0732-08	101.25
TCGA-AB-2803-03B-01W-0728-08	73.07	TCGA-AB-2905-03A-01D-0739-09	40.80
TCGA-AB-2803-11B-01W-0728-08	90.77	TCGA-AB-2905-11A-01D-0739-09	25.30
TCGA-AB-2804-03B-01W-0728-08	191.97	TCGA-AB-2906-03A-01D-0739-09	28.29
TCGA-AB-2804-11B-01W-0728-08	182.64	TCGA-AB-2906-11A-01D-0739-09	23.56
TCGA-AB-2805-03B-01W-0728-08	210.13	TCGA-AB-2907-03A-01D-0739-09	28.53
TCGA-AB-2805-11B-01W-0728-08	179.27	TCGA-AB-2907-11A-01D-0739-09	33.33
TCGA-AB-2806-03B-01W-0728-08	80.88	TCGA-AB-2908-03A-01W-0745-08	56.47
TCGA-AB-2806-11B-01W-0728-08	85.16	TCGA-AB-2908-11A-01W-0745-08	196.17
TCGA-AB-2807-03D-01W-0755-09	105.54	TCGA-AB-2909-03A-01W-0755-09	132.34
TCGA-AB-2807-11D-01W-0755-09	117.96	TCGA-AB-2909-11A-01W-0755-09	117.88
TCGA-AB-2808-03D-01W-0755-09	133.50	TCGA-AB-2910-03A-01W-0745-08	201.18
TCGA-AB-2808-11D-01W-0755-09	134.02	TCGA-AB-2910-11A-01W-0745-08	65.40
TCGA-AB-2809-03D-01W-0755-09	113.62	TCGA-AB-2911-03A-01W-0732-08	146.21
TCGA-AB-2809-11D-01W-0755-09	132.68	TCGA-AB-2911-11A-01W-0732-08	97.63
TCGA-AB-2810-03B-01W-0728-08	189.13	TCGA-AB-2912-03A-01W-0732-08	158.85
TCGA-AB-2810-11B-01W-0728-08	180.31	TCGA-AB-2912-11A-01W-0761-09	166.40
TCGA-AB-2811-03B-01W-0728-08	90.06	TCGA-AB-2913-03A-01W-0732-08	100.54
TCGA-AB-2811-11B-01W-0728-08	88.25	TCGA-AB-2913-11A-01W-0732-08	130.24
TCGA-AB-2812-03B-01W-0728-08	206.34	TCGA-AB-2914-03A-01W-0732-08	123.15
TCGA-AB-2812-11B-01W-0728-08	389.81	TCGA-AB-2914-11A-01W-0732-08	115.04
TCGA-AB-2813-03B-01W-0728-08	168.75	TCGA-AB-2915-03A-01W-0745-08	75.21
TCGA-AB-2813-11B-01W-0728-08	156.13	TCGA-AB-2915-11A-01W-0745-08	167.48
TCGA-AB-2814-03D-01W-0755-09	105.47	TCGA-AB-2916-03A-01W-0732-08	92.66

TCGA-AB-2814-11D-01W-0755-09	104.83	TCGA-AB-2916-11A-01W-0732-08	100.00
TCGA-AB-2815-03B-01W-0728-08	175.39	TCGA-AB-2917-03A-01W-0732-08	73.22
TCGA-AB-2815-11B-01W-0728-08	174.79	TCGA-AB-2917-11A-01W-0732-08	20.66
TCGA-AB-2816-03B-01W-0728-08	69.68	TCGA-AB-2918-03A-01W-0745-08	76.48
TCGA-AB-2816-11B-01W-0728-08	91.92	TCGA-AB-2918-11A-01W-0761-09	168.68
TCGA-AB-2817-03B-01W-0728-08	177.51	TCGA-AB-2919-03A-01W-0745-08	189.75
TCGA-AB-2817-11B-01W-0728-08	168.90	TCGA-AB-2919-11A-01W-0745-08	171.34
TCGA-AB-2818-03B-01W-0728-08	80.59	TCGA-AB-2920-03A-01W-0732-08	126.78
TCGA-AB-2818-11B-01W-0728-08	84.61	TCGA-AB-2920-11A-01W-0732-08	103.18
TCGA-AB-2819-03B-01W-0728-08	75.49	TCGA-AB-2921-03A-01W-0755-09	119.27
TCGA-AB-2819-11B-01W-0728-08	91.58	TCGA-AB-2921-11A-01W-0755-09	129.34
TCGA-AB-2820-03B-01W-0728-08	202.87	TCGA-AB-2922-03A-01W-0745-08	172.77
TCGA-AB-2820-11B-01W-0728-08	210.83	TCGA-AB-2922-11A-01W-0745-08	189.03
TCGA-AB-2821-03B-01W-0728-08	162.06	TCGA-AB-2923-03A-01W-0745-08	197.14
TCGA-AB-2821-11B-01W-0728-08	185.90	TCGA-AB-2923-11A-01W-0745-08	178.76
TCGA-AB-2822-03D-01W-0755-09	109.02	TCGA-AB-2924-03A-01W-0745-08	176.74
TCGA-AB-2822-11D-01W-0755-09	115.18	TCGA-AB-2924-11A-01W-0745-08	138.11
TCGA-AB-2823-03B-01W-0728-08	211.97	TCGA-AB-2925-03A-01W-0732-08	141.53
TCGA-AB-2823-11B-01W-0728-08	210.39	TCGA-AB-2925-11A-01W-0732-08	133.49
TCGA-AB-2824-03B-01W-0728-08	222.68	TCGA-AB-2926-03A-01W-0732-08	153.91
TCGA-AB-2824-11B-01W-0728-08	212.77	TCGA-AB-2926-11A-01W-0761-09	122.81
TCGA-AB-2825-03D-01W-0755-09	98.23	TCGA-AB-2927-03A-01W-0755-09	126.68
TCGA-AB-2825-11D-01W-0755-09	88.40	TCGA-AB-2927-11A-01W-0755-09	115.35
TCGA-AB-2826-03B-01W-0728-08	233.34	TCGA-AB-2928-03A-01W-0745-08	182.69
TCGA-AB-2826-11B-01W-0728-08	232.28	TCGA-AB-2928-11A-01W-0745-08	209.04
TCGA-AB-2827-03B-01W-0728-08	281.03	TCGA-AB-2929-03A-01W-0732-08	175.14
TCGA-AB-2827-11B-01W-0728-08	247.45	TCGA-AB-2929-11A-01W-0732-08	160.64
TCGA-AB-2828-03C-01W-0761-09	162.34	TCGA-AB-2930-03A-01W-0761-09	142.13
TCGA-AB-2828-11B-01W-0728-08	253.19	TCGA-AB-2930-11A-01W-0745-08	203.37
TCGA-AB-2829-03B-01W-0728-08	262.44	TCGA-AB-2931-03A-01W-0745-08	71.09
TCGA-AB-2829-11B-01W-0728-08	240.69	TCGA-AB-2931-11A-01W-0745-08	186.38

TCGA-AB-2830-03B-01W-0728-08	267.74	TCGA-AB-2932-03A-01W-0745-08	58.86
TCGA-AB-2830-11B-01W-0728-08	259.51	TCGA-AB-2932-11A-01W-0745-08	77.84
TCGA-AB-2831-03A-01W-0726-08	216.83	TCGA-AB-2933-03A-01W-0732-08	182.34
TCGA-AB-2831-11A-01W-0727-08	213.51	TCGA-AB-2933-11A-01W-0732-08	148.51
TCGA-AB-2832-03B-01W-0728-08	267.58	TCGA-AB-2934-03A-01W-0755-09	149.90
TCGA-AB-2832-11B-01W-0729-08	254.80	TCGA-AB-2934-11A-01W-0755-09	121.94
TCGA-AB-2833-03B-01W-0728-08	206.10	TCGA-AB-2935-03A-01W-0755-09	134.23
TCGA-AB-2833-11B-01W-0729-08	253.40	TCGA-AB-2935-11A-01W-0755-09	158.64
TCGA-AB-2834-03B-01W-0728-08	217.09	TCGA-AB-2936-03A-01W-0745-08	78.45
TCGA-AB-2834-11B-01W-0729-08	187.42	TCGA-AB-2936-11A-01W-0745-08	195.29
TCGA-AB-2835-03B-01W-0728-08	204.42	TCGA-AB-2937-03A-01W-0732-08	161.86
TCGA-AB-2835-11B-01W-0729-08	215.12	TCGA-AB-2937-11A-01W-0732-08	172.04
TCGA-AB-2836-03B-01W-0728-08	217.68	TCGA-AB-2938-03A-01W-0732-08	210.58
TCGA-AB-2836-11B-01W-0729-08	216.62	TCGA-AB-2938-11A-01W-0732-08	173.37
TCGA-AB-2837-03B-01W-0728-08	203.46	TCGA-AB-2939-03A-01W-0745-08	85.71
TCGA-AB-2837-11B-01W-0729-08	209.80	TCGA-AB-2939-11A-01W-0745-08	79.39
TCGA-AB-2838-03A-01W-0726-08	398.36	TCGA-AB-2940-03A-01W-0733-08	198.63
TCGA-AB-2838-11A-01W-0727-08	268.03	TCGA-AB-2940-11A-01W-0732-08	203.07
TCGA-AB-2839-03B-01W-0728-08	76.28	TCGA-AB-2941-03A-01W-0745-08	61.30
TCGA-AB-2839-11B-01W-0729-08	87.55	TCGA-AB-2941-11A-01W-0745-08	156.66
TCGA-AB-2840-03D-01W-0755-09	124.06	TCGA-AB-2942-03A-01W-0733-08	212.71
TCGA-AB-2840-11D-01W-0755-09	104.31	TCGA-AB-2942-11A-01W-0732-08	187.52
TCGA-AB-2841-03B-01W-0728-08	231.13	TCGA-AB-2943-03A-01W-0745-08	54.64
TCGA-AB-2841-11B-01W-0729-08	216.17	TCGA-AB-2943-11A-01W-0745-08	112.07
TCGA-AB-2842-03A-01W-0726-08	274.93	TCGA-AB-2944-03A-01W-0755-09	124.80
TCGA-AB-2842-11A-01W-0727-08	390.73	TCGA-AB-2944-11A-01W-0755-09	120.72
TCGA-AB-2843-03B-01W-0728-08	219.93	TCGA-AB-2945-03A-01W-0733-08	189.87
TCGA-AB-2843-11B-01W-0729-08	211.73	TCGA-AB-2945-11A-01W-0732-08	178.04
TCGA-AB-2844-03B-01W-0728-08	223.46	TCGA-AB-2946-03A-01W-0755-09	112.71
TCGA-AB-2844-11B-01W-0729-08	221.50	TCGA-AB-2946-11A-01W-0755-09	121.64
TCGA-AB-2845-03D-01W-0755-09	122.74	TCGA-AB-2947-03A-01W-0745-08	177.40

