Supplementary Materials: Pathology

METHODS

We analyzed material from 230 patients (Supplementary Table 1) who provided informed consent. Tumour and matched normal samples from patients with previously untreated lung adenocarcinoma were selected for analysis according to tumour percentage, availability of clinical data and sufficient nucleic acid as described previously[1]. The Biospecimen Core Resource began by reviewing 678 submitted cases. 120 cases were excluded by pathological assessment (e.g. purity < 60% or necrosis levels > 20%). 195 additional cases did not meet molecular metrics (e.g. RIN<7 or low yield). Six cases were excluded due to discordant tumour/normal genotypes.

We then classified the 289 remaining cases using the 2004 WHO and the 2011 IASLC/ATS/ERS lung adenocarcinoma classification criteria[2]. Neuroendocrine (n=15) and indeterminate (n=44) histologies were excluded. The remaining 230 samples represented the major histologic types of lung adenocarcinoma: 5% lepidic, 33% acinar, 9% papillary, 14% micropapillary, 25% solid, 4% invasive mucinous, 0.4% colloid and 8% unclassifiable adenocarcinoma (Supplementary Figure 1)[2, 3]. Due to artifacts mostly caused by frozen section material, 19 cases were regarded as adenocarcinoma, but histologic subtyping could not be performed. Tumors were classified according to architectural grade as proposed by Yoshizawa A, *et al.*[4].

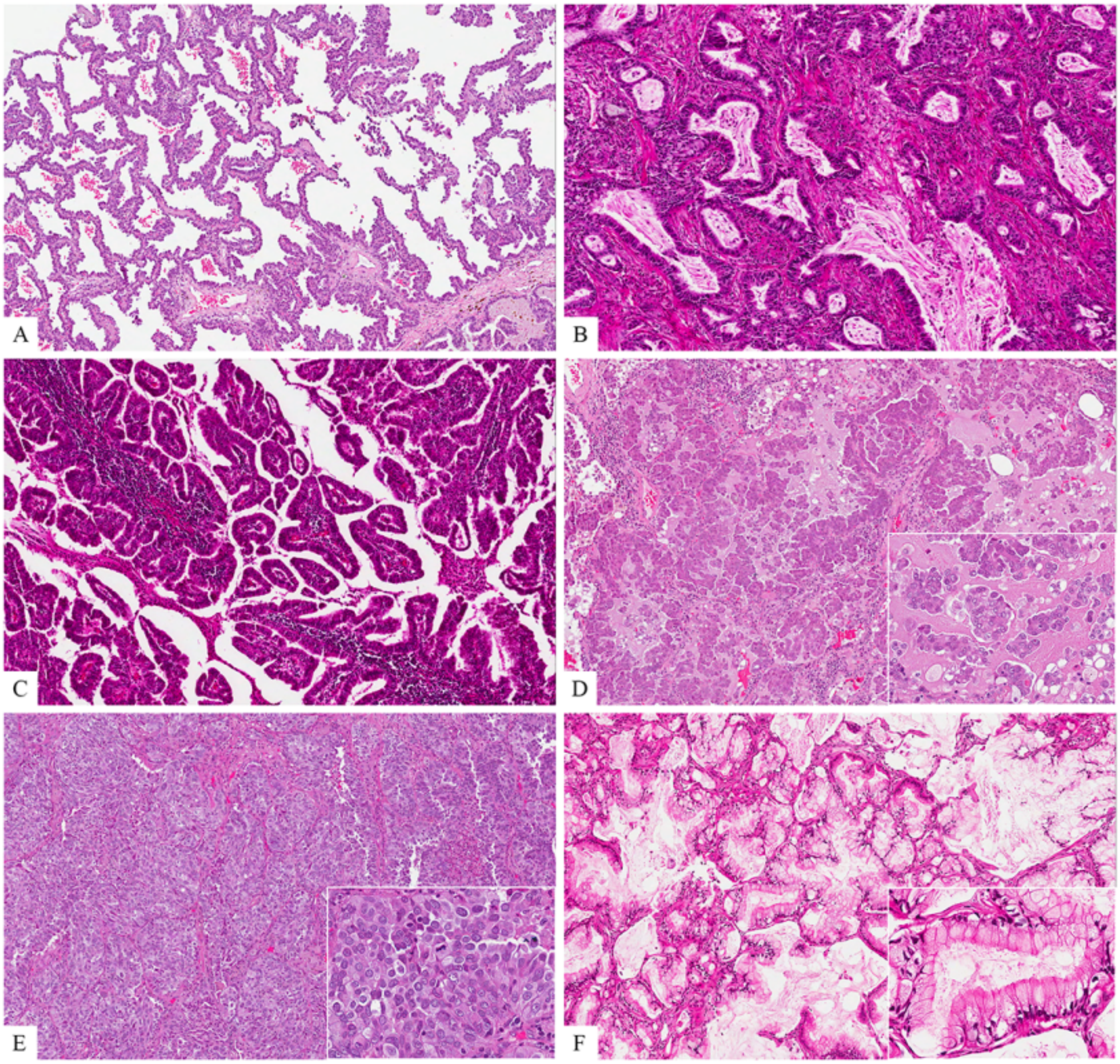
The majority of the tumours analyzed were stage I or II (n=174); the remainder were stage III or IV (n=56). Germline DNA was obtained from peripheral blood (n=133), from adjacent histologically normal tissue resected at surgery (n= 65), or from both (n=32). Supplementary Table 2 summarizes demographic data. Median follow-up was 19 months, and 163 patients were alive at the time of last follow-up. Eighty-one percent of patients reported past or present smoking (current: n=45, former: n=142, never-smoker: n=33, not known: n=10). DNA, RNA and protein were extracted from specimens and quality-control assessments were performed as described previously[1]. Supplementary Table 3 summarizes molecular estimates of tumour cellularity[5].

Histological assessment, molecular quality control, and genotype matching for all samples were performed at the Biospecimen Core Resource (BCR) as previously described[6]. Aperio© scanned hematoxylin and eosin stained slides were reviewed from 289 tumors according to the 2004 WHO classification and the 2011 IASLC/ATS/ERS lung adenocarcinoma classification criteria[2]. Tumors that were unclassifiable (n=44) or suspected to be large cell neuroendocrine carcinoma (n=15) were excluded resulting in 230 cases that were classified as adenocarcinoma. Whenever possible, comprehensive histologic subtyping was performed to determine a predominant subtype. In 161 (70%) of cases, the H&E slide was from a representative section of formalin fixed paraffin embedded tissue while in the remaining 69 (30%) cases, the only slide available for review was from the tissue processed for frozen section. Most of the tumors that were difficult to classify were in the cases were only frozen section material was available for review.

For the purpose of histology molecular correlations, the one case of colloid adenocarcinoma was grouped with unclassified adenocarcinoma. Histologic subtypes were analyzed comparing tumors classified in a specific subtype versus all other histologic subtypes.

RESULTS

Supplementary Figure 1: Histologic patterns of lung adenocarcinoma: A: lepidic pattern with atypical pneumocytes growing along alveolar walls; B: acinar pattern with tumor cells forming glands and tubules; C: papillary pattern with tumor cells growing in papillae along the surface of fibrovascular cores; D: micropapillary pattern with tumor cells growing in papillae lacking fibrovascular cores; E: solid pattern with diffuse sheets of tumor cells lacking any architectural patterns; F: invasive mucinous adenocarcinoma consists of tumor cells with abundant intracytoplasmic mucin, mostly in the apical cytoplasm with basally oriented nuclei.



REFERENCES:

1. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489, 519-25 (2012).
2. Travis, W.D. et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* 6, 244-85 (2011).
3. Travis, W.D., Brambilla, E. & Riely, G.J. New pathologic classification of lung cancer: relevance for clinical practice and clinical trials. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 31, 992-1001 (2013).
4. Yoshizawa, A. et al. Impact of proposed IASLC/ATS/ERS classification of lung adenocarcinoma: prognostic subgroups and implications for further revision of staging based on analysis of 514 stage I cases. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 24, 653-64 (2011).

5. Carter, S.L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* (2012).

6. Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609-15 (2011).

DNA sequencing, validation and data processing

METHODS

Whole exome sequencing was performed as previously described[1]. Briefly, 0.5-3 micrograms of DNA from each sample was used for library preparation, which included shearing and ligation of sequencing adaptors. Exome capture was performed using the Agilent SureSelect Human All Exon 50Mb kit. Captured DNA was sequenced using the Illumina HiSeq platform, and paired-end sequencing reads were generated for each sample. Initial alignment and quality control were performed using the Picard and Firehose pipelines at the Broad Institute[2]. Picard generates a single BAM file for each sample that includes reads, calibrated quantities, and alignments to the genome. Firehose represents a set of tools for analyzing sequencing data from tumor and matched normal DNA. The pipeline uses GenePattern16 as its execution engine, and performs quality control, local realignment, mutation calling using MuTect[3], small insertion and deletion identification using Indelocator[2, 4] rearrangement detection, and coverage calculations, among other analyses. Complete details of this pipeline can be found in Stransky *et al*.[1].

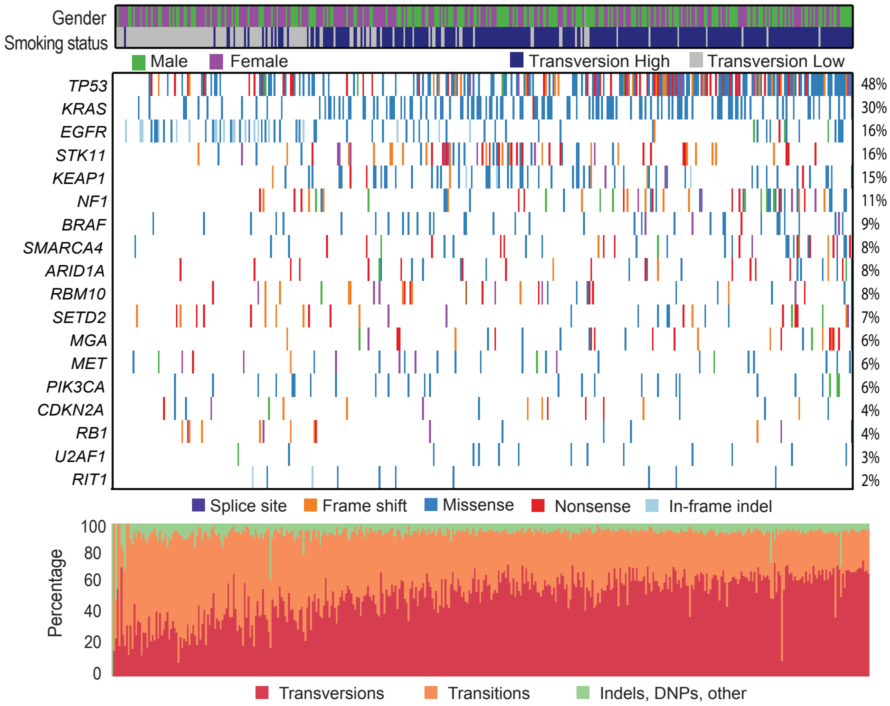
Mutation Significance Analysis

Mutation significance was performed using the MutSig2CV algorithm[5]. In brief, this algorithm takes into account recurrence of mutations, nucleotide context, gene-expression, replication time, and somatic background mutation rate. For this analysis, the cohort of TCGA samples (n=230) was combined with a previously published cohort of lung adenocarcinomas with corresponding WES data (n=182)[6]. Genes with a Bonferroni-corrected p-value less than 0.025 were deemed significant. See **Figure 1A** for the co-mutation plot for the TCGA samples (n=230), **Supplementary Figure 2** for the co-mutation plot for all samples used in the analysis (n=412) and **Supplementary Table 4** for a list of all genes and their respective p-values. For the significant genes, we applied a Fisher's exact test to determine if the proportions of mutated samples differed between Transversion-High and Transversion-Low or male and female sample subsets. These p-values were corrected using the Benjamini-Hochberg multiple test correction method.

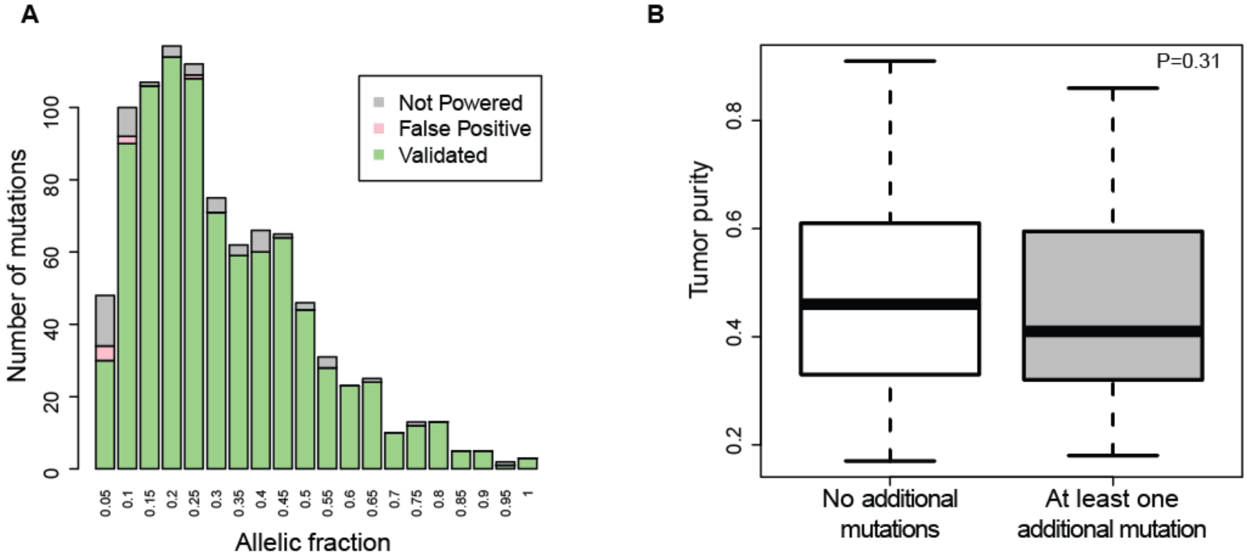
Mutation and indel validation

Using novel baits, we used hybrid captured to resequence the exons of 70 candidate genes to an average depth of 501 reads for 164 of the samples used in the original cohort. Validation rates were calculated as previously described[1, 6] and are shown in **Supplementary Figure 3** and **Supplementary Table 5**. Briefly, we used a binomial model to first determine the minimum alternate read count required for validation based on noise in the normal. We then adopted a second binomial to calculate the power to detect a mutation or indel and analyzed only those sites whose power exceeded 95%. This model takes into consideration the allele fraction of the event, the depth of coverage, and the expected error rate at that site.

Supplementary Figure 2: Co-mutation plot for whole exome sequencing analysis of 412 lung adenocarcinomas. Significant genes were identified using the MutSig2CV algorithm (Bonferroni-corrected p < 0.025) and are ranked in order of decreasing significance. See Figure 1A for a co-mutation plot of the 230 TCGA samples.



Supplementary Figure 3: Validation rates for SNPs and Indels. (A) Considering sites with a 95% power of detection, the validation rate was 99% and 100% for SNPs and indels, respectively. Additionally, 235 mutations and 9 indels were identified *de novo* in the validation data suggesting false negative rates of 23% and 17%, respectively. The majority of these additional mutations (64%) lacked adequate coverage for detection in the whole exome sequencing data (power < 95%). **(B)** Samples with at least one additional mutation identified (n=107) did not have significantly lower tumor purity than samples that had no additional mutations identified (p=0.31; Wilcoxon rank-sum test).



References:

1. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-25 (2012).
2. Stransky, N. et al. The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-60 (2011).
3. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-9 (2013).
4. Banerji, S. et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405-9 (2012).
5. Lawrence, M.S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-8 (2013).
6. Imielinski, M. et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107-20 (2012).

Supplementary Materials: Transversion High/Low Analysis

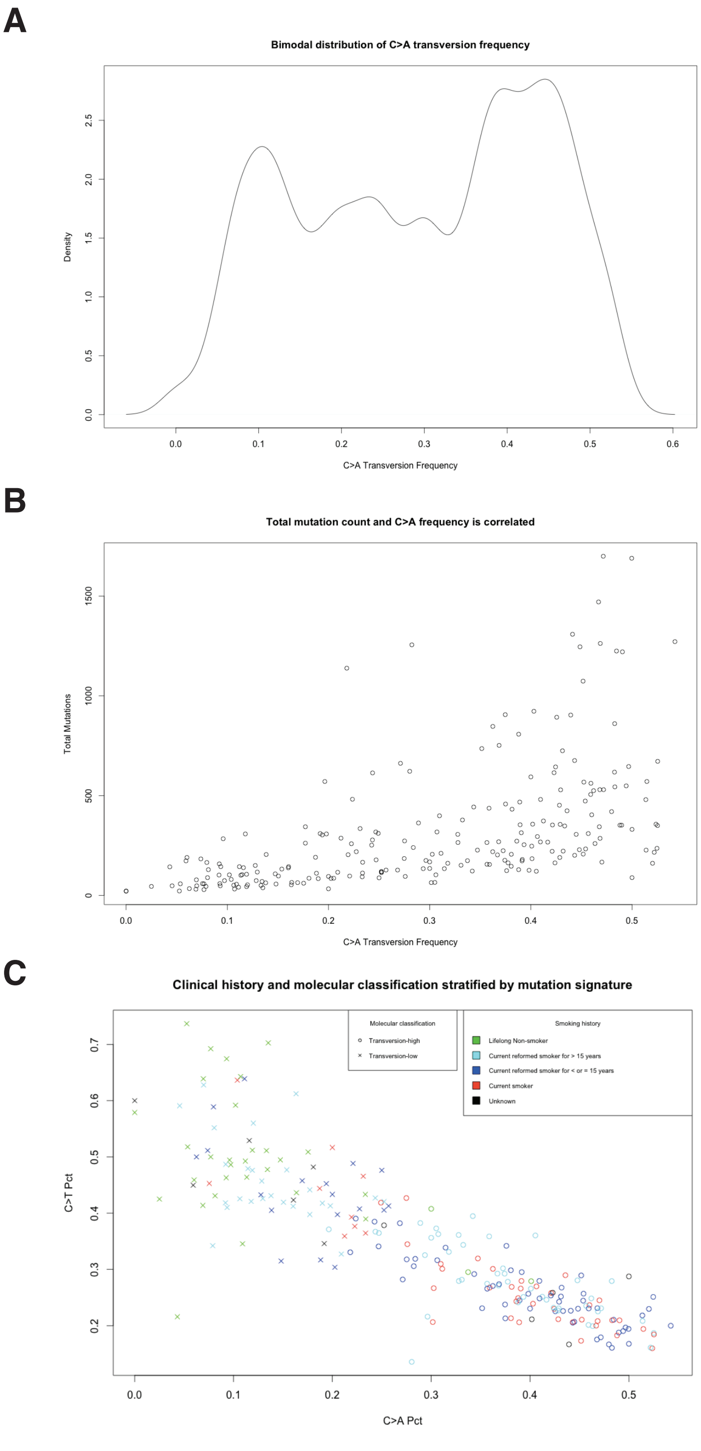
METHODS

230 TCGA samples were merged with 183 lung adenocarcinoma samples from a previously published study[1]. Mutation spectrum for each sample was calculated as the percentage of each of six possible single nucleotide changes (AT>CG, AT>GC, AT>TA, GC>AT, GC>CG, GC>TA) among all single nucleotide substitutions. A training set was established with transversion low (TL) samples being tumors from lifelong never-smokers and transversion high (TH) samples as tumors from patients with 60 or more pack years of smoking history. A linear discriminant analysis based on GC>AT frequency, GC>TA frequency, and total mutation count (using the R MASS library[2]) was performed based on this training set to classify all samples as belonging to either the TH or TL categories.

RESULTS

Among 413 total samples, 144 were classified as transversion low and 269 as transversion high. For 59 lifelong never-smokers in the merged sample set, 54 (92%) were classified as TL. For 163 current or heavy smokers, 137 (84%) were classified as TH. Significantly mutated genes were calculated separately for each sample subset with selected genes shown in Figure 1B.

Supplementary Figure 4: Transversion frequency correlates with smoking history. **A)** Distribution of C>A transversion frequency among lung adenocarcinoma samples showing two peaks. **B)** Total mutation count and C>A transversion frequency is correlated, R^2=0.30. **C)** C>A transversion frequency and C>T transition frequency are inversely correlated (R^2=0.75). Transversion High and Transversion Low classification denoted by point shape, annotated clinical smoking history denoted by point color



REFERENCES

1. Imielinski, M. *et al.* Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing. *Cell* **150,** 1107–1120 (2012).

2. W. N. Venables & B. D. Ripley. *Modern Applied Statistics with S*. (Springer, 2002). at <http://www.stats.ox.ac.uk/pub/MASS4>

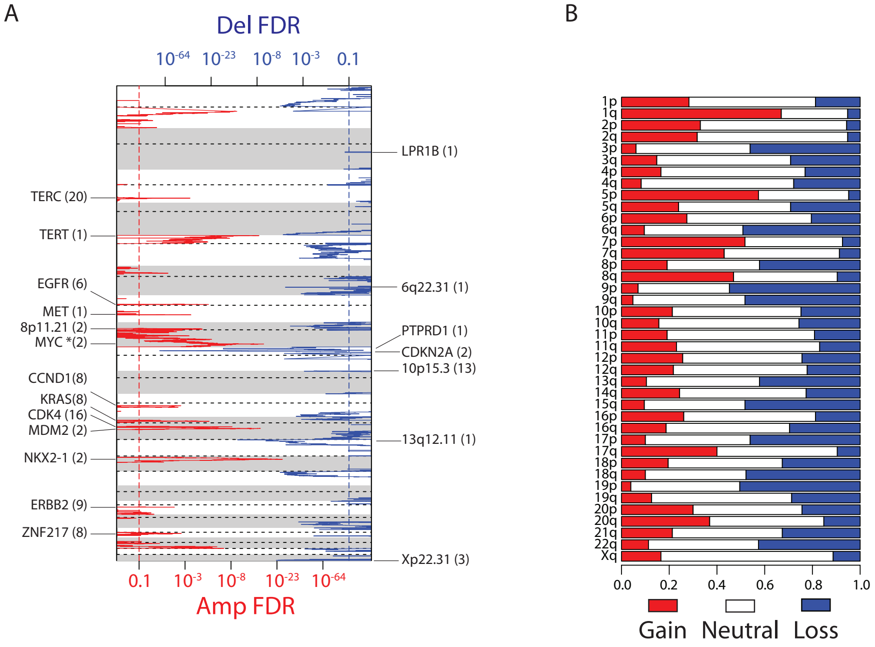
Supplementary Materials: Copy Number Analysis & low pass Whole Genome Sequencing

METHODS

SNP Array-Based Copy Number Analysis: DNA from each tumor or germline-derived sample was hybridized to the Affymetrix SNP 6.0 arrays using protocols at the Genome Analysis Platform of the Broad Institute[1]. From raw .CEL files, Birdseed was used to infer a preliminary copy-number at each probe locus[2]. For each tumor, genome-wide copy number estimates were refined using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumor[3] (and Tabak B and Beroukhim R. Manuscript in preparation). This linear combination of normal samples tends to match the noise profile of the tumor better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy-number profile. Individual copy-number estimates then undergo segmentation using Circular Binary Segmentation[4]. As part of this process of copy-number assessment and segmentation, regions corresponding to germline copy-number alterations were removed by applying filters generated from either the TCGA germline samples from the ovarian cancer analysis or from samples from this collection. Segmented copy number profiles for tumor and matched control DNAs were analyzed using Ziggurat Deconstruction, an algorithm that parsimoniously assigns a length and amplitude to the set of inferred copy number changes underlying each segmented copy number profile[4]. Analysis of broad copy number alterations was then conducted as previously described[2]. Significant focal copy number alterations were identified from segmented data using GISTIC 2.0[5]. Allelic copy number, and purity and ploidy estimates were calculated using the ABSOLUTE algorithm[6].

RESULTS

Supplementary Figure 5: Copy Number Aberrations in Lung Adenocarcinoma. (**A**) Statistically significant focally amplified (red) and deleted (blue) regions plotted along the genome. Annotated regions are FDR<0.05. (**B**) Arm level copy number changes with red representing arm level gains and blue representing arm level losses.



WGS (low-pass) Based Analysis of Structural Variations

METHODS

From 500 to 700 ng of each sample gDNA were sheared using Covaris E220 to about 250 bp fragments, than converted to a pair-end Illumina library using KAPA Bio kits with Caliper (PerkinElmer) robotic NGS Suite according to manufacturers' protocols. All libraries were sequenced by HiSeq2000 using one sample, one lane, pair-end 2 x 51bp setup. Tumor and its matching normal were usually loaded to the same flowcell. Average sequence coverage was found to be 5.6X, read quality 37.3, 92.13% reads mapped. Raw data were converted to FASTQ format then were fed to BWA alignment software to generate .bam files.