TCGA-AB-2845-11D-01W-0755-09	105.27	TCGA-AB-2947-11A-01W-0745-08	68.52
TCGA-AB-2846-03B-01W-0728-08	184.79	TCGA-AB-2948-03A-01W-0755-09	122.10
TCGA-AB-2846-11B-01W-0729-08	190.46	TCGA-AB-2948-11A-01W-0755-09	122.45
TCGA-AB-2847-03B-01W-0728-08	195.12	TCGA-AB-2949-03A-01W-0733-08	174.09
TCGA-AB-2847-11B-01W-0729-08	187.91	TCGA-AB-2949-11A-01W-0732-08	163.16
TCGA-AB-2848-03B-01W-0728-08	194.63	TCGA-AB-2950-03A-01W-0733-08	177.30
TCGA-AB-2848-11B-01W-0729-08	192.54	TCGA-AB-2950-11A-01W-0732-08	181.95
TCGA-AB-2849-03B-01W-0728-08	192.95	TCGA-AB-2952-03A-01W-0733-08	172.64
TCGA-AB-2849-11B-01W-0729-08	177.39	TCGA-AB-2952-11A-01W-0732-08	183.02
TCGA-AB-2850-03B-01W-0728-08	92.68	TCGA-AB-2954-03A-01W-0733-08	198.44
TCGA-AB-2850-11B-01W-0729-08	70.37	TCGA-AB-2954-11A-01W-0732-08	190.63
TCGA-AB-2851-03B-01W-0728-08	214.01	TCGA-AB-2955-03A-01W-0733-08	194.77
TCGA-AB-2851-11B-01W-0729-08	203.38	TCGA-AB-2955-11A-01W-0732-08	367.66
TCGA-AB-2853-03D-01W-0755-09	182.34	TCGA-AB-2956-03A-01W-0733-08	196.67
TCGA-AB-2853-11D-01W-0755-09	124.08	TCGA-AB-2956-11A-01W-0732-08	206.57
TCGA-AB-2854-03B-01W-0728-08	183.47	TCGA-AB-2957-03A-01W-0733-08	192.96
TCGA-AB-2854-11B-01W-0729-08	192.60	TCGA-AB-2957-11A-01W-0732-08	197.97
TCGA-AB-2855-03B-01W-0728-08	205.64	TCGA-AB-2959-03A-01W-0733-08	164.91
TCGA-AB-2855-11B-01W-0729-08	203.96	TCGA-AB-2959-11A-01W-0732-08	194.59
TCGA-AB-2856-03A-01W-0726-08	172.94	TCGA-AB-2963-03A-01D-0739-09	25.46
TCGA-AB-2856-11A-01W-0727-08	170.25	TCGA-AB-2963-11A-01D-0739-09	26.62
TCGA-AB-2857-03B-01W-0728-08	207.85	TCGA-AB-2964-03A-01D-0739-09	25.61
TCGA-AB-2857-11B-01W-0729-08	208.65	TCGA-AB-2964-11A-01D-0739-09	31.46
TCGA-AB-2858-03D-01W-0755-09	142.90	TCGA-AB-2965-03A-01D-0739-09	35.17
TCGA-AB-2858-12A-01W-0755-09	119.35	TCGA-AB-2965-11A-01D-0739-09	28.21
TCGA-AB-2859-03B-01W-0728-08	225.96	TCGA-AB-2966-03A-01D-0739-09	26.45
TCGA-AB-2859-11B-01W-0729-08	210.48	TCGA-AB-2966-11A-01D-0739-09	25.88
TCGA-AB-2860-03B-01W-0728-08	229.16	TCGA-AB-2967-03A-01D-0739-09	29.00
TCGA-AB-2860-11B-01W-0729-08	224.24	TCGA-AB-2967-11A-01D-0739-09	30.69
TCGA-AB-2861-03B-01W-0728-08	225.32	TCGA-AB-2968-03A-01D-0739-09	39.93
TCGA-AB-2861-11B-01W-0729-08	219.88	TCGA-AB-2968-11A-01D-0739-09	39.46

TCGA-AB-2862-03B-01W-0728-08	207.80
TCGA-AB-2862-11B-01W-0729-08	104.43
TCGA-AB-2863-03D-01W-0755-09	112.86
TCGA-AB-2863-11B-01W-0729-08	240.98
TCGA-AB-2864-03D-01W-0755-09	119.47
TCGA-AB-2864-11D-01W-0755-09	103.03
TCGA-AB-2865-03B-01W-0728-08	90.76
TCGA-AB-2865-11B-01W-0729-08	74.23
TCGA-AB-2866-03B-01W-0728-08	221.43
TCGA-AB-2866-11B-01W-0729-08	212.64
TCGA-AB-2867-03B-01W-0728-08	239.43
TCGA-AB-2867-11B-01W-0729-08	189.40
TCGA-AB-2868-03B-01W-0728-08	82.16
TCGA-AB-2868-11B-01W-0729-08	83.46
TCGA-AB-2869-03A-01W-0761-09	198.76
TCGA-AB-2869-11A-01W-0732-08	231.64
TCGA-AB-2870-03A-01W-0732-08	224.57
TCGA-AB-2870-11A-01W-0732-08	210.88
TCGA-AB-2871-03A-01W-0732-08	210.19
TCGA-AB-2871-11A-01W-0732-08	216.07
TCGA-AB-2872-03A-01W-0732-08	220.65
TCGA-AB-2872-11A-01W-0761-09	157.59
TCGA-AB-2873-03A-01W-0732-08	221.89
TCGA-AB-2873-11A-01W-0732-08	227.09
TCGA-AB-2874-03A-01W-0732-08	220.08
TCGA-AB-2874-11A-01W-0732-08	221.60
TCGA-AB-2875-03A-01W-0732-08	229.85
TCGA-AB-2875-11A-01W-0732-08	208.51
TCGA-AB-2876-03A-01W-0732-08	221.03
TCGA-AB-2876-11A-01W-0732-08	218.05
TCGA-AB-2877-03A-01W-0732-08	225.15
TCGA-AB-2969-03A-01D-0739-09	32.27
TCGA-AB-2969-11A-01D-0739-09	40.57
TCGA-AB-2970-03A-01D-0739-09	29.70
TCGA-AB-2970-11A-01D-0739-09	29.90
TCGA-AB-2971-03A-01D-0739-09	34.86
TCGA-AB-2971-11A-01D-0739-09	31.64
TCGA-AB-2972-03A-01D-0739-09	42.22
TCGA-AB-2972-11A-01D-0739-09	31.87
TCGA-AB-2973-03A-01D-0739-09	32.80
TCGA-AB-2973-11A-01D-0739-09	36.96
TCGA-AB-2974-03A-01D-0739-09	31.79
TCGA-AB-2974-11A-01D-0739-09	26.99
TCGA-AB-2975-03A-01D-0739-09	26.90
TCGA-AB-2975-11A-01D-0739-09	32.00
TCGA-AB-2976-03A-01D-0739-09	25.51
TCGA-AB-2976-11A-01D-0739-09	26.67
TCGA-AB-2977-03A-01D-0739-09	28.22
TCGA-AB-2977-11A-01D-0739-09	30.17
TCGA-AB-2978-03A-01D-0739-09	34.03
TCGA-AB-2978-11A-01D-0739-09	31.35
TCGA-AB-2979-03A-01D-0739-09	34.26
TCGA-AB-2979-11A-01D-0739-09	26.20
TCGA-AB-2980-03A-01D-0739-09	34.14
TCGA-AB-2980-11A-01D-0739-09	28.33
TCGA-AB-2981-03A-01D-0739-09	25.13
TCGA-AB-2981-11A-01D-0739-09	40.36
TCGA-AB-2982-03A-01D-0739-09	28.84
TCGA-AB-2982-11A-01D-0739-09	24.53
TCGA-AB-2983-03A-01D-0739-09	26.28
TCGA-AB-2983-11A-01D-0739-09	41.04
TCGA-AB-2984-03A-01D-0739-09	21.71