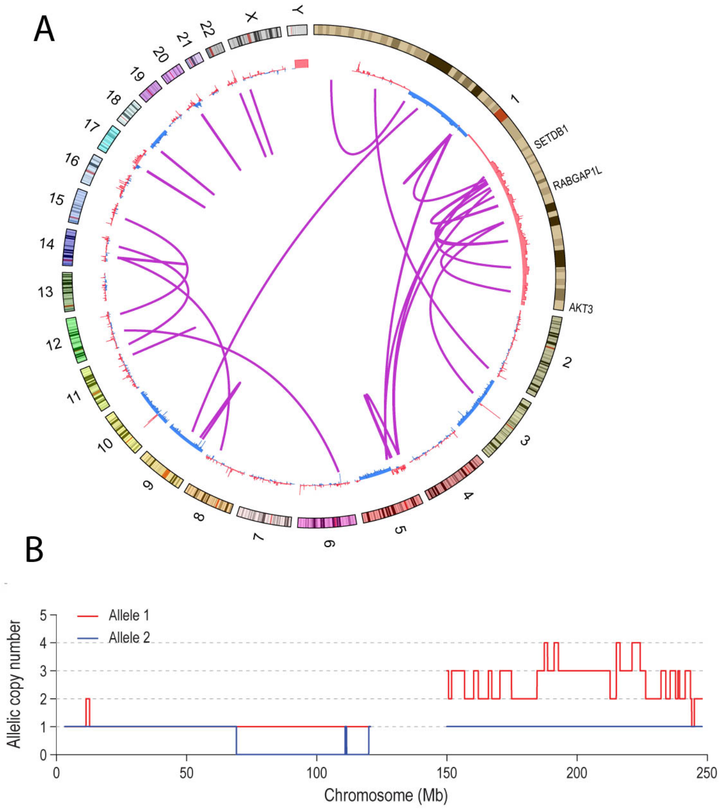
Identification of Copy Number Variants. To characterize somatic copy number alterations in the tumor genome, we applied a new algorithm called BIC-seq[7] to low-coverage whole-genome sequencing data. First, we counted the uniquely aligned reads in fixed-size, non-overlapping windows along the genome. Given these bins with read counts for tumor and matched normal genomes, BIC-seq attempts to iteratively combine neighboring bins with similar copy numbers. Whether the two neighboring bins should be merged is based on Bayesian Information Criteria (BIC), a statistical criterion measuring both fitness and complexity of a statistical model. Segmentation stops when no merging of windows improves BIC, and the boundaries of the windows are reported as a final set of copy number breakpoints. Segments with copy ratio difference smaller than 0.1 (log2 scale) between tumor and normal genomes were merged in the post-processing step to avoid excessive refinement of altered regions with high read counts.

Discovery of Rearrangements with BreakDancer and MEERKAT. Structural variation detection is performed with the program BreakDancer on a .bam file constructed from HiSeq sequencing of each tumor pair[8]. The first step requires a configuration file of each bam file for each tumor pair with the bam2cfg.pl perl module of the program. After the configuration file, the perl module BreakDancerMax.pl is run on the configuration file in order to call structural variants in the tumor and control files. Each tumor structural variant file is filtered with its matched normal to remove any false positives. Structural variations are also detected by Meerkat[9], which requires at least two discordant read pairs supporting one event and at least one read covering the breakpoint junction. Each variant detected from tumor genome is filtered with all normal genomes to remove germline events. The final calls are also filtered out if both breakpoints fall into simple repeats or satellite repeats.

Validation of Rearrangement Hits. We attempted to validate the translocations using two different approaches. MEERKAT determines translocations on the basis of discordant reads as well as reads that span the translocation junction (split reads). We also attempted to validate several translocations by attempting to PCR amplify the junctions of the translocation and sequencing the products. Based on these two approaches we validated 25/46 (54%) of translocations. Therefore, it is possible that the false discovery is 46 percent.

RESULTS

Supplementary Figure 6: Chromothripsis in LUAD 05-5715. A. Circos diagram depicting the region of Chromosome 1 that suffered chromothripsis. Inner tracks display copy number histograms for each chromosome, amplifications (red) and deletions (blue), and structural variation links are purple. Some genes annotated on chromosome 1 include *SETDB1*, *RABGAP1L*, and *AKT3*. Part B. Chromothripsis allele graphs display alternating copy states on Allele 1 (top panel) and no copy state alteration on Allele 2 (bottom panel).



REFERENCES

1 McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics* **40**, 1166-1174, doi:10.1038/ng.238 (2008).

2 Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**, 1253-1260 (2008).

3 The Cancer Genome Atlas Research Network & Perou, C. M. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615, doi:10.1038/nature10166 (2011).

4 Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572, doi:10.1093/biostatistics/kxh008 (2004).

5 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology* **12**, R41, doi:10.1186/gb-2011-12-4-r41 (2011).

6 Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* **30**, 413-421, doi:10.1038/nbt.2203 (2012).

7 Xi, R. *et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proceedings of the National Academy of Sciences of the United States of America* **108**, E1128-1136, doi:10.1073/pnas.1110574108 (2011).

8 Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods* **6**, 677-681, doi:10.1038/nmeth.1363 (2009).

9 Lixing Yang, L. J. L., Nils Gehlenborg, Ruibin Xi, Psalm S. Haseley, Chih-Heng Hsieh,, Chengsheng Zhang, X. R., Alexei Protopopov, Lynda Chin, Raju Kucherlapati, Charles Lee, & Park, a. P. J. Diverse Mechanisms of Somatic Structural Variations in Human Cancer Genomes. *Cell* **153**, 1-11 (2013).

Supplementary Materials: RNA Sequencing

METHODS

Expression quantification

RNA was extracted, prepared into mRNA libraries, and sequenced by Illumina HiSeq resulting in paired 50nt reads, and subjected to quality control as previously described[1]. RNA reads were aligned to the hg19 genome assembly using Mapsplice.[2] Gene expression was quantified for the transcript models corresponding to the TCGA GAF2.1[3], using RSEM[4] and normalized within-sample to a fixed upper quartile. For further details on this processing, refer to Description file at the DCC data portal under the V2_MapSpliceRSEM workflow.[5] For gene level analyses, expression values of zero were set to the overall minimum value, and all data were log2 transformed.[6]

For splicing analyses in Fig 2A, exonic RPKM values (described in [5]) were utilized. For a gene, exons were standardized to z-scores within a sample and then across exons.

Fusion transcript detection

RNA fusion events were automatically detected by MapSplice as previously described[1,7]. A secondary search for fusion events was performed for the genes *RET*, *ROS1*, and *ALK*. This secondary search, by *SigFuge* ([8] and http://bioconductor.org/packages/release/bioc/html/SigFuge.html), consisted of clustering tumor specimens by base-wise expression profiles, separately for each gene, such that highly discordant expression classes indicative of fusion events might be found. This process confirmed the automatic detections of MapSplice and additionally made one further detection of a *RET* fusion in TCGA-75-6203. Manual analysis confirmed the presence of aligned RNA reads bridging this locus and its upstream partner *CCDC6*.

RNA-seq mutation confirmation

Each tumor's sequence mutations detected by DNA whole exome sequencing (DNA-WES) were interrogated for confirmation in their RNA sequencing (RNA-seq) using the tool *UNCeqR* ([9] and http://lbg.med.unc.edu/tools/unceqr/) as previously described[1.] First, DNA-WES mutation positions having a minimum of 1X RNA depth were evaluated for the presence of at least one read confirming the variant allele (Supplementary Figure 7A - top panel). With this definition of expression, 64% of mutation loci were expressed, 40% were confirmed, and 62% were confirmed if the locus was expressed, on average among samples. Then, DNA-WES mutation positions having a minimum of 5X RNA depth were evaluated for the presence of at least one confirming read with the variant allele (Supplementary Figure 7A – middle panel). With this definition of expression, 49% of mutation loci were expressed, 36% were confirmed, and 73% were confirmed if the locus was expressed, on average among samples. As expected, the confirmation rate is greater among positions with minimum 5X coverage because there were more RNA reads and greater statistical power to possibly confirm the variant allele than there was in the positions with minimum 1X coverage. Overall, these results demonstrate a high, independent validation rate of DNA-WES mutations by RNA-seq conditioned on moderate RNA expression.

Across all specimens, mutation validation rates by RNA-seq were not widely different among mutations with different predicted protein coding ramifications (Supplementary Figure 7A – bottom panel). DNA mutations having non-sense, mis-sense and silent predicted protein coding effects displayed 37%, 41%, and 40% validation rates by RNA-seq, overall irrespective of locus expression. Limiting to mutation sites with at least 5X RNA read depth, non-sense, mis-sense and silent mutations had 67%, 78%, and 76% respective validation rates. These results support that RNA-seq is effective at validating DNA mutations of many varieties in lung adenocarcinoma, similar to that previously shown for lung squamous cell carcinoma[1].

RNA-seq validated mutations are listed in[10].

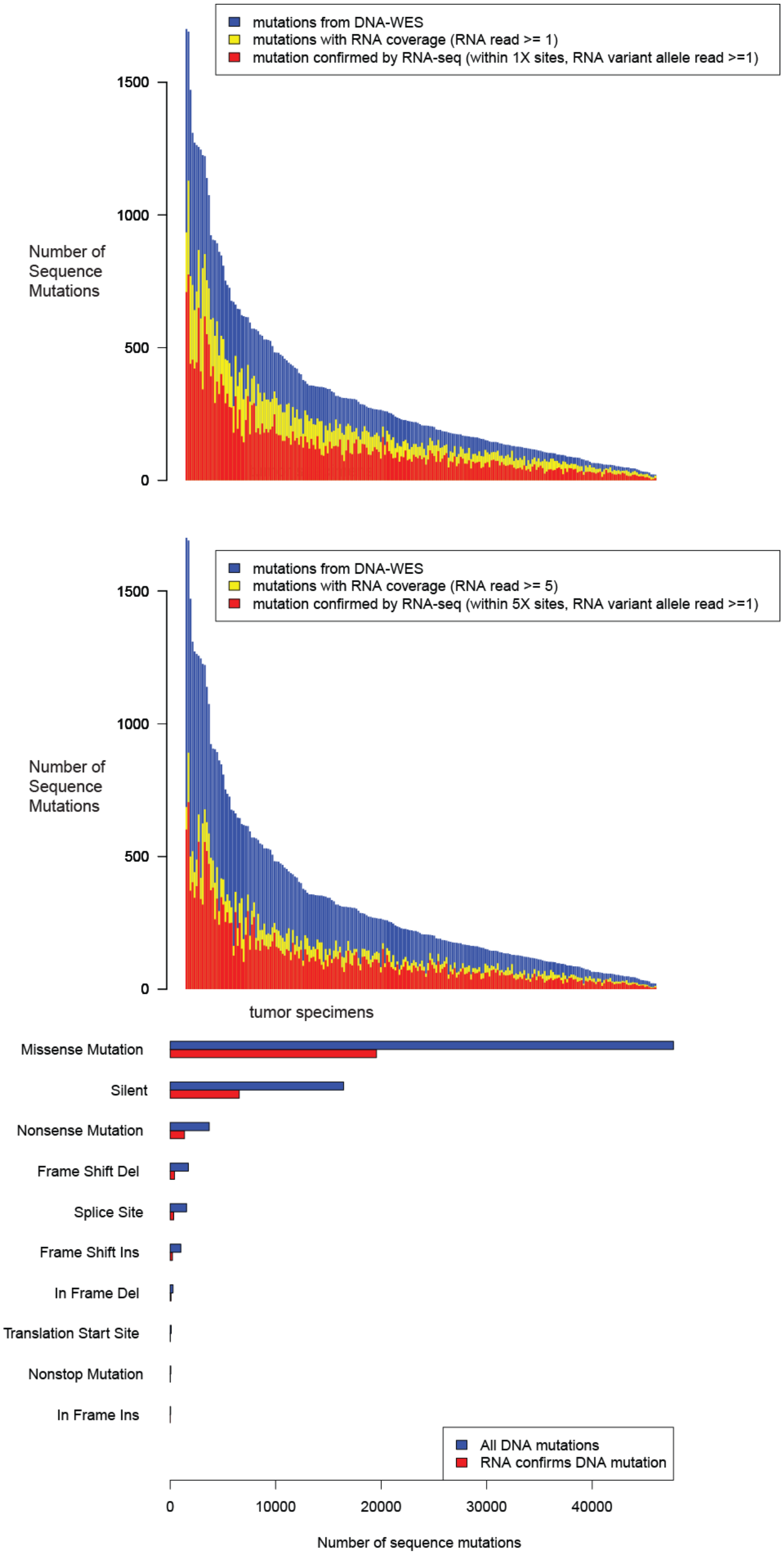
Expression subtype detection

Previously validated gene expression subtypes of lung adenocarcinoma[11,12] were detected in the TCGA lung adenocarcinoma cohort. Gene expression data were gene median centered. Using previously published predictor centroids[12,13] subtype was assigned to each TCGA tumor specimen using a nearest centroid predictor[13], limiting to the genes common to the predictor and the TCGA cohort (*n* = 489) and using Pearson correlation as the similarity metric, with the maximum correlation coefficient providing the subtype prediction for a tumor (subtype calls in [14]). To empirically assess the quality of these subtype detections similar to earlier studies[1,12], expression of the predictor genes was compared between the TCGA cohort and the previously published Wilkerson et al. cohort (Supplementary Figure 7B). Subtype expression patterns were highly concordant between the cohorts, indicating that the subtypes are a similar stratification of the TCGA cohort as in earlier cohorts.

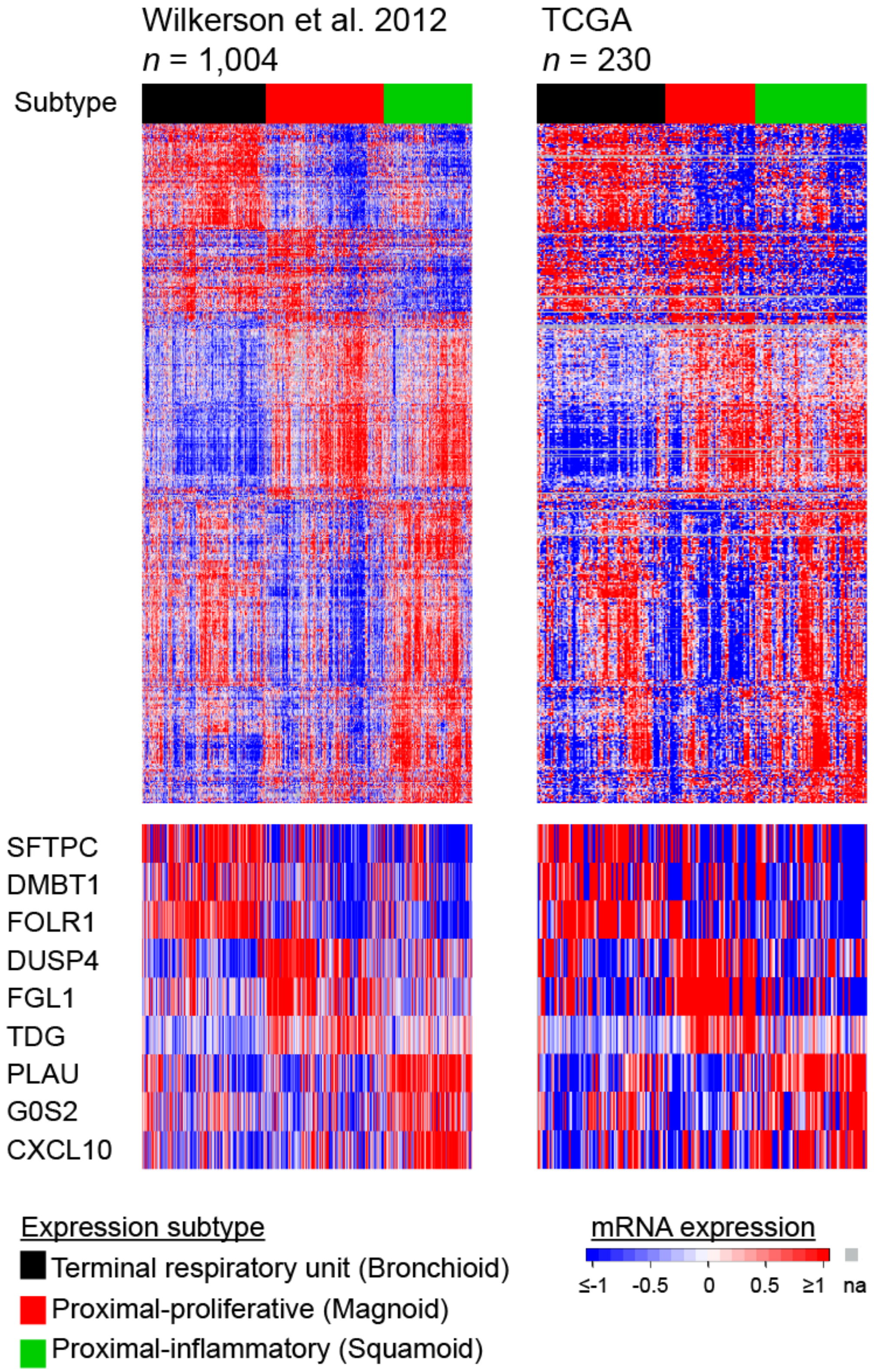
The subtypes' names were updated to be consistent with recent changes to morphological classification and to reflect distinct expression properties of the subtypes, as follows: Bronchioid to Terminal Respiratory Unit (TRU), Magnoid to Proximal-proliferative (PP), and Squamoid to Proximal-inflammatory (PI). Patient overall survival was compared among the subtypes, limited to those patients with follow up information and censoring patients having died within 30 days of surgery. The TRU subtype exhibited a superior outcome relative to the other two subtypes (Supplementary Figure 7C), consistent with earlier studies[11,12]. Limiting cases to N0 or N0 and N1, TRU continued to have superior outcome.

RESULTS

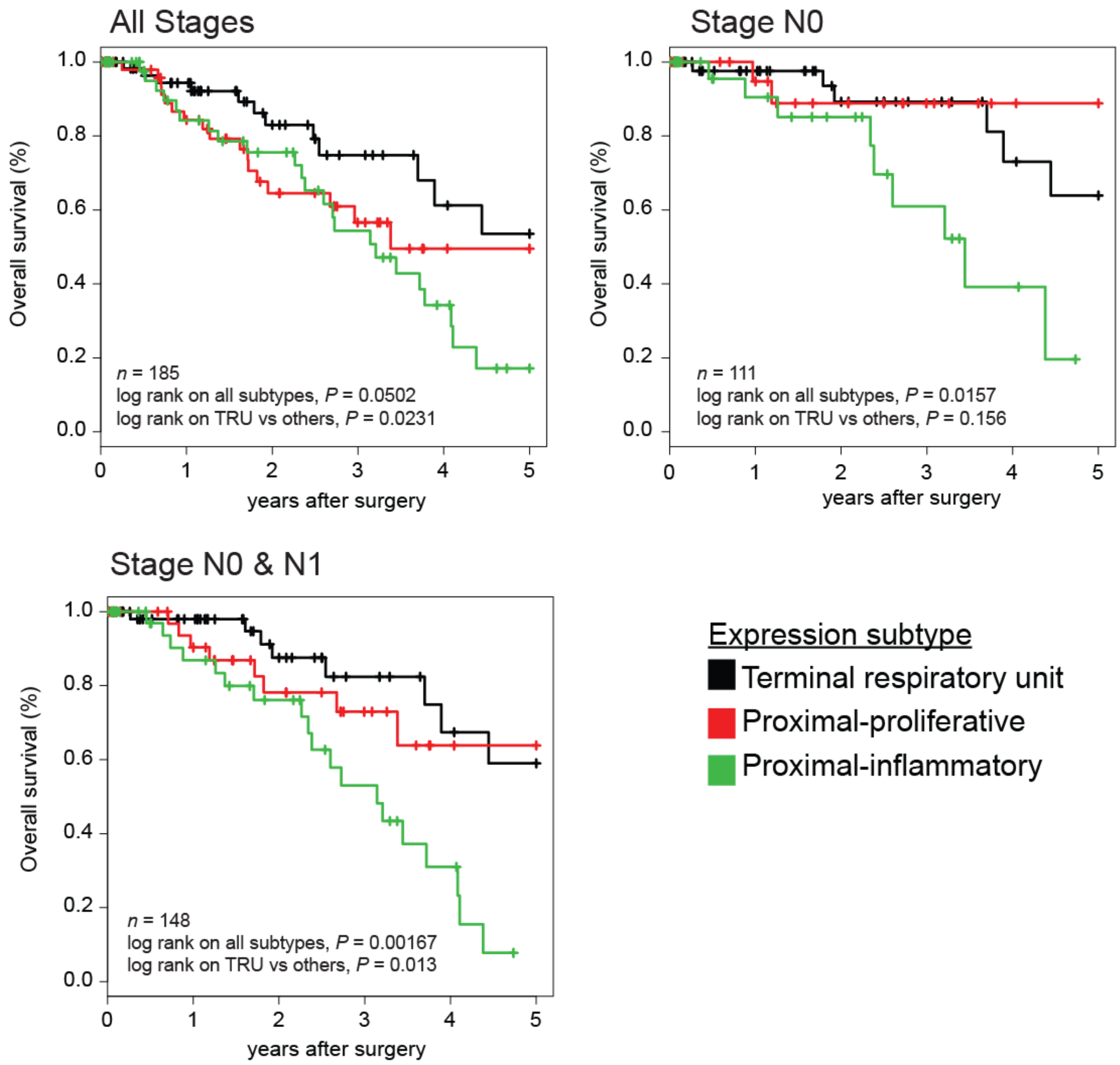
Supplementary Figure 7A: DNA-WES sequence mutation validation by RNA-seq.



Supplementary Figure 7B: Expression subtype detection.



Supplementary Figure 7C: Survival outcome of expression subtypes.



REFERENCES:

1. The Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012 Sep 27;489(7417):519-25
2. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J. (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Research, 2010 Oct;38(18):e178
3. http://tcga-data.nci.nih.gov/docs/GAF/GAF.hg19.June2011.bundle/outputs/TCGA.hg19.June2011.gaf
4. Li B, Dewey CN. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Aug 4;12:323.
5. https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/luad/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_LUAD.IlluminaHiSeq_RNASeqV2.mage-tab.1.9.0/DESCRIPTION.txt
6. https://tcga-data.nci.nih.gov/docs/publications/luad_2014/tcga.luad.rnaseq.20121025.csv.zip
7. https://tcga-data.nci.nih.gov/docs/publications/luad_2014/tcga.luad.unc.rna.fusions.20121029.csv.zip
8. Kimes PK, Cabanski CR , Wilkerson MD, Zhao N, Johnson AR, Perou CM, Makowski L, Maher CA, Liu Y, Marron JS, Hayes DN. (2014) SigFuge: single gene clustering of RNA-seq reveals differential isoform usage among cancer samples. Nucleic Acids Research, In Press.
9. Wilkerson MD, Cabanski CR, Sun W, Hoadley KA, Walter V, Mose LE, Troester MA, Hammerman PS, Parker JS, Perou CM, Hayes DN. (2014) Integrated RNA and DNA sequencing improves mutation detection in low purity tumors Nucleic Acids Research. first published online June 26, 2014 doi:10.1093/nar/gku489
10. https://tcga-data-secure.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/tcga4yeo/other/publications/luad_2014/mafX.20131024.csv.zip
11. Hayes DN, Monti S, Parmigiani G, Gilks CB, Naoki K, Bhattacharjee A, Socinski MA, Perou C, Meyerson M. (2006). Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. J Clin Oncol. 2006 Nov 1;24(31):5079-90
12. Wilkerson MD, Yin X, Walter V, Zhao N, Cabanski CR, Hayward MC, Miller CR, Socinski MA, Parsons AM, Thorne LB, Haithcock BE, Veeramachaneni NK, Funkhouser WK, Randell SH, Bernard PS, Perou CM, Hayes DN. (2012) Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. PLoS One. 2012;7(5):e36530
13. http://cancer.unc.edu/nhayes/publications/adenocarcinoma.2012/wilkerson.2012.LAD.predictor.centroids.csv.zip
14. https://tcga-data.nci.nih.gov/docs/publications/luad_2014/tcga.luad.gene.expression.subtypes.20121025.csv