TCGA-AB-2877-11A-01W-0732-08	225.92	TCGA-AB-2984-11A-01D-0739-09	18.94
TCGA-AB-2878-03A-01W-0732-08	194.99	TCGA-AB-2985-03A-01D-0739-09	37.69
TCGA-AB-2878-11A-01W-0732-08	62.91	TCGA-AB-2985-11A-01D-0739-09	28.31
TCGA-AB-2879-03A-01W-0732-08	186.85	TCGA-AB-2986-03A-01D-0739-09	34.62
TCGA-AB-2879-11A-01W-0732-08	196.42	TCGA-AB-2986-11A-01D-0739-09	28.72
TCGA-AB-2880-03A-01W-0732-08	180.38	TCGA-AB-2987-03A-01D-0739-09	26.01
TCGA-AB-2880-11A-01W-0732-08	210.41	TCGA-AB-2987-11A-01D-0739-09	30.63
TCGA-AB-2881-03A-01W-0732-08	170.96	TCGA-AB-2988-03A-01D-0739-09	34.86
TCGA-AB-2881-11A-01W-0732-08	185.32	TCGA-AB-2988-11A-01D-0739-09	47.29
TCGA-AB-2882-03A-01W-0761-09	157.75	TCGA-AB-2989-03A-01D-0739-09	29.08
TCGA-AB-2882-11A-01W-0732-08	184.84	TCGA-AB-2989-11A-01D-0739-09	32.68
TCGA-AB-2883-03A-01W-0732-08	172.53	TCGA-AB-2990-03A-01D-0739-09	29.10
TCGA-AB-2883-11A-01W-0732-08	194.23	TCGA-AB-2990-11A-01D-0739-09	27.81
TCGA-AB-2884-03A-01W-0732-08	105.80	TCGA-AB-2991-03A-01D-0739-09	27.25
TCGA-AB-2884-11A-01W-0732-08	145.96	TCGA-AB-2991-11A-01D-0739-09	24.84
TCGA-AB-2885-03A-01W-0732-08	125.73	TCGA-AB-2992-03A-01D-0739-09	31.74
TCGA-AB-2885-11A-01W-0732-08	159.68	TCGA-AB-2992-11A-01D-0739-09	30.82
TCGA-AB-2886-03A-01W-0732-08	164.75	TCGA-AB-2993-03A-01D-0739-09	26.38
TCGA-AB-2886-11A-01W-0732-08	188.17	TCGA-AB-2993-11A-01D-0739-09	24.34
TCGA-AB-2887-03A-01W-0732-08	156.32	TCGA-AB-2994-03A-01D-0739-09	30.62
TCGA-AB-2887-11A-01W-0732-08	189.84	TCGA-AB-2994-11A-01D-0739-09	26.36
TCGA-AB-2888-03A-01W-0732-08	172.71	TCGA-AB-2995-03A-01D-0739-09	31.51
TCGA-AB-2888-11A-01W-0732-08	199.67	TCGA-AB-2995-11A-01D-0739-09	28.09
TCGA-AB-2889-03A-01W-0732-08	163.90	TCGA-AB-2996-03A-01D-0739-09	36.17
TCGA-AB-2889-11A-01W-0761-09	164.51	TCGA-AB-2996-11A-01D-0739-09	26.39
TCGA-AB-2890-03A-01W-0732-08	127.98	TCGA-AB-2997-03A-01D-0739-09	26.17
TCGA-AB-2890-11A-01W-0732-08	144.93	TCGA-AB-2997-11A-01D-0739-09	26.58
TCGA-AB-2891-03A-01W-0733-08	140.01	TCGA-AB-2998-03A-01D-0739-09	30.21
TCGA-AB-2891-11A-01W-0732-08	148.66	TCGA-AB-2998-11A-01D-0739-09	28.59
TCGA-AB-2892-03A-01W-0733-08	162.45	TCGA-AB-2999-03A-01D-0739-09	25.47
TCGA-AB-2892-11A-01W-0732-08	154.12	TCGA-AB-2999-11A-01D-0739-09	29.51

TCGA-AB-2893-03A-01W-0733-08	101.60	TCGA-AB-3000-03A-01D-0739-09	26.34
TCGA-AB-2893-11A-01W-0732-08	165.02	TCGA-AB-3000-11A-01D-0739-09	29.87
TCGA-AB-2894-03A-01W-0733-08	166.68	TCGA-AB-3001-03A-01D-0739-09	26.83
TCGA-AB-2894-11A-01W-0732-08	174.47	TCGA-AB-3001-11A-01D-0739-09	28.33
TCGA-AB-2895-03A-01W-0733-08	151.35	TCGA-AB-3002-03A-01D-0739-09	27.39
TCGA-AB-2895-11A-01W-0732-08	171.12	TCGA-AB-3002-11A-01D-0739-09	30.16
TCGA-AB-2896-03A-01W-0733-08	146.60	TCGA-AB-3005-03A-01D-0739-09	36.13
TCGA-AB-2896-11A-01W-0732-08	176.46	TCGA-AB-3005-11A-01D-0739-09	32.87
TCGA-AB-2897-03A-01W-0733-08	157.86	TCGA-AB-3006-03A-01D-0739-09	31.14
TCGA-AB-2897-11A-01W-0732-08	172.35	TCGA-AB-3006-11A-01D-0739-09	33.49
TCGA-AB-2898-03A-01W-0733-08	196.42	TCGA-AB-3007-03A-01D-0739-09	26.46
TCGA-AB-2898-11A-01W-0732-08	196.84	TCGA-AB-3007-11A-01D-0739-09	26.92
TCGA-AB-2899-03A-01W-0733-08	184.91	TCGA-AB-3008-03A-01D-0739-09	39.05
TCGA-AB-2899-11A-01W-0732-08	182.22	TCGA-AB-3008-11A-01D-0739-09	28.01
TCGA-AB-2900-03A-01W-0733-08	185.84	TCGA-AB-3009-03A-01D-0739-09	32.95
TCGA-AB-2900-11A-01W-0732-08	193.20	TCGA-AB-3009-11A-01D-0739-09	31.63
TCGA-AB-2901-03A-01W-0733-08	183.65	TCGA-AB-3011-03A-01D-0739-09	50.40
TCGA-AB-2901-11A-01W-0732-08	188.39	TCGA-AB-3011-11A-01D-0739-09	30.12
TCGA-AB-2903-03A-01W-0761-09	160.61	TCGA-AB-3012-03A-01D-0739-09	25.76
TCGA-AB-2903-11A-01W-0732-08	137.26	TCGA-AB-3012-11A-01D-0739-09	33.38

Table S4: Summary of assays performed for each sample

See separate file
https://tcga-data.nci.nih.gov/docs/publications/aml_2012/SupplementalTable04.tsv

Table S5: Segments of copy number amplification and deletion

See separate file
https://tcga-data.nci.nih.gov/docs/publications/aml_2012/SupplementalTable05.tsv

Table S6: All somatic mutations with annotation and readcounts from DNA and RNA sequencing
See separate file
https://tcga-data.nci.nih.gov/docs/publications/aml_2012/SupplementalTable06.tsv

Table S7: Significantly Mutated Genes

Significantly mutated genes, as determined by the MuSiC package. Columns 7-9 give p-values for a Fisher's Combined P-value Test (FCPT) a Likelihood Ratio Test (LRT) and a Convolution Test (CT). The subsequent 3 columns give FDR values for the same, after multiple-testing correction.

Gene	Indels	SNVs	Tot Muts	Covd Bps	Muts pMbp	P-value FCPT	P-value LRT	P-value CT	FDR FCPT	FDR LRT	FDR CT
CEBPA	16	3	19	41306	459.98	0	0	0	0	0	0
DNMT3A	5	52	57	629565	90.54	0	0	0	0	0	0
FLT3	37	17	54	657876	82.08	0	0	0	0	0	0
IDH1	0	19	19	253738	74.88	0	0	0	0	0	0
IDH2	0	20	20	243278	82.21	0	0	0	0	0	0
NPM1	54	1	55	183897	299.08	0	0	0	0	0	0
NRAS	0	15	15	116904	128.31	0	0	0	0	0	0
RUNX1	6	15	21	283105	74.18	0	0	0	0	0	0
TET2	16	12	29	1110805	26.11	0	0	0	0	0	0
TP53	4	14	18	264994	67.93	0	0	0	0	0	0
WT1	9	4	13	233938	55.57	0	0	0	0	0	0
KRAS	0	8	8	138617	57.71	2.6068E-13	1.11022E-16	7.35648E-19	8.11911E-10	3.19189E-13	2.29124E-15
U2AF1	0	8	8	164572	48.61	4.35985E-13	0	1.59177E-18	1.16392E-09	0	4.57634E-15
KIT	3	7	10	635486	15.74	3.61156E-13	2.22045E-16	2.53343E-18	1.03832E-09	5.9278E-13	6.76335E-15
PTPN11	0	9	9	456145	19.73	2.14748E-11	2.27596E-14	2.21698E-16	5.35081E-08	5.67093E-11	5.52397E-13
PHF6	2	4	6	326403	18.38	1.09882E-07	2.19761E-10	3.07091E-12	0.000256678	5.13349E-07	7.17346E-09

SMC3	0	7	7	745533	9.39	6.28548E-07	9.74997E-08	1.05739E-10	0.001381882	0.000173526	2.3247E-07
FAM5C	0	6	6	462924	12.96	9.49337E-06	4.52893E-09	2.76489E-09	0.018674464	9.95698E-06	5.74099E-06
SMC1A	0	7	7	852590	8.21	8.59176E-06	4.75931E-08	3.92292E-09	0.017839838	8.89397E-05	7.71679E-06
RAD21	3	2	5	387435	12.91	2.55389E-05	1.42727E-08	7.76104E-09	0.047725814	2.96357E-05	1.45034E-05
STAG2	0	6	6	816412	7.35	4.38476E-05	3.69498E-08	3.24599E-08	0.078038344	7.26842E-05	5.77709E-05
HNRNPK	3	0	3	337099	8.9	0.009064315	2.4535E-07	5.84965E-06	1	0.000398694	0.009505687
EZH2	2	2	4	496328	8.06	0.005591501	1.92418E-05	8.95329E-06	1	0.024085704	0.013942879