Supplementary Material: RNA splicing

METHODS

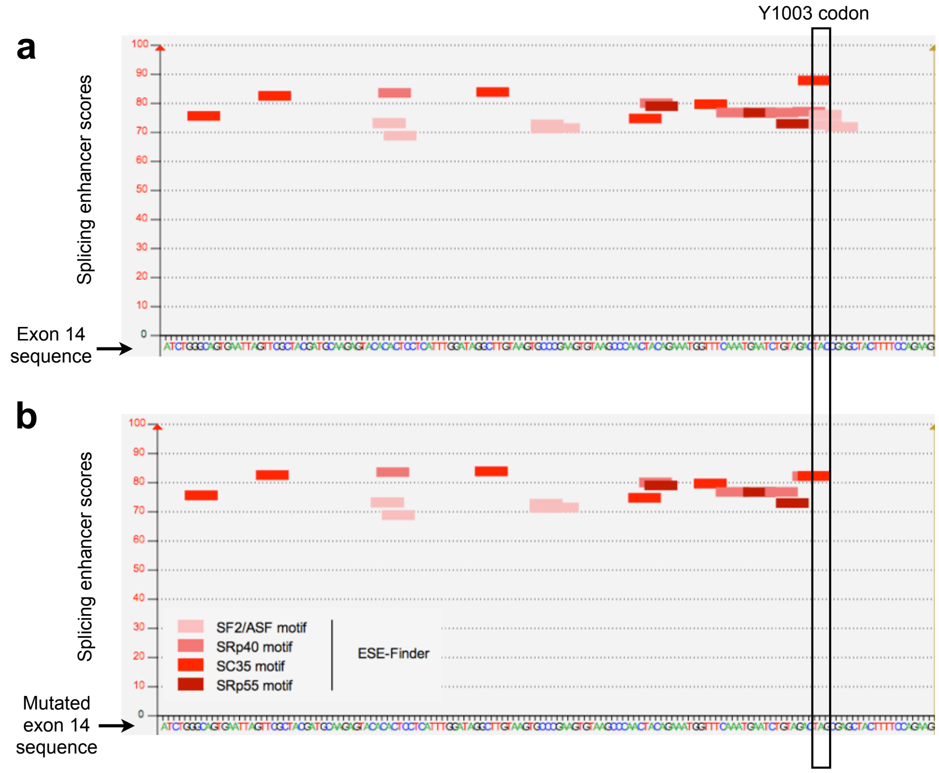
BAM files from 230 tumor samples of MapSplice alignments[1] were used as input to JuncBASE for analysis of alternative splicing[2]. Briefly, JuncBASE identifies and quantifies alternative splicing using exon-exon junctions present in the BAM files and also exon coordinates from annotated or *de novo* transcript assembly sources. Cufflinks[3] transcript assembly was performed to identify novel exons to use as input for JuncBASE. JuncBASE analysis was run with a junction sequence length of 88 and a junction entropy cutoff of 2. Reference exon coordinates were derived from UCSC known genes hg19[4]. Further analysis was done on "percent spliced in" values (inclusion isoform abundance/total abundance) given by JuncBASE for each of the 29,857 alternative splicing event identified in the 230 samples.

To identify alternative splicing events that were significantly differentially expressed in the presence of a *U2AF1* S34F mutation, we used a Mann-Whitney test comparing "percent spliced in" values between 8 *U2AF1* S34F tumors and 222 *U2AF1* WT tumors (Benjamini-Hochberg correction, FDR < 0.05). We further filtered for stronger effects where the difference in median "percent spliced in" values of the *U2AF1* WT and *U2AF1* S34F tumors were greater than 10%.

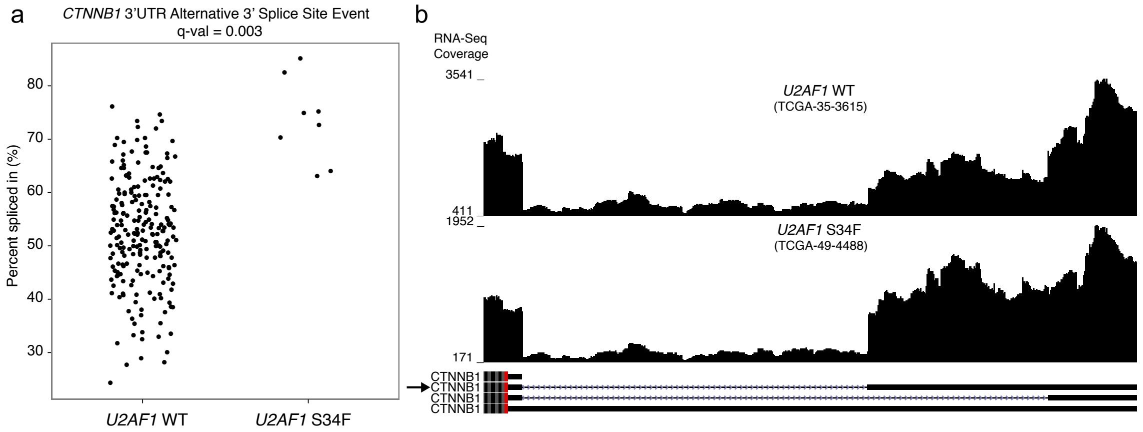
JuncBASE identified and quantified the skipping of exon 14 of *MET* in 9/230 samples. One tumor sample (TCGA-50-6597) contained an exon 14 splice site mutation; however, it had insufficient coverage around exon 14 for JuncBASE quantification. We were able to confirm skipping of exon 14 in this sample by the presence of multiple exon-exon junction reads spanning exon 13 and exon 15. The ten samples described above were the only samples that contained any exon-exon junction reads spanning exon 13 and exon 15 of *MET*.

RESULTS

Supplementary Figure 8A: Scores for predicted splicing enhancer sequences within (**a**) wild-type MET exon 14 and (**b**) Y1003* (c.3009C>G) as determined by Human Splicing Finder [Human Splicing Finder. Desmet *et al.* NAR 2009]. Rectangles correspond to the splicing enhancer scores using ESE Finder matrices for n-mers. N-mers that score above the default threshold values are shown.



Supplementary Figure 8B: Increased expression of a *CTNNB1* isoform associated with *U2AF1* S34F tumors. (**a**) Percent spliced in values of a *CTNNB1* 3'UTR alternative 3' splice site event shown for *U2AF1* WT tumors and *U2AF1* S34F tumors. (**b**) RNA-Seq coverage from a representative *U2AF1* WT sample and *U2AF1* S34F sample of the alternatively spliced region (http://genome.ucsc.edu[5]). The isoform with increased expression in *U2AF1* S34F samples is indicated with the arrow.



REFERENCES:

1. Wang, K. et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research* **38**, e178 (2010).

2. Brooks, A.N. et al. Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res* **21**, 193-202 (2011).

3. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-5 (2010).

4. Meyer, L.R. et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic acids research* **41**, D64-9 (2013).

5. Kent, W.J. et al. The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).

Supplementary materials: Oncogene Discovery Analysis

To identify mutations enriched in the subset of tumors lacking canonical lung adenocarcinoma mutations, we first defined which tumors lacked canonical driver mutations via the following analysis:

Definition of canonical driver mutations: We curated canonical driver genes from two sources[1, 2]. We considered only genes with mutually exclusive mutation patterns (excluding *PIK3CA*). Therefore the genes considered "drivers" were: *KRAS, EGFR, ERBB2, BRAF, MET, ALK* fusion genes*, RET* fusion genes*, ROS1* fusion genes*, HRAS, NRAS,* and *MAP2K1.*

Next, mutations were filtered to include only those with either evidence of recurrence within the COSMIC database[3] (>3 independent mutations at the same site) or evidence of functional impact (e.g. *MAP2K1* p.C121S[4] and *MET* exon 14 deletions[5, 6]). Supplementary Figure 9a displays the mutations identified in each gene and those that were considered of known significance (black) or of unknown significance (gray).

After mutation filtering, we considered any sample having a mutation in one of the above listed genes listed as belonging to the "oncogene-positive" group (n = 143). Samples lacking any of the mutations were considered "oncogene-negative" (n = 87). Supplementary Table 7 shows the oncogene-positive or negative classification for each sample as well as mutation status for the oncogenes listed above.

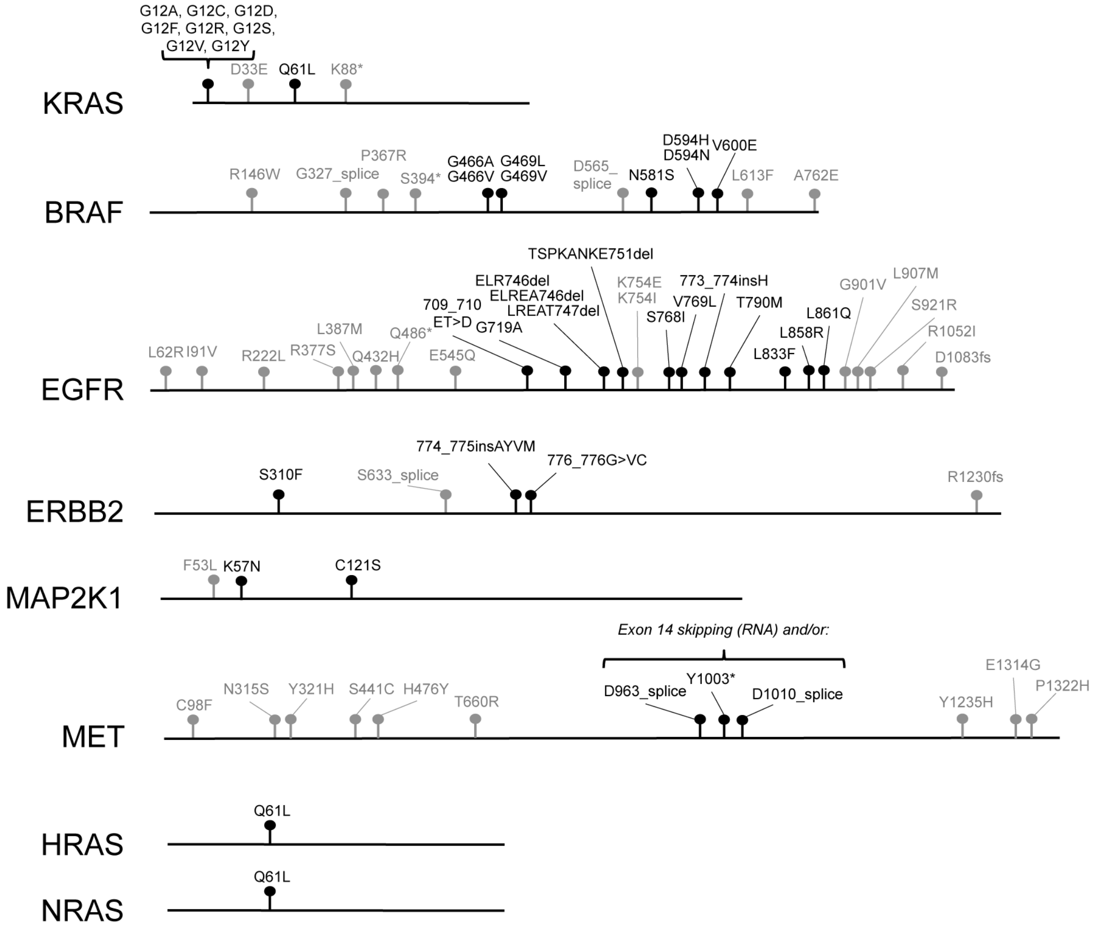
We identified new candidate mutually exclusive mutated "driver" genes by applying a Fisher's exact test p-value for the difference in percentage of samples mutated in that gene between the "oncogene-negative" vs. "oncogene-positive" sample sets. These p-values were then corrected using the Benjamini-Hochberg multiple test correction method. See Supplementary Table 12 for the ranked list.

To identify novel mutually exclusive somatic copy number alterations, GISTIC analysis was performed on the "oncogene-positive" and "oncogene-negative" subsets as defined above and significant focal amplification peaks compared between the two, revealing *MET* and *ERBB2* amplification peaks as specific to the oncogene-negative subset (Figure 3b). While moderate amplification of *MET* and *ERBB2* sometimes co-occurred with other somatic driver events (Supplementary Table 7), amplifications with accompanying high-level overexpression (RNASeq V2 RSEM z-score > 10) of *MET* or *ERBB2* were mutually exclusive with other driver events (Figure 3C).

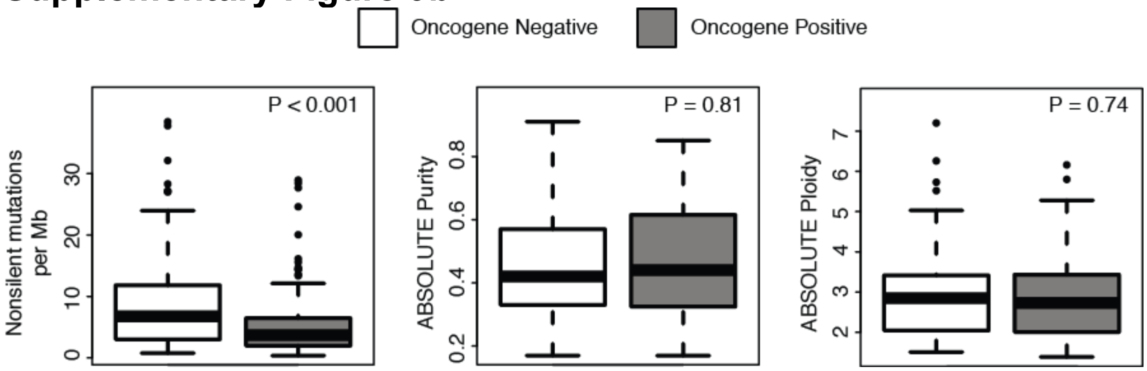
RESULTS

Supplementary Figure 9 – Oncogene definition and nonsilent mutation rate, purity and ploidy between oncogene-positive and oncogene-negative tumors. a, Schematics of known driver oncogenes showing nonsilent somatic variants observed in this dataset. Each lollipop indicates a different somatic variant. Black, variants of known significance as described above. Gray, variants of unknown significance excluded from the "oncogene-positive" classification. **b,** No significant differences were observed in ABSOLUTE purity or ploidy between oncogene-positive and oncogene-negative tumors (p > 0.05; Wilcoxon rank-sum test). Oncogene-negative tumors did however exhibit a higher overall mutation rate (p < 0.001). See Figure 3 in the main text for a description of alterations identified in oncogene-positive or -negative groups.

Supplementary Figure 9a



Supplementary Figure 9b



References

1. Pao, W. & Girard, N. New driver mutations in non-small-cell lung cancer. *Lancet Oncol* **12**, 175-80 (2011).
2. Pao, W. & Hutchinson, K.E. Chipping away at the lung cancer genome. *Nature medicine* **18**, 349-51 (2012).
3. Forbes, S.A. et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research* **39**, D945-50 (2011).
4. Wagle, N. et al. Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **29**, 3085-96 (2011).
5. Kong-Beltran, M. et al. Somatic mutations lead to an oncogenic deletion of met in lung cancer. *Cancer research* **66**, 283-9 (2006).
6. Onozato, R. et al. Activation of MET by gene amplification or by splice mutations deleting the juxtamembrane domain in primary resected lung cancers. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **4**, 5-11 (2009).

Supplementary Material: Reverse Phase Protein Array Data

METHODS

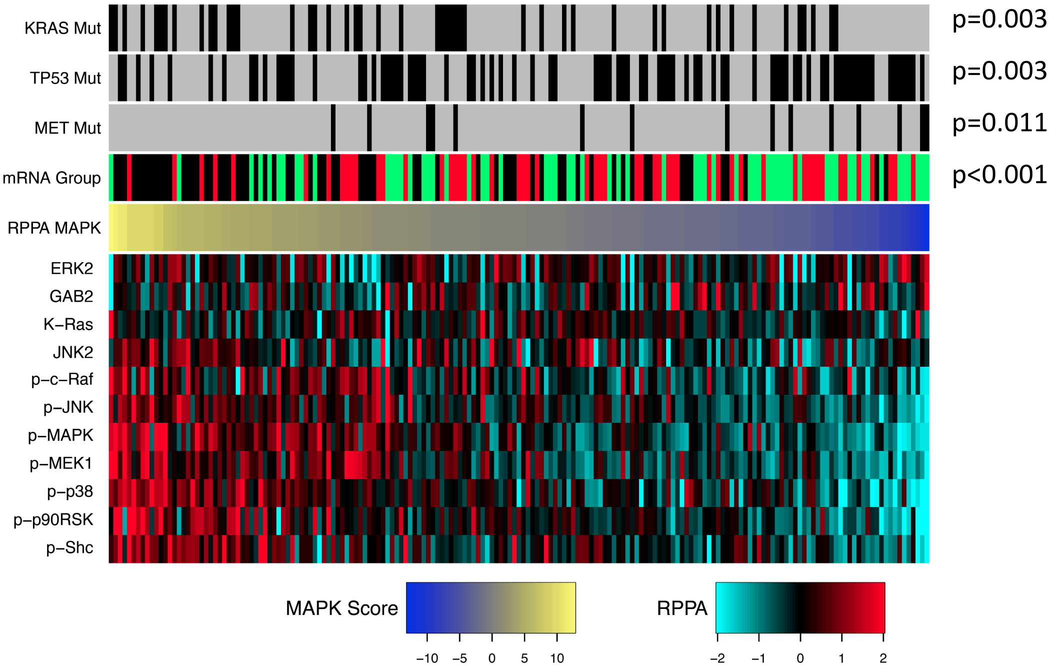
Protein lysate was prepared and analyzed by reverse phase protein array (RPPA) as previously described [1-5]. Briefly, protein was extracted using RPPA lysis buffer (1% Triton X-100, 50 nmol/L Hepes (pH 7.4), 150 nmol/L NaCl, 1.5 nmol/L MgCl2, 1 mmol/L EGTA, 100 nmol/L NaF, 10 nmol/L NaPPi, 10% glycerol, 1 nmol/L phenylmethylsulfonyl fluoride, 1 nmol/L Na3VO4, and aprotinin 10 Ag/mL) from human tumors and RPPA was performed. Lysis buffer was used to lyse frozen tumors by Precellys homogenization. Tumor lysates were adjusted to 1 µg/µL concentration and boiled with 1% SDS. Tumor lysates were manually diluted in five-fold serial dilutions with lysis buffer. An Aushon Biosystems 2470 arrayer (Burlington, MA) printed 1,056 samples on nitrocellulose-coated slides (Grace Bio-Labs). Slides were probed with 160 primary antibodies (**Supplementary Table 13**) followed by corresponding secondary antibodies (Goat anti-Rabbit IgG, Goat anti-Mouse IgG or Rabbit anti-Goat IgG). Signal was captured using a DakoCytomation-catalyzed system and DAB colorimetric reaction. Spot intensities were analyzed and quantified using Microvigene software (VigeneTech Inc., Carlisle, MA), to generate spot signal intensities (Level 1 data). The software SuperCurveGUI[3,5], available at http://bioinformatics.mdanderson.org/Software/supercurve/, was used to estimate the EC50 values of the proteins in each dilution series (in log2 scale). Briefly, a fitted curve ("supercurve") was plotted with the signal intensities on the Y-axis and the relative log2 concentration of each protein on the X-axis using the non-parametric, monotone increasing B-spline model[1]. During the process, the raw spot intensity data were adjusted to correct spatial bias before model fitting. A QC metric[5] was returned for each slide to help determine the quality of the slide: if the score is less than 0.8 on a 0-1 scale, the slide was dropped. In most cases, the staining was repeated to obtain a high quality score. If more than one slide was stained for an antibody, the slide with the highest QC score was used for analysis (Level 2 data). Protein measurements were corrected for loading as described[3,5,6] using median centering across antibodies (level 3 data). In total, 160 antibodies and 237 samples were used (183 of which were represented in the core sample set). Final selection of antibodies was also driven by the availability of high quality antibodies that consistently pass a strict validation process as previously described[7]. These antibodies are assessed for specificity, quantification and sensitivity (dynamic range) in their application for protein extracts from cultured cells or tumor tissue. Antibodies are labeled as validated and use with caution based on degree of validation by criteria previously described[7]. Raw data (level 1), SuperCurve nonparameteric model fitting on a single array (level 2), and loading corrected data (level 3) were deposited at the DCC.

Lung adenocarcinoma samples were examined for mTOR pathway signature, defined as the average of phosphoprotein levels of S6K, S6 and 4EBP1 (all proteins levels being first normalized to standard deviations from the median across tumors). Tumors were sorted into five groups: first by *PIK3CA* activating mutation (E453Q, E542K, E545K, and H1047R), then by *STK11* inactivating (nonsilent) mutation, then by high p-AKT levels (normalized levels of Akt-pS473>0.5), then by low combined LKB1/p-AMPK protein levels (average normalized levels of [LKB1+AMPK-pT172]<-0.5), and then by tumors not falling into any of the above groups. mTOR pathway signature scores for tumors in each of the first four groups were compared with those of the unaligned group (using two-sided t-test). Lung adenocarcinoma samples were then examined for activation of the MAPK pathway, calculated by taking the average of phosphorylated pathway proteins (JNK, MAPK, MEK1, p38, p90RSK, Shc, and cRaf) and total levels of ERK2, GAB2, KRAS, and JNK2. Tumors were then sorted based on KRAS mutation status and MAPK pathway protein score.

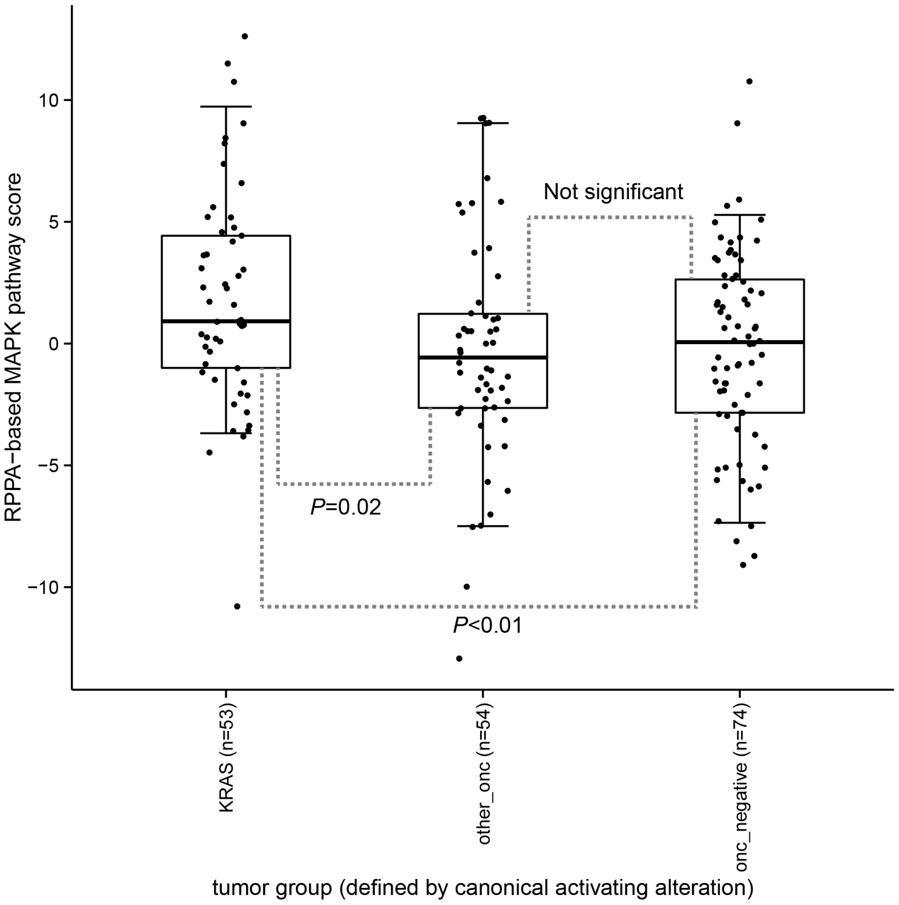
RESULTS

Supplementary Figure 10. Lung adenocarcinoma samples were examined for activation of the **(A&B)** MAPK pathway and expression of **(C)** cell cycle proteins. **(A)** A MAPK pathway protein score was calculated by taking the average of phosphorylated pathway proteins (JNK, MAPK, MEK1, p38, p90RSK, Shc, and cRaf) and total levels of ERK2, GAB2, KRAS, and JNK2. Tumors were then sorted based on MAPK pathway protein score. KRAS mutant tumors had higher MAPK pathway protein score (p=0.003 by t-test), while tumors with TP53 or Met mutations had lower MAPK protein scores (p=0.003 and 0.011, respectively). **(B)** Tumors harboring activating *KRAS* mutations are enriched for higher MAPK pathway signature score, as compared to oncogene negative tumors (P<0.01, two-sided t-test). Proteins represented in MAPK score include JNK, MAPK, MEK1, p38, p90RSK, Shc, and c-Raf. Box plots represent 5%, 25%, 75%, median, and 95%. **(C)** The proteomic cell cycle score was calculated by taking the average protein expression of cell cycle proteins shown in the figure. A cell cycle mRNA score was computed by taking the average of the normalized expression values (standard deviations from the median across samples), for genes previously found correlated with cell cycle progression[8]. mRNA subtypes and an mRNA-based cell cycle score were significantly associated with protein scores, with p19del subgroup (red) having higher cell cycle scores and T subgroup (black) having higher MAPK scores.