Table S8: Recurrent mutations in non-genic regions

Chr	St	Sp	Ref	Var	Sample	Gene	Distance To Gene	ENCODE_ChromHMM	Conservation_score
1	80743586	80743586	C	A	AML12	ENSG00000217175	9080	13_Heterochrom/lo	0.000
1	80743777	80743777	C	T	AML43	ENSG00000217175	9271	13_Heterochrom/lo	0.000
1	175674904	175674904	T	A	AML43	LOC400796	0	12_Repressed	0.017
1	175674914	175674914	A	C	AML05	LOC400796	0	12_Repressed	0.005
1	199603019	199603019	C	T	AML40	TNNT2	0	13_Heterochrom/lo	0.001
1	199603188	199603188	C	T	AML18	TNNT2	0	13_Heterochrom/lo	0.000
1	207560587	207560587	G	A	AML25	ENSG00000219440	52396	13_Heterochrom/lo	-
1	207560745	207560745	A	C	AML20	ENSG00000219440	52554	13_Heterochrom/lo	-
1	216303921	216303921	T	C	AML28	ENSG00000201493	65838	13_Heterochrom/lo	-
1	216304035	216304035	G	A	AML43	ENSG00000201493	65724	13_Heterochrom/lo	-
1	240678783	240678783	G	A	AML28	PLD5	0	13_Heterochrom/lo	0.809
1	240678902	240678902	G	A	AML45	PLD5	0	13_Heterochrom/lo	0.151
2	2787616	2787616	C	T	AML08	LOC389024	235461	13_Heterochrom/lo	-
2	2787616	2787616	C	T	AML44	LOC389024	235461	13_Heterochrom/lo	-
2	23136138	23136138	G	T	AML10	ENSG00000222616	197893	13_Heterochrom/lo	-
2	23136226	23136226	C	T	AML36	ENSG00000222616	197981	13_Heterochrom/lo	-
2	23315322	23315322	C	T	AML48	KLHL29	143880	13_Heterochrom/lo	-
2	23315413	23315413	C	T	AML10	KLHL29	143789	13_Heterochrom/lo	-
2	143217115	143217115	T	C	AML29	KYNU	134421	13_Heterochrom/lo	-
2	143217232	143217232	A	C	AML50	KYNU	134304	13_Heterochrom/lo	-
2	158403247	158403247	A	C	AML44	ACVR1	0	13_Heterochrom/lo	0.000
2	158403328	158403328	C	A	AML12	ACVR1	0	13_Heterochrom/lo	0.996
2	215677037	215677037	C	T	AML46	ABCA12	0	11_Weak_Txn	0.032
2	215677160	215677160	C	T	AML48	ABCA12	0	11_Weak_Txn	0.124
3	21310539	21310539	C	A	AML20	VENTXP7	111682	13_Heterochrom/lo	-

3	21310708	21310708	C	A	AML31	VENTXP7	111513	13_Heterochrom/lo	-
3	27043413	27043413	G	A	AML03	NEK10	83985	13_Heterochrom/lo	-
3	27043554	27043554	T	C	AML11	NEK10	83844	13_Heterochrom/lo	-
3	53488484	53488484	G	T	AML06	CACNA1D	15631	12_Repressed	0.000
3	53488649	53488649	C	G	AML12	CACNA1D	15466	12_Repressed	0.000
4	14719735	14719735	A	-	AML21	LOC100129903	0	13_Heterochrom/lo	0.000
4	14719735	14719735	A	G	AML29	LOC100129903	0	13_Heterochrom/lo	0.000
4	18022446	18022446	A	G	AML12	ENSG00000209956	163397	1_Active_Promoter	-
4	18022627	18022627	A	T	AML07	ENSG00000209956	163216	2_Weak_Promoter	-
4	23984368	23984368	G	A	AML51	ENSG00000222262	47528	13_Heterochrom/lo	0.001
4	23984451	23984451	G	T	AML03	ENSG00000222262	47445	13_Heterochrom/lo	0.022
4	31779477	31779477	G	C	AML28	LOC100130644	0	13_Heterochrom/lo	0.004
4	31779598	31779598	G	A	AML34	LOC100130644	0	13_Heterochrom/lo	0.002
4	76113392	76113392	G	A	AML09	DKFZP564O0823,uc003hih.1	0	13_Heterochrom/lo	0.000
4	76113412	76113412	C	A	AML28	DKFZP564O0823,uc003hih.1	0	13_Heterochrom/lo	0.034
4	99892327	99892327	G	T	AML44	ENSG00000174991	10855	13_Heterochrom/lo	0.012
4	99892405	99892405	T	C	AML29	ENSG00000174991	10933	13_Heterochrom/lo	0.000
4	144623658	144623658	A	G	AML40	GAB1	12929	12_Repressed	0.018
4	144623720	144623720	C	A	AML08	GAB1	12991	12_Repressed	0.025
5	29385295	29385295	T	-	AML06	LOC100130803	251175	13_Heterochrom/lo	-
5	29385298	29385298	T	C	AML49	LOC100130803	251172	13_Heterochrom/lo	-
5	45550440	45550440	G	A	AML02	HCN1	0	13_Heterochrom/lo	0.003
5	45550574	45550574	C	-	AML45	HCN1	0	13_Heterochrom/lo	0.000
5	116635650	116635650	C	T	AML23	LOC728342	143453	13_Heterochrom/lo	-
5	116635767	116635767	C	T	AML51	LOC728342	143336	13_Heterochrom/lo	-
5	140320477	140320477	T	C	AML21	PCDHA1, PCDHA10, PCDHA11, PCDHA12, PCDHA13, PCDHA2, PCDHA3, PCDHA4, PCDHA5, PCDHA6, PCDHA7, PCDHA8, PCDHA9, PCDHAC1	0	13_Heterochrom/lo	0.814

5	140320536	140320536	G	A	AML06	PCDHA1, PCDHA10, PCDHA11, PCDHA12, PCDHA13, PCDHA2, PCDHA3, PCDHA4, PCDHA5, PCDHA6, PCDHA7, PCDHA8, PCDHA9, PCDHAC1	0	13_Heterochrom/lo	0.971
5	156250002	156250002	G	A	AML33	TIMD4	28941	13_Heterochrom/lo	0.000
5	156250002	156250002	G	A	AML43	TIMD4	28941	13_Heterochrom/lo	0.000
6	318610	318610	G	A	AML38	LOC730077	0	14_Repetitive/CNV	0.000
6	318785	318785	G	A	AML39	LOC730077	0	14_Repetitive/CNV	0.000
6	18875178	18875178	G	T	AML18	MIRN548A1	195088	13_Heterochrom/lo	-
6	18875323	18875323	A	-	AML02	MIRN548A1	195233	13_Heterochrom/lo	-
6	22698663	22698663	A	G	AML23	HDGFL1	19934	13_Heterochrom/lo	0.007
6	22698699	22698699	T	A	AML06	HDGFL1	19970	13_Heterochrom/lo	0.874
6	95328747	95328747	T	C	AML45	ENSG00000209311	114979	13_Heterochrom/lo	-
6	95328814	95328814	A	G	AML49	ENSG00000209311	115046	13_Heterochrom/lo	-
7	9440110	9440110	T	G	AML48	ENSG00000218135	181022	13_Heterochrom/lo	-
7	9440306	9440306	T	A	AML51	ENSG00000218135	180826	13_Heterochrom/lo	-
7	70789706	70789706	C	G	AML44	WBSCR17	0	13_Heterochrom/lo	0.003
7	70789795	70789795	C	G	AML01	WBSCR17	0	13_Heterochrom/lo	0.000
7	136384858	136384858	C	T	AML06	LOC100128744	24142	13_Heterochrom/lo	0.002
7	136384880	136384880	C	T	AML30	LOC100128744	24120	13_Heterochrom/lo	0.058
7	156474488	156474488	G	A	AML37	MNX1	15819	10_Txn_Elongation	0.001
7	156474493	156474493	C	T	AML03	MNX1	15814	10_Txn_Elongation	0.000
8	36074299	36074299	G	T	AML39	ENSG00000210631	180371	13_Heterochrom/lo	-
8	36074393	36074393	G	A	AML10	ENSG00000210631	180277	13_Heterochrom/lo	-
9	13363849	13363849	G	A	AML49	FLJ41200	32529	13_Heterochrom/lo	0.272
9	13363866	13363866	G	C	AML11	FLJ41200	32512	13_Heterochrom/lo	0.002
10	31544255	31544255	C	T	AML02	ENSG00000209675	44278	13_Heterochrom/lo	0.007
10	31544410	31544410	C	T	AML07	ENSG00000209675	44123	13_Heterochrom/lo	0.002