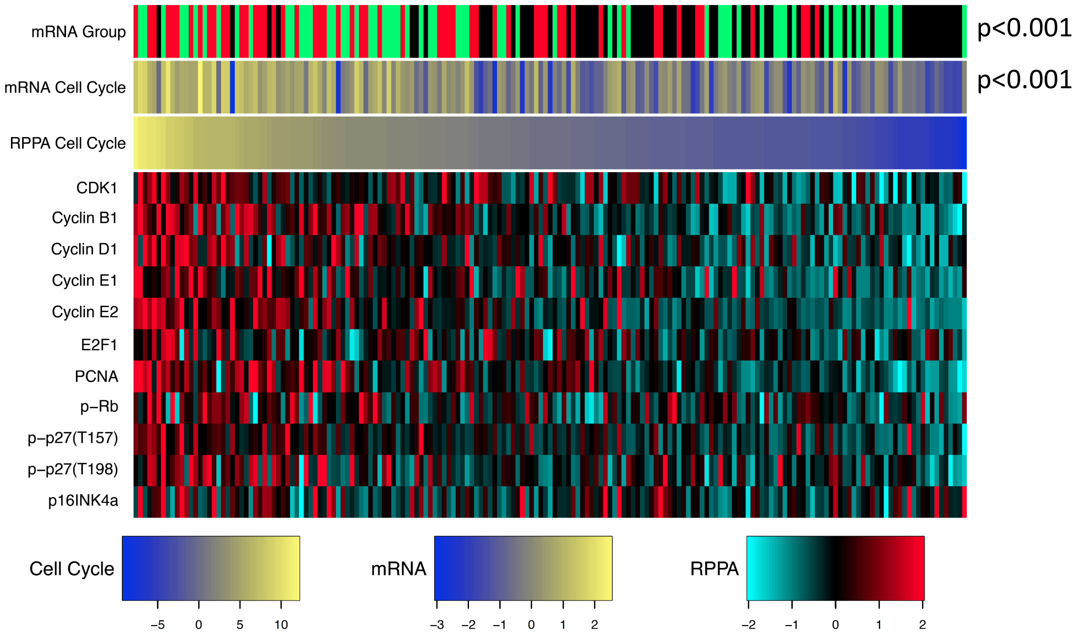
S10a



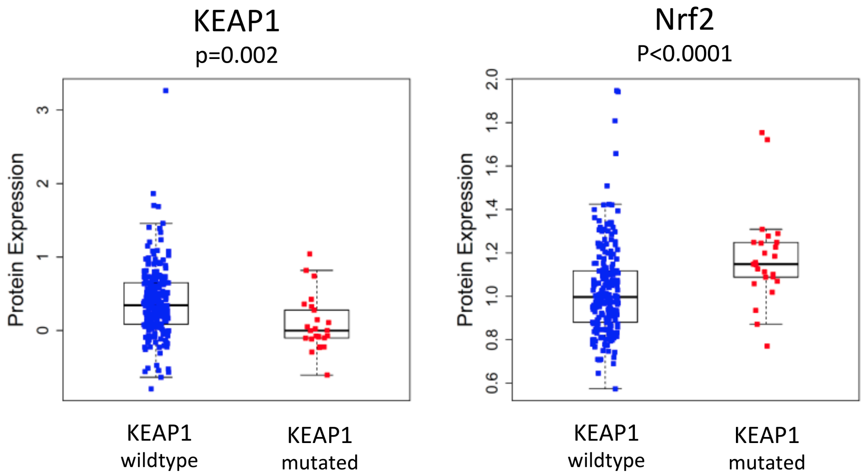
S10b


RPPA−based MAPK pathway score
10
5
0
−5
−10
Not significant
P=0.02
P<0.01
KRAS (n=53)
other_onc (n=54)
onc_negative (n=74)
tumor group (defined by canonical activating alteration)

S10c


mRNA Group
mRNA Cell Cycle
RPPA Cell Cycle
CDK1
Cyclin B1
Cyclin D1
Cyclin E1
Cyclin E2
E2F1
PCNA
p−Rb
p−p27(T157)
p−p27(T198)
p16INK4a
p<0.001
p<0.001
Cell Cycle
−5
0
5
10
mRNA
−3
−2
−1
0
1
2
RPPA
−2
−1
0
1
2

Supplementary Figure 11. *KEAP1* mutated samples had significantly lower expression of KEAP1 protein (p=0.002) and higher NRF2 protein (p<0.001), which is normally targeted for proteasomal degradation by KEAP1. Differences in protein expression between wildtype and mutated samples assessed by t-test.



REFERENCES

1. Tibes R, Qiu Y, Lu Y, et al: Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Molecular Cancer Therapeutics 5:2512-2521, 2006

2. Liang J, Shao SH, Xu Z-X, et al: The energy sensing LKB1-AMPK pathway regulates p27kip1 phosphorylation mediating the decision to enter autophagy or apoptosis. Nat Cell Biol 9:218-224, 2007

3. Hu J, He X, Baggerly KA, et al: Non-parametric quantification of protein lysate arrays. Bioinformatics 23:1986-1994, 2007

4. Hennessy BT, Lu Y, Poradosu E, et al: Pharmacodynamic Markers of Perifosine Efficacy. Clinical Cancer Research 13:7421-7431, 2007

5. Coombes K, Neeley S, Joy C, et al: SuperCurve: SuperCurve Package. R package version 1.4.1. 2011

6. Gonzalez-Angulo A, Hennessy B, Meric-Bernstam F, et al: Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. Clin Proteomics 8:11

7. Hennessy B, Lu Y, Gonzalez-Angulo A, et al: A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Non-microdissected Human Breast Cancers. Clin Proteomics 6:129-151

8. Whitfield ML, Sherlock G, Saldanha AJ, et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell. 2002 Jun;13(6):1977-2000

Supplementary Material: microRNA sequencing analysis

METHODS

MicroRNA sequence (miRNA-seq) data were generated using reported methods[1] for 230 tumor samples and 32 matched tissue normals. Unsupervised non-negative matrix factorization (NMF) consensus clustering was done as reported.[1] For the 32 miRNAs that were most discriminatory (i.e. had scores above the 95th percentile in each of the five NMF metagenes),[2,3] normalized (reads per million, RPM) abundance profiles were log$_2$-transformed and mean-centred, then were hierarchically clustered with Cluster 3.0 (bonsai.hgc.jp/~mdehoon/software/cluster/), using an absolute centered correlation and average linkage. The resulting heatmap, with tumor samples in NMF order, was visualized with Java TreeView (jtreeview.sourceforge.net/). Differentially abundant 5p/3p strands for miRBase v16 annotations were identified with two-group exact tests, using edgeR[4] v3.2.1 and R v3.0.0. The calculation was done first for the 32 matched tumor-normal pairs, then for each unsupervised tumor group, using all 230 tumors and the 32 matched normals. These calculations used read-count input matrices, TMM normalization and tagwise dispersions, and assigned Benjamini-Hochberg-corrected *P*-values. 5p and 3p strand names were assigned using miRBase v19. Purity and ploidy were reported by ABSOLUTE.[5]

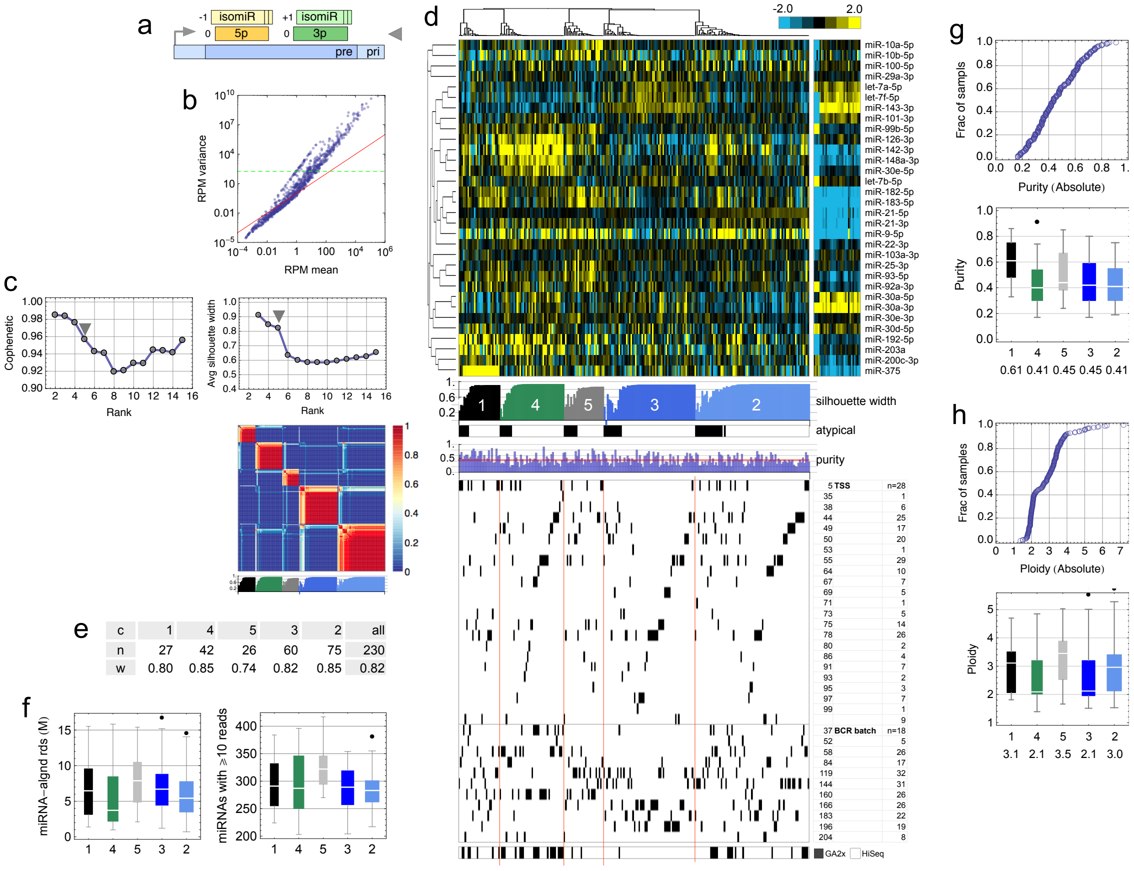
RESULTS

Cophenetic and silhouette profiles for unsupervised consensus NMF clustering suggested a five-group solution (Supplementary figure 12). Technical covariates, i.e. tissue source sites, Biospecimen Core Resource (BCR) batches, sequencing platforms, and reads aligned to miRBase annotations were consistent with the data having no strong technical biases.

miRNAs that were differentially abundant between the 32 matched tumor/adjacent normal samples were consistent with those reported in a recent meta-analysis that considered data from 598 tumor and 528 control samples across 20 studies.[6] Differentially abundant miRNAs included six of the seven upregulated miRNAs from that work (miR-21, Benjamini-Hochberg-corrected *P*=1.1e-8; miR-210, 2.7e-20; miR-182, 1.3e-12; miR-31, 5.6e-6; miR-200b, 1.2e-6; miR-205, 3.3e-4) and all eight downregulated miRNAs (miR-126-3p, *P*=3.6e-5; miR-30a, 4.0e-18; miR-30d, 1.1e04; miR-486-5p, 2.5e-22; miR-451a, 1.4e-10; miR-126-5p, 3.6e-5; miR-143, 5.2e-8; and miR-145, 1.2e-7). (Underlined miRNAs were included in the 32 most discriminatory from NMF clustering, Supplemental figure 12d.)