10	56803960	56803960	A	T	AML20	PCDH15	0	13_Heterochrom/lo	0.009
10	56804084	56804084	G	A	AML04	PCDH15	0	13_Heterochrom/lo	0.000
10	84453341	84453341	G	A	AML35	NRG3	0	13_Heterochrom/lo	0.000
10	84453471	84453471	G	T	AML43	NRG3	0	13_Heterochrom/lo	0.000
10	127873288	127873288	A	G	AML30	ADAM12	0	13_Heterochrom/lo	0.000
10	127873311	127873311	C	T	AML19	ADAM12	0	13_Heterochrom/lo	0.000
11	117101176	117101176	C	T	AML37	DSCAML1	0	9_Txn_Transition	0.000
11	117101343	117101343	A	G	AML01	DSCAML1	0	5_Strong_Enhancer	0.000
11	117869938	117869938	T	C	AML06	MLL	0	11_Weak_Txn	0.005
11	117870121	117870121	G	C	AML17	MLL	0	11_Weak_Txn	0.003
11	131710591	131710591	C	A	AML49	HNT,NTM	0	13_Heterochrom/lo	0.000
11	131710683	131710683	G	A	AML39	HNT,NTM	0	13_Heterochrom/lo	0.000
12	5326141	5326141	C	T	AML39	LOC100133418	0	13_Heterochrom/lo	0.002
12	5326326	5326326	G	A	AML16	LOC100133418	0	13_Heterochrom/lo	0.000
12	14089388	14089388	C	T	AML23	GRIN2B	65069	13_Heterochrom/lo	-
12	14089532	14089532	G	A	AML29	GRIN2B	65213	13_Heterochrom/lo	-
12	96997823	96997823	G	A	AML06	LOC732096	323397	13_Heterochrom/lo	-
12	96997914	96997914	A	-	AML33	LOC732096	323488	13_Heterochrom/lo	-
12	125520708	125520708	C	A	AML40	ENSG00000214043	0	13_Heterochrom/lo	0.014
12	125520889	125520889	G	A	AML29	ENSG00000214043	0	13_Heterochrom/lo	0.028
13	18518150	18518150	C	T	AML44	ENSG00000218236	0	.	0.000
13	18518170	18518170	G	A	AML33	ENSG00000218236	0	.	0.000
13	55550363	55550363	T	-	AML06	HNF4GP1	77780	13_Heterochrom/lo	-
13	55550461	55550462	-	T	AML08	HNF4GP1	77879	13_Heterochrom/lo	-
14	53624173	53624173	G	A	AML37	ENSG00000219778	95710	13_Heterochrom/lo	-
14	53624180	53624180	C	T	AML50	ENSG00000219778	95717	13_Heterochrom/lo	-
14	78991968	78991968	T	A	AML23	NRXN3	0	13_Heterochrom/lo	0.020
14	78992069	78992069	C	T	AML30	NRXN3	0	13_Heterochrom/lo	0.011
14	81806925	81806925	C	T	AML17	ENSG00000210370	159042	11_Weak_Txn	-
14	81807121	81807121	G	A	AML38	ENSG00000210370	158846	11_Weak_Txn	-
15	24601542	24601543	-	T	AML43	GABRB3	0	13_Heterochrom/lo	0.000

15	24601667	24601667	G	A	AML20	GABRB3	0	13_Heterochrom/lo	0.000
15	25047599	25047599	G	A	AML34	GABRG3	0	13_Heterochrom/lo	0.000
15	25047792	25047792	C	T	AML27	GABRG3	0	13_Heterochrom/lo	0.002
15	51192393	51192393	G	A	AML46	LOC645693	174345	13_Heterochrom/lo	-
15	51192438	51192438	G	A	AML15	LOC645693	174390	13_Heterochrom/lo	-
15	69244719	69244719	G	A	AML51	THSD4	0	5_Strong_Enhancer	1.000
15	69244831	69244831	G	A	AML50	THSD4	0	7_Weak_Enhancer	0.023
16	6129410	6129410	G	A	AML02	A2BP1	0	13_Heterochrom/lo	0.000
16	6129458	6129458	C	T	AML30	A2BP1	0	13_Heterochrom/lo	0.000
16	46232164	46232164	G	C	AML15	PHKB	0	.	0.587
16	46232202	46232202	C	T	AML49	PHKB	0	.	0.679
17	43672036	43672036	T	A	AML03	SKAP1	0	13_Heterochrom/lo	0.734
17	43672126	43672126	C	G	AML04	SKAP1	0	13_Heterochrom/lo	0.004
17	53256697	53256697	C	T	AML20	MRPS23	15143	13_Heterochrom/lo	0.000
17	53256723	53256723	C	T	AML30	MRPS23	15117	13_Heterochrom/lo	0.005
18	13606615	13606615	G	A	AML03	C18orf1	0	13_Heterochrom/lo	0.001
18	13606633	13606633	C	T	AML31	C18orf1	0	13_Heterochrom/lo	0.004
19	34569692	34569692	G	A	AML10	LOC284395	25868	13_Heterochrom/lo	0.000
19	34569752	34569752	C	A	AML03	LOC284395	25808	13_Heterochrom/lo	0.000
20	15371459	15371459	G	A	AML23	C20orf133,MACROD2	0	13_Heterochrom/lo	0.000
20	15371644	15371644	C	A	AML14	C20orf133,MACROD2	0	13_Heterochrom/lo	0.000
20	43250606	43250606	C	T	AML36	PI3	12007	13_Heterochrom/lo	0.103
20	43250759	43250760	-	C	AML32	PI3	12161	13_Heterochrom/lo	0.000
21	24117139	24117139	C	A	AML43	ENSG00000199698	390666	13_Heterochrom/lo	-
21	24117145	24117145	T	C	AML40	ENSG00000199698	390672	13_Heterochrom/lo	-
21	43545785	43545785	G	A	AML34	C21orf136	27220	12_Repressed	0.000
21	43545886	43545886	C	T	AML33	C21orf136	27119	12_Repressed	0.009
22	33530064	33530064	T	G	AML40	ENSG00000220899	212794	13_Heterochrom/lo	-
22	33530100	33530100	C	T	AML44	ENSG00000220899	212830	13_Heterochrom/lo	-
22	38398117	38398117	C	T	AML51	CACNA1I	0	13_Heterochrom/lo	0.000
22	38398294	38398294	G	A	AML29	CACNA1I	0	13_Heterochrom/lo	0.001

Table S9: Mitochondrial mutations

TCGA_id	WGS_id	chromosome_name	start	stop	reference	variant	type
TCGA-AB-2802		MT	13059	13059	C	-	DEL
TCGA-AB-2802		MT	14767	14767	T	C	SNP
TCGA-AB-2803		MT	7207	7207	G	A	SNP
TCGA-AB-2811		MT	2779	2779	G	A	SNP
TCGA-AB-2814		MT	7962	7962	T	C	SNP
TCGA-AB-2817		MT	10493	10493	T	C	SNP
TCGA-AB-2817		MT	1311	1311	A	G	SNP
TCGA-AB-2817		MT	13116	13116	T	C	SNP
TCGA-AB-2817		MT	3434	3434	T	C	SNP
TCGA-AB-2819		MT	11810	11810	T	C	SNP
TCGA-AB-2819		MT	530	530	T	C	SNP
TCGA-AB-2820		MT	8103	8103	G	A	SNP
TCGA-AB-2845		MT	1149	1149	G	A	SNP
TCGA-AB-2859		MT	14767	14767	T	C	SNP
TCGA-AB-2885		MT	8156	8156	G	A	SNP
TCGA-AB-2886		MT	8066	8066	G	A	SNP
TCGA-AB-2897		MT	3919	3919	G	A	SNP
TCGA-AB-2900		MT	14964	14964	G	A	SNP
TCGA-AB-2903		MT	7782	7782	A	G	SNP
TCGA-AB-2904		MT	15014	15014	A	C	SNP
TCGA-AB-2908		MT	1872	1872	A	G	SNP
TCGA-AB-2908		MT	6721	6721	A	G	SNP
TCGA-AB-2916		MT	14767	14767	T	C	SNP
TCGA-AB-2919		MT	12256	12256	T	G	SNP
TCGA-AB-2921		MT	651	651	A	G	SNP
TCGA-AB-2927		MT	13370	13370	T	C	SNP
TCGA-AB-2927		MT	13763	13763	T	C	SNP

TCGA-AB-2927		MT	9731	9731	T	C	SNP
TCGA-AB-2931		MT	913	913	T	C	SNP
TCGA-AB-2934		MT	15651	15651	G	A	SNP
TCGA-AB-2935		MT	7622	7622	T	C	SNP
TCGA-AB-2936		MT	9554	9554	G	A	SNP
TCGA-AB-2937		MT	11462	11462	A	G	SNP
TCGA-AB-2939		MT	7655	7655	T	C	SNP
TCGA-AB-2939		MT	9968	9968	T	C	SNP
TCGA-AB-2941		MT	13692	13692	A	G	SNP
TCGA-AB-2943		MT	4598	4598	T	C	SNP
TCGA-AB-2948		MT	8706	8706	T	C	SNP
TCGA-AB-2955		MT	7920	7920	G	A	SNP
TCGA-AB-3009	AML43	MT	3919	3919	G	A	SNP

Table S10: miRNA mutations

Sample	Chr	Start	Stop	Ref	Var	Type	miRNA
TCGA-AB-2805	17	53763621	53763621	C	G	SNP	MIRN142
TCGA-AB-2807	17	53763624	53763624	T	C	SNP	MIRN142
TCGA-AB-2859	19	58931994	58931995	TT	-	DEL	MIRN516B1
TCGA-AB-2926	17	53763622	53763622	A	G	SNP	MIRN142
TCGA-AB-2972	13	106981565	106981565	G	T	SNP	MIRN1267
TCGA-AB-2977	X	144917082	144917082	G	T	SNP	MIRN891A
TCGA-AB-3002	17	27701282	27701282	C	A	SNP	MIRN632
TCGA-AB-3002	17	53763621	53763621	C	G	SNP	MIRN142
TCGA-AB-3002	17	53763624	53763624	T	C	SNP	MIRN142

Table S11: Germline Variants

See separate file

https://tcga-data.nci.nih.gov/docs/publications/aml_2012/SupplementalTable11.xls

Supplementary Table 12: Mutual Exclusivity and Co-Occurence

a) All pairs of genes, gene groups, and cytogenetic risk categories that are exclusive with $P < 0.05$ using Fisher's exact test. A description of how Fisher's exact test was performed is available in the supplementary methods. Because pairs of cytogenetic categories are by definition fully exclusive, we do not include them in this table.

Gene / Gene Group 1	Gene / Gene Group 2	Fisher's p-value
Favorable	NPM1	2.07E-06
DNMT3A	Favorable	4.90E-06
Intermediate	PML-RARA	1.27E-05
Intermediate	TP53	3.11E-05
Intermediate	MYH11-CBFB	5.50E-05
NPM1	Unfavorable	3.26E-04
NPM1	RUNX1	1.26E-03
Intermediate	RUNX1-RUNX1T1	2.16E-03
NPM1	PML-RARA	5.13E-03
FLT3	TP53	5.83E-03
DNMT3A	PML-RARA	7.24E-03
NPM1	TP53	7.25E-03
FLT3	RUNX1	9.35E-03
Intermediate	KIT	1.12E-02
Intermediate	Unknown	1.29E-02
Favorable	IDH2	1.32E-02
Favorable	RUNX1	1.32E-02
Favorable	IDH1	1.67E-02
PML-RARA	Unfavorable	1.75E-02
MYH11-CBFB	NPM1	2.82E-02
DNMT3A	MYH11-CBFB	3.56E-02
FLT3	Unfavorable	3.73E-02
Favorable	TP53	4.10E-02
RUNX1	Unfavorable	4.31E-02

FLT3	IDH2	4.42E-02
ABL1,DYRK4,EPHA2,EPHA3,JAK3,MST1R,OBSCN,PDGFRB,WEE1	FLT3	4.84E-02

b) All pairs of genes, gene groups, and cytogenetic risk categories that are co-occurring with *P* < 0.05 using Fisher's exact test. A description of how Fisher's exact test was performed is available in the supplement. Because pairs of cytogenetic categories are by definition fully exclusive, we do not include them in this table.

Gene / Gene Group 1	Gene / Gene Group 2	Fisher's p-value
Favorable	PML-RARA	9.10E-12
ADAM3A	uc003xne.1	1.21E-11
ENSG00000219705	PRMT2	1.21E-11
Intermediate	NPM1	1.08E-09
Favorable	MYH11-CBFB	2.20E-09
ADAM3A	ADAM5P	1.41E-08
ADAM5P	uc003xne.1	1.41E-08
C15orf23	DISP2	1.55E-08
TP53	Unfavorable	3.15E-08
ENSG00000219705	S100B	2.32E-07
PRMT2	S100B	2.32E-07
DNMT3A	NPM1	6.28E-07
FLT3	NPM1	2.00E-06
Favorable	RUNX1-RUNX1T1	4.51E-06
C15orf23	LOC100128885	1.21E-05
DISP2	LOC100128885	1.21E-05
DNMT3A	Intermediate	1.78E-05
ENSG00000219705	TP53	2.89E-04
PRMT2	TP53	2.89E-04
ASXL1	IDH2	3.50E-04

ASXL1	RUNX1	3.50E-04
NF1	SUZ12	4.51E-04
BCR-ABL1	MT-ND5	8.98E-04
LOC728807	UGT2B17	8.98E-04
IDH2	RUNX1	1.17E-03
Intermediate	RUNX1	1.19E-03
DNMT3A	SMC3	1.23E-03
S100B	TP53	1.32E-03
DNMT3A	IDH1	1.62E-03
IDH2	MIR142	3.25E-03
Intermediate	MLL-PTD	5.98E-03
ADAM5P	TP53	6.02E-03
KIT	MYH11-CBFB	6.02E-03
KDM6A	RUNX1-RUNX1T1	6.12E-03
Favorable	KIT	6.19E-03
BCR-ABL1	Unfavorable	9.40E-03
MLL-ELL	Unfavorable	9.40E-03
NPM1	PTPN11	1.29E-02
DNMT3A	FLT3	1.36E-02
PHF6	RUNX1	1.42E-02
STAG2	TET2	1.46E-02
DNMT3A	RPS6KA6	1.59E-02
DNAH9	TET2	1.95E-02
ENSG00000219705	Unfavorable	2.02E-02
PRMT2	Unfavorable	2.02E-02
GATA2,CBFB,ETV6,ETV3,GLI1,IKZF1,MYB,MYC,MLLT10-CEP164	LOC100128885	2.18E-02
C15orf23	GATA2,CBFB,ETV6,ETV3,GLI1,IKZF1,MYB,MYC,MLLT10-	2.18E-02
DISP2	GATA2,CBFB,ETV6,ETV3,GLI1,IKZF1,MYB,MYC,MLLT10-	2.18E-02
IDH1	RPS6KA6	2.43E-02
LOC100130472	MYH11-CBFB	2.52E-02
IDH2	Intermediate	2.53E-02

KIT	RUNX1-RUNX1T1	2.67E-02
DIS3	RUNX1	2.69E-02
MT-CO3	RUNX1	2.69E-02
IDH2	SMG1	2.69E-02
RUNX1	SUZ12	2.69E-02
S100B	Unfavorable	3.19E-02
ABL1,DYRK4,EPHA2,EPHA3,JAK3,MST1R,OBSCN,PDGFRB,WEE1	RUNX1-RUNX1T1	3.37E-02
PTPN11	STAG2	3.37E-02
IDH2	KRAS	3.48E-02
RUNX1	U2AF1	3.48E-02
IDH1	NPM1	3.79E-02
DNMT3A	MT-CYB	3.79E-02
PHF6	WT1	4.34E-02
BCR-ABL1	EZH2	4.45E-02
C8orf33	NF1	4.45E-02
C8orf33	SUZ12	4.45E-02
CACNA1B	HLA-DRB1	4.45E-02
CACNA1B	LOC728807	4.45E-02
CACNA1E	CROCC	4.45E-02
CACNA1E	LOC728807	4.45E-02
CROCC	PTPRT	4.45E-02
DIS3	DNAH9	4.45E-02
DIS3	EZH2	4.45E-02
DIS3	MT-CO3	4.45E-02
FCGBP	GPR128-TFG	4.45E-02
FCGBP	SMG1	4.45E-02
GPR128-TFG	NUP98-NSD1	4.45E-02
GRIK2	LOC732248	4.45E-02
GRIK2	RPS6KA6	4.45E-02
GRIK2	SPEN	4.45E-02
LOC732248	RPS6KA6	4.45E-02

MT-CO2	NF1	4.45E-02
MT-CO2	SUZ12	4.45E-02
PHACTR1	SPEN	4.45E-02
FLT3	Intermediate	4.48E-02
LOC100130472	NRAS	4.61E-02
IDH2	MLL-PTD	4.88E-02
LOC100130472	NRAS	4.61E-02

Table S13: Gene Fusions

See separate file

https://tcga-data.nci.nih.gov/docs/publications/aml_2012/SupplementalTable13.tsv

Table S14: Univariate analyses of DNA methylation changes (*P* < 0.05, FDR < 0.05 overall)

	Hypermethylated vs. CD34+ cells	Hypomethylated vs. CD34+ cells
PML-RARA fusion (n=15)	9171	8314
MYH11-CBFB fusion (n=11)	25537	11855
RUNX1-RUNX1T1 fusion (n=7)	4679	3125
MLL fusions (all partners, n=11)	7180	4319
NUP98-NSD1 fusion (n=3)	716	51
FLT3 mutation (n=55)	10669	5474
NPM1 mutation (n=53)	15087	6144
DNMT3A mutation (n=48)	11707	5520
NPM1/FLT3/DNMT3A (n=17)	19934	14746
TP53 mutation (n=15)	3484	2104
IDH1 mutation (n=19)	14451	2583
IDH2 mutation (n=18)	11871	2549
TET2 mutation (n=16)	5931	2107
WT1 mutation (n=11)	11395	1998
CEBPA mutation (n=13)	8543	1216
RUNX1 mutation (n=16)	7634	2139
ASXL1 mutation (n=4)	0	0
KDM6A mutation (n=4)	0	0
Cohesin mutations (n=24)	10380	3657
Spliceosome mutations (n=22)	9503	2541
	Hypermethylated vs. all normals	**Hypomethylated vs. all normals**
PML-RARA fusion (n=15)	88650	52844
MYH11-CBFB fusion (n=11)	126070	63799
RUNX1-RUNX1T1 fusion (n=7)	96867	51332
MLL fusions (all partners, n=11)	67794	59847
NUP98-NSD1 fusion (n=3)	82897	21427
FLT3 mutation (n=55)	76678	54755
NPM1 mutation (n=53)	79321	60692
DNMT3A mutation (n=48)	61457	54002
NPM1/FLT3/DNMT3A (n=17)	77877	83031
TP53 mutation (n=15)	65381	64249

IDH1 mutation (n=19)	115089	36256
IDH2 mutation (n=18)	119620	35519
TET2 mutation (n=16)	69426	44652
WT1 mutation (n=11)	106840	31397
CEBPA mutation (n=13)	105028	27038
RUNX1 mutation (n=16)	85092	38235
ASXL1 mutation (n=4)	79913	21350
KDM6A mutation (n=4)	46560	20838
Cohesin mutations (n=24)	78554	52895
Spliceosome mutations (n=22)	79055	42305

Table S15: Multivariate analysis of DNA methylation changes (*P* < 0.05, FDR < 0.05)

	Hypermethylated loci (fraction)	Hypomethylated loci (fraction)
AML patients (n=192) vs. donors	107336 (27.9%)	53183 (13.8%)
PML-RARA fusion (n=15)	10085 (2.6%)	14673 (3.8%)
MYH11-CBFB fusion (n=11)	7134 (1.9%)	25620 (6.7%)
RUNX1-RUNX1T1 fusion (n=7)	2226 (0.6%)	12435 (3.2%)
MLL fusions (all partners, n=11)	1745 (0.5%)	30212 (7.9%)
NUP98-NSD1 fusion (n=3)	428 (0.1%)	127 (<0.1%)
FLT3 mutation (n=55)	0 (0%)	0 (0%)
NPM1 mutation (n=53)	4568 (1.2%)	35617 (9.3%)
DNMT3A mutation (n=48)	1442 (0.4%)	39775 (10.4%)
…with FLT3 mutation (n=21)	333 (0.1%)	1506 (0.4%)
…and NPM1 mutation (n=17)	881 (0.2%)	190 (<0.1%)
TP53 mutation (n=15)	3561 (0.9%)	45764 (11.9%)
IDH1 mutation (n=19)	78798 (20.5%)	4119 (1.1%)
IDH2 mutation (n=18)	72600 (18.9%)	540 (0.1%)
TET2 mutation (n=16)	1534 (0.4%)	579 (0.2%)
WT1 mutation (n=11)	11131 (2.9%)	360 (0.1%)
CEBPA mutation (n=13)	16443 (4.3%)	1617 (0.4%)
RUNX1 mutation (n=16)	104 (<0.1%)	132 (<0.1%)
ASXL1 mutation (n=4)	31 (<0.1%)	12 (<0.1%)
KDM6A mutation (n=4)	116 (<0.1%)	648 (0.2%)
Cohesin mutations (n=24)	52 (<0.1%)	19 (<0.1%)
Spliceosome mutations (n=22)	0 (0%)	0 (0%)

Table S16: Univariate analyses of differentially methylated regions (p < 0.05, 1000nt+)

	Hypermethylated vs. CD34+ cells	**Hypomethylated vs. CD34+ cells**
PML-RARA fusion (n=15)	261	152
MYH11-CBFB fusion (n=11)	198	286
RUNX1-RUNX1T1 fusion (n=7)	223	230
MLL fusions (all partners, n=11)	60	60
NUP98-NSD1 fusion (n=3)	205	39
FLT3 mutation (n=55)	47	8
NPM1 mutation (n=53)	66	27
DNMT3A mutation (n=48)	12	1
NPM1/FLT3/DNMT3A (n=17)	54	328
TP53 mutation (n=15)	40	3
IDH1 mutation (n=19)	55	1
IDH2 mutation (n=18)	90	0
TET2 mutation (n=16)	15	0
WT1 mutation (n=11)	237	5
CEBPA mutation (n=13)	242	1
RUNX1 mutation (n=16)	55	1
ASXL1 mutation (n=4)	24	0
KDM6A mutation (n=4)	13	2
Cohesin mutations (n=24)	50	10
Spliceosome mutations (n=22)	61	0
	Hypermethylated vs. all normals	Hypomethylated vs. all normals
PML-RARA fusion (n=15)	1294	315
MYH11-CBFB fusion (n=11)	864	246
RUNX1-RUNX1T1 fusion (n=7)	1694	717
MLL fusions (all partners, n=11)	751	497
NUP98-NSD1 fusion (n=3)	1804	112
FLT3 mutation (n=55)	921	195
NPM1 mutation (n=53)	911	346
DNMT3A mutation (n=48)	503	205
NPM1/FLT3/DNMT3A (n=17)	326	861

TP53 mutation (n=15)	1057	85
IDH1 mutation (n=19)	1414	23
IDH2 mutation (n=18)	2737	27
TET2 mutation (n=16)	635	26
WT1 mutation (n=11)	2603	103
CEBPA mutation (n=13)	2672	60
RUNX1 mutation (n=16)	1851	39
ASXL1 mutation (n=4)	1830	49
KDM6A mutation (n=4)	1366	547
Cohesin mutations (n=24)	1143	288
Spliceosome mutations (n=22)	1547	74

Table S17: The 4 subnetworks identified by HotNet.
For each gene, the number in parentheses indicates the number of patients with an alteration in the gene. For pathway/protein complex enrichments, the significance of the overlap between genes in the subnetwork and those in the pathway/protein complex is reported (*p*-values are from a hypergeometric test with Bonferroni correction for multiple hypotheses).

SUBNETWORK	KEGG PATHWAYS ENRICHMENTS		PROTEIN COMPLEXES (PINdb) ENRICHMENTS	
	Name	p-value	Name	p-value
SMC1A(7), CYLD(1), WAPAL(1), RAD21(5), STAG2(6), SMC3(7), PDS5B(1)	Cell cycle	10^{-3}	cohesin-2 SNF2h/cohesin cohesin-1	$9x10^{-7}$ $6x10^{-5}$ $2x10^{-4}$
DNMT3B(2), EZH2(3), DNMT3A(51), EED(2), MYC(1), DNMT1 (1)	Cysteine and methionine metabolism	$9x10^{-4}$		
RARA(16), TCF4(1), HDAC3(1), SUMO2(1), THRB(1), RUNX1(26), RUNX1T1(8), PML(16),	Acute myeloid leukemia	10^{-3}		
MLL3(2), E2F6(1), KDM6A(4), LEO1(1), MLL2(1), CDC73(1), MLL(20)			PTIP HMT hPAF	$2x10^{-4}$ $4x10^{-2}$

Table S18. List of genes in each of the manually curated functional gene groups.

Functional gene group	Genes in the group
Spliceosome	CSTF2T, DDX1, DDX23, DHX32, HNRNPK, METTL3, PLRG1, PRPF3, PRPF8, RBMX, SF3B1, SNRNP200, SRRM2, SRSF6, SUPT5H, TRA2B, U2AF1, U2AF1L4, U2AF2
Cohesin complex	SMC1A, SMC3, SMC5, STAG2, RAD21
MLL-X fusions	MLL-ELL, MLL-MLLT4, MLL-MLLT3, MLLT10-MLL
Other myeloid transcription factors	GATA2, CBFB, ETV6, ETV3, GLI1, IKZF1, MYB, MYC, MLLT10-CEP164
Other epigenetic modifiers	ARID4B, ASXL2, ASXL3, BRPF1, CBX5, CBX7, EED, HDAC2, HDAC3, JMJD1C, KAT6B, KDM2B, KDM3B, MLL2, MLL3, MTA2, PRDM9, PRDM16, RBBP4, SAP130, SCML2, SUDS3, SUZ12, ZBTB33, ZBTB7B, CREBBP-KAT6A, RPN1-MECOM, RUNX1-MECOM
Other Tyrosine Kinase	ABL1, DYRK4, EPHA2, EPHA3, JAK3, MST1R, OBSCN, PDGFRB, WEE1
Serine/Threonine Kinase	ACVR2B, ADRBK1, AKAP13, BUB1, CPNE3, DCLK1, MAPK1, YLK2, MYO3A, NRK, PRKCG, RPS6KA6, SMG1, STK32A, STK33, STK36, TRIO, TTBK1, WNK3, WNK4
Protein tyrosine phosphatase	PTPN11, PTPRT, PTPN14
RAS protein	KRAS, NRAS

Table S19. Pairs of genes with significant (*p* < 0.04) exclusivity (a) and co-occurrence (b) for all genes and gene groups with at least 4 mutations.

We removed all three cases of significant co-occurrence where the total number of co-occurring mutations between a pair of genes was two or less. We also calculated the co-occurrence and exclusivity of cytogenetic risk (favorable, intermediate, unfavorable) with genes and gene groups, and included the top four interactions. Highlighted rows indicate pairs with FDR < 0.1 according to the Benjamini-Hochberg-Yekutieli procedure. A graphical representation of this table is shown in **Figure 3**.

a) Significantly exclusive pairs

Gene 1	Gene 2	*p*-value
NPM1	RUNX1	1.80×10^{-3}
PML-RARA	NPM1	5.13×10^{-3}
FLT3	TP53	5.83×10^{-3}
PML-RARA	DNMT3A	7.24×10^{-3}
NPM1	TP53	7.25×10^{-3}
FLT3	RUNX1	1.28×10^{-2}
KRAS/NRAS	FLT3	1.93×10^{-2}
FLT3	Ser/Thr kinases	1.93×10^{-2}
MLL-X fusions	NPM1	2.82×10^{-2}
NPM1	MYH11-CBFB	2.82×10^{-2}
MLL-X fusions	DNMT3A	3.56×10^{-2}
DNMT3A	MYH11-CBFB	3.56×10^{-2}

b) Significantly co-occurring pairs

Gene 1	Gene 2	*p*-value
PML-RARA	Favorable	9.10x10^{-12}
NPM1	Intermediate	1.08x10^{-9}
Favorable	MYH11-CBFB	2.20x10^{-9}
TP53	Unfavorable	3.15x10^{-8}
DNMT3A	NPM1	6.28x10^{-7}
FLT3	NPM1	2.00x10^{-6}
RUNX1-RUNX1T1	Favorable	4.51x10^{-6}
ASXL1	RUNX1	2.81x10^{-4}
ASXL1	IDH2	3.50x10^{-4}
RUNX1	IDH2	8.13x10^{-4}
DNMT3A	IDH1	1.62x10^{-3}
MYH11-CBFB	KIT	6.02x10^{-3}
Cohesin	NPM1	6.32x10^{-3}
RUNX1	PHF6	1.04x10^{-2}
Cohesin	PTPs	1.06x10^{-2}
DNMT3A	Cohesin	1.19x10^{-2}
Ser/Thr kinases	Spliceosome	1.28x10^{-2}
DNMT3A	FLT3	1.36x10^{-2}
NPM1	PTPs	1.85x10^{-2}
MLLT10-PICALM	TET1	1.99x10^{-2}
RUNX1-RUNX1T1	KIT	2.67x10^{-2}
Cohesin	FLT3	2.72x10^{-2}
Ser/Thr kinases	TET2	3.11x10^{-2}
RUNX1-RUNX1T1	Other Tyr kinases	3.37x10^{-2}
NPM1	IDH1	3.79x10^{-2}

Table S20. Small RNA annotation priorities.

For small RNA sequencing, the table shows annotation priorities that are used to resolve multiple database matches for a single alignment location and multiple alignment locations for a read.

Priority	Annotation type	Database
1 2 3 4 5	mature strand star strand precursor miRNA stemloop, from 1 to 6 bases outside the mature strand, between the mature and star strands "unannotated", any region other than the mature strand in miRNAs where no star strand is annotated	miRBase v16
6 7 8 9 10 11 12	snoRNA tRNA rRNA snRNA scRNA srpRNA Other RNA repeats	UCSC small RNAs, RepeatMasker
13 14 15 16,17	coding exons with zero annotated CDS region length 3' UTR 5' UTR coding exon, intron	UCSC genes
18 19 20 21 22 23	LINE SINE LTR Satellite RepeatMasker DNA RepeatMasker Low complexity	UCSC RepeatMasker

24	RepeatMasker Simple Repeat	
25	RepeatMasker Other	
26	RepeatMasker Unknown	

D. References

1 Walter, M. J. *et al.* Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proc Natl Acad Sci U S A* **106**, 12950-12955, doi:10.1073/pnas.0903091106 (2009).

2 Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361**, 1058-1066, doi:NEJMoa0903840 [pii]10.1056/NEJMoa0903840 (2009).

3 Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome biology* **12**, R1, doi:10.1186/gb-2011-12-1-r1 (2011).

4 Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-472, doi:10.1038/nature09837 (2011).

5 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

6 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:btp352 [pii]10.1093/bioinformatics/btp352 (2009).

7 Larson, D. E. *et al.* SomaticSniper: Identification of Somatic Point Mutations in Whole Genome Sequencing Data. *Bioinformatics*, doi:10.1093/bioinformatics/btr665 (2011).

8 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871, doi:btp394 [pii]10.1093/bioinformatics/btp394 (2009).

9 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

10 Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677-681, doi:nmeth.1363 [pii]10.1038/nmeth.1363 (2009).

11 Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506-510, doi:10.1038/nature10738 (2012).

12 Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264-278, doi:10.1016/j.cell.2012.06.023 (2012).

13 Dees, N. D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome Res* **22**, 1589-1598, doi:10.1101/gr.134635.111 (2012).

14 Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817-825, doi:10.1038/nbt.1662 (2010).

15 Wallace, D. C. Mitochondria and cancer. *Nat Rev Cancer* **12**, 685-698, doi:10.1038/nrc3365 (2012).

16 He, Y. *et al.* Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* **464**, 610-614, doi:10.1038/nature08802 (2010).

17 Ruiz-Pesini, E. *et al.* An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res* **35**, D823-828, doi:10.1093/nar/gkl927 (2007).

18 Larman, T. C. *et al.* Spectrum of somatic mitochondrial mutations in five cancers. *Proc Natl Acad Sci U S A* **109**, 14087-14091, doi:10.1073/pnas.1211502109 (2012).

19 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828, doi:10.1126/science.1215040 (2012).

20 Van Vlierberghe, P. *et al.* PHF6 mutations in adult acute myeloid leukemia. *Leukemia* **25**, 130-134, doi:10.1038/leu.2010.247 (2011).

21 Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

22 Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525, doi:10.1038/nature11404 (2012).

23 Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).

24 Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615, doi:10.1038/nature10166 (2011).

25 Payton, J. E. *et al.* High throughput digital quantification of mRNA abundance in primary human acute myeloid leukemia samples. *J Clin Invest* **119**, 1714-1726, doi:10.1172/JCI38248 (2009).

26 Vandin, F., Clay, P., Upfal, E. & Raphael, B. J. Discovery of mutated subnetworks associated with clinical data in cancer. *Pac Symp Biocomput*, 55-66 (2012).

27 Razick, S., Magklaras, G. & Donaldson, I. M. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9**, 405, doi:10.1186/1471-2105-9-405 (2008).

28 Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38**, D355-360, doi:10.1093/nar/gkp896 (2010).

29 Luc, P. V. & Tempst, P. PINdb: a database of nuclear protein complexes from human and yeast. *Bioinformatics* **20**, 1413-1415, doi:10.1093/bioinformatics/bth114 (2004).

30 Vandin, F., Upfal, E. & Raphael, B. J. De novo discovery of mutated driver pathways in cancer. *Genome Res* **22**, 375-385, doi:10.1101/gr.120477.111 (2012).

31 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

32 Morin, R. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45**, 81-94, doi:10.2144/000112900 (2008).

33 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628, doi:10.1038/nmeth.1226 (2008).

34 Khattra, J. *et al.* Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells, and cell lines. *Genome Res* **17**, 108-116, doi:10.1101/gr.5488207 (2007).

35 de Hoon, M. J. *et al.* Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome Res* **20**, 257-264, doi:10.1101/gr.095273.109 (2010).

36 Birol, I. *et al.* De novo transcriptome assembly with ABySS. *Bioinformatics* **25**, 2872-2877, doi:10.1093/bioinformatics/btp367 (2009).

37 Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nat Methods* **7**, 909-912, doi:10.1038/nmeth.1517 (2010).

38 Houseley, J. & Tollervey, D. Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One* **5**, e12271, doi:10.1371/journal.pone.0012271 (2010).

39 Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367, doi:10.1186/1471-2105-11-367 (2010).

40 Rousseeuw, P. J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. . *Journal of Computational and Applied Mathematics.* **20**, 53-65 (1987).

41 Hoon, D. S. Are circulating tumor cells an independent prognostic factor in patients with high-risk melanoma? *Nat Clin Pract Oncol* **1**, 74-75, doi:10.1038/ncponc0041 (2004).

42 Dabney, A. R. ClaNC: point-and-click software for classifying microarrays to nearest centroids. *Bioinformatics* **22**, 122-123, doi:bti756 [pii]10.1093/bioinformatics/bti756 (2006).

43 Mehrian-Shai, R. *et al.* Insulin growth factor-binding protein 2 is a candidate biomarker for PTEN status and PI3K/Akt pathway activation in glioblastoma and prostate cancer. *Proc Natl Acad Sci U S A* **104**, 5563-5568, doi:10.1073/pnas.0609139104 (2007).

44 Mullokandov, G. *et al.* High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nat Methods* **9**, 840-846, doi:10.1038/nmeth.2078 (2012).

45 Verhaak, R. G. *et al.* Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica* **94**, 131-134, doi:10.3324/haematol.13299 (2009).

46 Li, Z. *et al.* Identification of a 24-Gene Prognostic Signature That Improves the European LeukemiaNet Risk Classification of Acute Myeloid Leukemia: An International Collaborative Study. *J Clin Oncol*, doi:10.1200/JCO.2012.44.3184 (2013).

47 Sun, S. M. *et al.* The prognostic relevance of miR-212 expression with survival in cytogenetically and molecularly heterogeneous AML. *Leukemia* **27**, 100-106, doi:10.1038/leu.2012.158 (2013).

48 Rockova, V. *et al.* Risk stratification of intermediate-risk acute myeloid leukemia: integrative analysis of a multitude of gene mutation and gene expression markers. *Blood* **118**, 1069-1076, doi:10.1182/blood-2011-02-334748 (2011).

49 Li, Z. *et al.* Up-regulation of a HOXA-PBX3 homeobox-gene signature following down-regulation of miR-181 is associated with adverse prognosis in patients with cytogenetically abnormal AML. *Blood* **119**, 2314-2324, doi:10.1182/blood-2011-10-386235 (2012).

50 Marcucci, G., Mrozek, K., Radmacher, M. D., Garzon, R. & Bloomfield, C. D. The prognostic and functional role of microRNAs in acute myeloid leukemia. *Blood* **117**, 1121-1129, doi:10.1182/blood-2010-09-191312 (2011).

51 Zhang, X. *et al.* The tumor suppressive role of miRNA-370 by targeting FoxM1 in acute myeloid leukemia. *Molecular cancer* **11**, 56, doi:10.1186/1476-4598-11-56 (2012).

52 Havelange, V. *et al.* Functional implications of microRNAs in acute myeloid leukemia by integrating microRNA and messenger RNA expression profiling. *Cancer* **117**, 4696-4706, doi:10.1002/cncr.26096 (2011).

53 Zhu, Y. D. *et al.* Distinctive microRNA signature is associated with the diagnosis and prognosis of acute leukemia. *Med Oncol* **29**, 2323-2331, doi:10.1007/s12032-011-0140-5 (2012).

54 Faraoni, I. *et al.* MiR-424 and miR-155 deregulated expression in cytogenetically normal acute myeloid leukaemia: correlation with NPM1 and FLT3 mutation status. *Journal of hematology & oncology* **5**, 26, doi:10.1186/1756-8722-5-26 (2012).

55 Zantinge, E. M., Verhaak, P. F., de Bakker, D. H., van der Meer, K. & Bensing, J. M. Does burnout among doctors affect their involvement in patients' mental health problems? A study of videotaped consultations. *BMC family practice* **10**, 60, doi:10.1186/1471-2296-10-60 (2009).

56 Rosales-Avina, J. A. *et al.* MEIS1, PREP1, and PBX4 are differentially expressed in acute lymphoblastic leukemia: association of MEIS1 expression with higher proliferation and chemotherapy resistance. *Journal of experimental & clinical cancer research : CR* **30**, 112, doi:10.1186/1756-9966-30-112 (2011).

57 Mendler, J. H. *et al.* RUNX1 mutations are associated with poor outcome in younger and older patients with cytogenetically normal acute myeloid leukemia and with distinct gene and MicroRNA expression signatures. *J Clin Oncol* **30**, 3109-3118, doi:10.1200/JCO.2011.40.6652 (2012).

58 Schwind, S. *et al.* Prognostic significance of expression of a single microRNA, miR-181a, in cytogenetically normal acute myeloid leukemia: a Cancer and Leukemia Group B study. *J Clin Oncol* **28**, 5257-5264, doi:10.1200/JCO.2010.29.2953 (2010).

59 Seoudi, A. M., Lashine, Y. A. & Abdelaziz, A. I. MicroRNA-181a - a tale of discrepancies. *Expert reviews in molecular medicine* **14**, e5, doi:10.1017/S1462399411002122 (2012).

60 Campan, M., Weisenberger, D. J., Trinh, B. & Laird, P. W. MethyLight. *Methods Mol Biol* **507**, 325-337, doi:10.1007/978-1-59745-522-0_23 (2009).

61 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80, doi:10.1186/gb-2004-5-10-r80 (2004).

62 Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**, 1412-1417, doi:10.1073/pnas.0510310103 (2006).

63 Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat Med* **9**, 811-818 (1990).