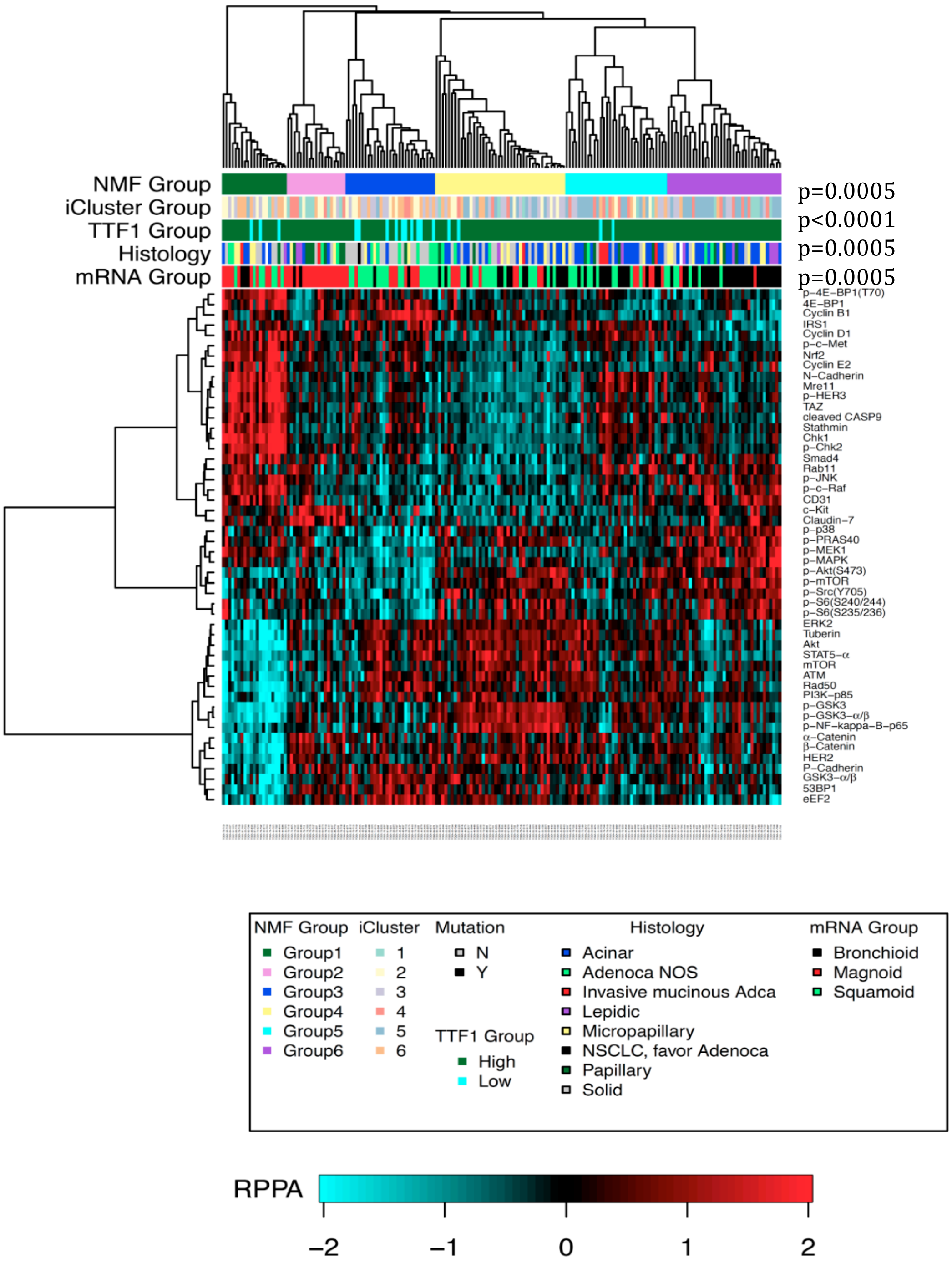
miRNAs that were differentially abundant between samples in each unsupervised group, compared to all other tumor samples and the 32 matched adjacent normal samples as a single ‘other’ group, were consistent with those reported for lung and other cancers (data not shown). These included miR-372, which was relatively abundant in group 2, and has been reported for hepatocellular[7] and gastric[8] cancers, and assessed with proteomic profiling in a lung cancer cell line.[9] miR-381 was relatively abundant in group 3, and has been reported for lung adenocarcinoma.[10] In this group, the most upregulated miRNAs include members of the miRNA-379/656 cluster that is within the imprinted *DLK-DIO3* region on chromosome 14.[11] miR-196b and miR-9 were relatively abundant in group 5. The former has been associated with proliferation and invasion in non-small cell lung cancer[12] and reported from a zonal profiling study in squamous lung cancer.[13] The latter is MYC/MYCN-activated and regulates E-cadherin.[14]

Supplementary Figure 12: Unsupervised NMF consensus clustering of miRNA-seq 5p/3p RPM abundance data for 230 tumor samples. **a)** Schematic of a miRNA primary transcript (pri), the trimmed pre-miRNA (pre), reference miRBase 5p and 3p strands, and 5' and 3' isomiR variation. The gray triangle indicates the 5p/3p-strand data representation used. **b)** Scatterplot of mean RPM vs. RPM variance, showing the expected overdispersion[4] relative to the red line, whose slope is 1. The horizontal dashed green line shows the 75th percentile of variance in normalized abundance (RPM). The input to NMF was an RPM abundance matrix for the 304 5p and 3p strands with variances above this threshold. **c)** Above: The rank survey silhouette width profile was more informative than the cophenetic correlation coefficient profile, and suggested a five-group solution. Below: Heatmap of consensus membership values for the five-group solution. **d)** Normalized $\log_2$ abundance of the 32 most discriminatory miRNAs. Beneath the heatmap is a profile of silhouette width. A sample that is a good fit in a dense, distinct group will have a high silhouette value. Beneath the profile is a track showing 'atypical' samples, defined as those with silhouette widths below 0.9 of the maximum in a group. Beneath this is a profile of purity, then tracks showing tissue source sites, BCR batches, and GAIIx and HiSeq sequencing platforms. **e)** Summary table showing group numbers, the number of tumor samples in each group, and the average silhouette width r each group. **f)** Distributions of the number of post-filter reads aligned to miRBase v16 annotations, and the number of miRNA annotations with at least 10 aligned reads. **g)** Sample purity and **h)** ploidy. Upper: distribution function. Lower: distributions in each group, with a table of median values.



c	1	4	5	3	2	all
n	27	42	26	60	75	230
w	0.80	0.85	0.74	0.82	0.85	0.82

Supplementary Figure 13. Unsupervised clustering of reverse phase protein array data by NMF clustering. NMF clustering was applied to the reverse phase protein array data, identifying 6 subgroups of tumors with distinct protein expression patterns, using the methods described in the miRNA-seq supplemental section. RPPA subgroups were significantly associated with subtypes identified in independent data types (ex., mRNA and iCluster subtypes), histology, and TTF1 expression levels (low/high) (by ANOVA). The top 50 protein markers differentially expressed between RPPA groups are shown.



REFERENCES:

1. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. *Nature* **490**, 61-70 (2012).

2. Gaujoux, R. and Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).

3. Devarajan K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput Biol*. **4**:e1000029 (2008).

4. Robinson, M.D. et al. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

5. Carter, S.L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology* **30,** 413-21 (2012).

6. Võsa, U., et al. Meta-analysis of microRNA expression in lung cancer. *International Journal of Cancer* **132**, 2884-2893 (2013).

7. Gu, H., et al. Upregulation of microRNA-372 associates with tumor progression and prognosis in hepatocellular carcinoma. *Molecular and Cellular Biochemistry* **375**, 23-30 (2013).

8. Zhou, C., et al. microRNA-372 maintains oncogene characteristics by targeting TNFAIP1 and affects NFκB signaling in human gastric carcinoma cells. *International Journal of Oncology* **42**, 635-642 (2013).

9. Lai, J.H., et al. Comparative proteomic profiling of human lung adenocarcinoma cells (CL 1-0) expressing miR-372. *Electrophoresis* **33**, 675-688 (2012).

10. Rothschild, S.I. et al. MicroRNA-381 represses ID1 and is deregulated in lung adenocarcinoma. *Journal of Thoracic Oncology* **7**, 1069-1077 (2012).

11. Glazov, E.A. et al. Origin, evolution, and biological role of miRNA cluster in DLK-DIO3 genomic region in placental mammals. *Molecular Biology and Evolution* **25**, 939-948 (2008).

12. Liu, X.H. et al. MicroRNA-196a promotes non-small cell lung cancer cell proliferation and invasion through targeting HOXA5. *BMC Cancer* **12**, 348 (2012).

13. Wu, H. et al. Tumor-microenvironment interactions studied by zonal transcriptional profiling of squamous cell lung carcinoma. *Genes Chromosomes Cancer* **52**, 250-264 (2013).

14. Ma, L. et al. miR-9, a MYC/MYCN-activated microRNA, regulates E-cadherin and cancer metastasis. *Nature Cell Biology* **12**, 247-256 (2010).

DNA Methylation Supplementary Material

METHODS

Sample preparation and hybridization

Two high-throughput DNA platforms were used for TCGA LUAD samples. The Infinium HM450 [1] assay probe set includes probes for more than 480,000 CpG sites, spanning 99% of RefSeq. In total, 96% of CpG islands and 92% of CpG shores are represented by at least one probe. This array is an expansion of the Illumina Infinium HM27 array [2], which interrogates 27,578 CpG dinucleotides spanning 14,495 unique gene regions, heavily concentrated near CpG islands. Sample preparation and hybridization protocols are identical for the two platforms, the crucial difference being that the Infinium HM27 array exclusively uses the Type-I chemistry described below, while the Infinium HM450 array employs both Type-I and Type-II chemistries for different CpG loci. [1]

Genomic DNA (1000 ng) for each sample was treated with sodium bisulfite, recovered using the Zymo EZ DNA methylation kit (Zymo Research, Irvine, CA) according to the manufacturer's specifications and eluted in 18 ul volume. An aliquot (3 ul) is removed for MethyLight-based quality control testing of bisulfite conversion completeness and the amount of bisulfite converted DNA available for the Infinium DNA Methylation assay as described in [3]. All TCGA DNA samples passed quality control and proceeded to the Infinium DNA methylation assay. Each bisulfite-converted DNA sample was whole genome amplified (WGA) followed by enzymatic fragmentation as specified by the manufacturer. The bisulfite-converted, fragmented WGA-DNA samples were then hybridized overnight to a 12 sample BeadChip. During this hybridization, the WGA-DNA molecules anneal to methylation-specific DNA oligomers linked to individual bead types, with each bead type corresponding to a specific DNA CpG site and methylation state. The oligomer probe designs follow the Infinium I and II chemistries, in which locus-specific base extension follows hybridization to a methylation-specific oligomer. There are two different bead types for each locus, one with an oligomer that anneals specifically to the methylated version of the locus, while the other oligomer anneals to the unmethylated version of the locus. The Infinium I probes terminate complementary to the interrogated CpG site for methylated loci, or complementary to the TpG for unmethylated alleles. A matched oligomer-template DNA molecule hybrid will allow for the incorporation of a labeled nucleotide immediately adjacent to the interrogated CpG (or TpG) site. However, if the probe and template are mismatched, then primer extension will not occur. Adenine and thymine nucleotides are labeled with cy5 (red), while cytosine nucleotides are labeled with cy3 (green). No insertion of guanine nucleotides occurs in Inifnium I assays. Of note, the identity of the dye is representative of the nucleotide adjacent to the CpG dinucleotide. The methylation discrimination is derived from separate measurements from the two different types of beads present for each locus. For some loci, both measurements will be cy3, and for others both will be cy5. The Infinium type II chemistry is a true two-color system. A matched oligomer-template DNA molecule hybrid will allow for the incorporation of a labeled nucleotide immediately 3' to the interrogated CpG (or TpG) site. Adenine nucleotides labeled with cy5 (red) are incorporated at unmethylated (TpG) sites, while guanine nucleotides labeled with cy3 (green) are incorporated at methylated (CpG) sites. The intensities of both cy3 and cy5 are obtained after scanning each hybridized array. BeadArrays are scanned and the raw data are imported into custom programs in R computing language for pre-processing and calculation of beta value DNA methylation scores for each probe and sample.

In addition, CDKN2A (p16) promoter methylation was measured in TCGA LUAD samples from batches 34, 37, 52 and 58 using the MethyLight assay [4] with assay primers and probe for *CDKN2A* (CDKN2A-M2; HB-081) as described previously [5].

MethyLight data are reported as a ratio between the value derived from the real-time PCR standard curve plotted as log (quantity) versus threshold C(t) value for the *CDKN2A* DNA methylation reaction and likewise for a methylation-independent control reaction (*ALU*). M.*Sss*I-treated genomic DNA is used as a reference sample to determine this ratio and to derive the standard curve. From these measurements, the Percent of Methylated Reference (PMR) is calculated as 100*(CDKN2A-M2 methylated reaction / *ALU* control reaction)$_{\text{sample}}$ / (CDKN2A-M2 methylated reaction / *ALU* control reaction)$_{\text{M.}\textit{Sss}\text{I-Reference}}$, in which the CDKN2A-M2 methylated reaction refers to the DNA methylation measurement at the *CDKN2A* promoter and the *ALU* control reaction refers to the methylation-independent measurement using a control reaction based on *ALU* repetitive elements [6].

Data Processing

For both platforms, raw image files were imported into the R (http://www.r-project.org) for pre-processing and calculation of beta value DNA methylation scores, using the methylumi Bioconductor package [7]. Pre-processing steps include background correction, dye-bias normalization, and calculation of beta values and detection p-values.

Analysis

Primary analyses including clustering to identify subtypes, comparison to other data types, and most gene specific DNA methylation estimates were confined to the clear majority of data freeze samples (181/230=79%) that were hybridized to the HM450 platform.

Clustering Analysis

DNA methylation clusters were based on CpG sites meeting the following criteria:

1) probes must be within CpG islands
2) they must be within 1500 bases of the transcription start site or in the 5' UTR of a gene
3) sample to sample variation must be in the top 1% of all probes.

Consensus clustering as implemented in the Bioconductor package ConsensusClusterPlus [8], with Euclidean distance and partitioning around medoids (pam) was used to derive clusters.

Additional Analyses

Student's t-tests were used to test for association between mutations and DNA methylation patterns and Fisher's exact test to test for associations between methylation subtype and other molecular factors including mRNA, miRNA and iCluster subtypes. Correlation between DNA methylation and expression were evaluated by Spearman correlation. Gene specific methylation for CDKN2A, RASSF1, RASAL1 and PITX1 was determined by selecting a probe for each gene on the basis of position respective of CpG islands and gene promoters as well as inverse correlation to expression.

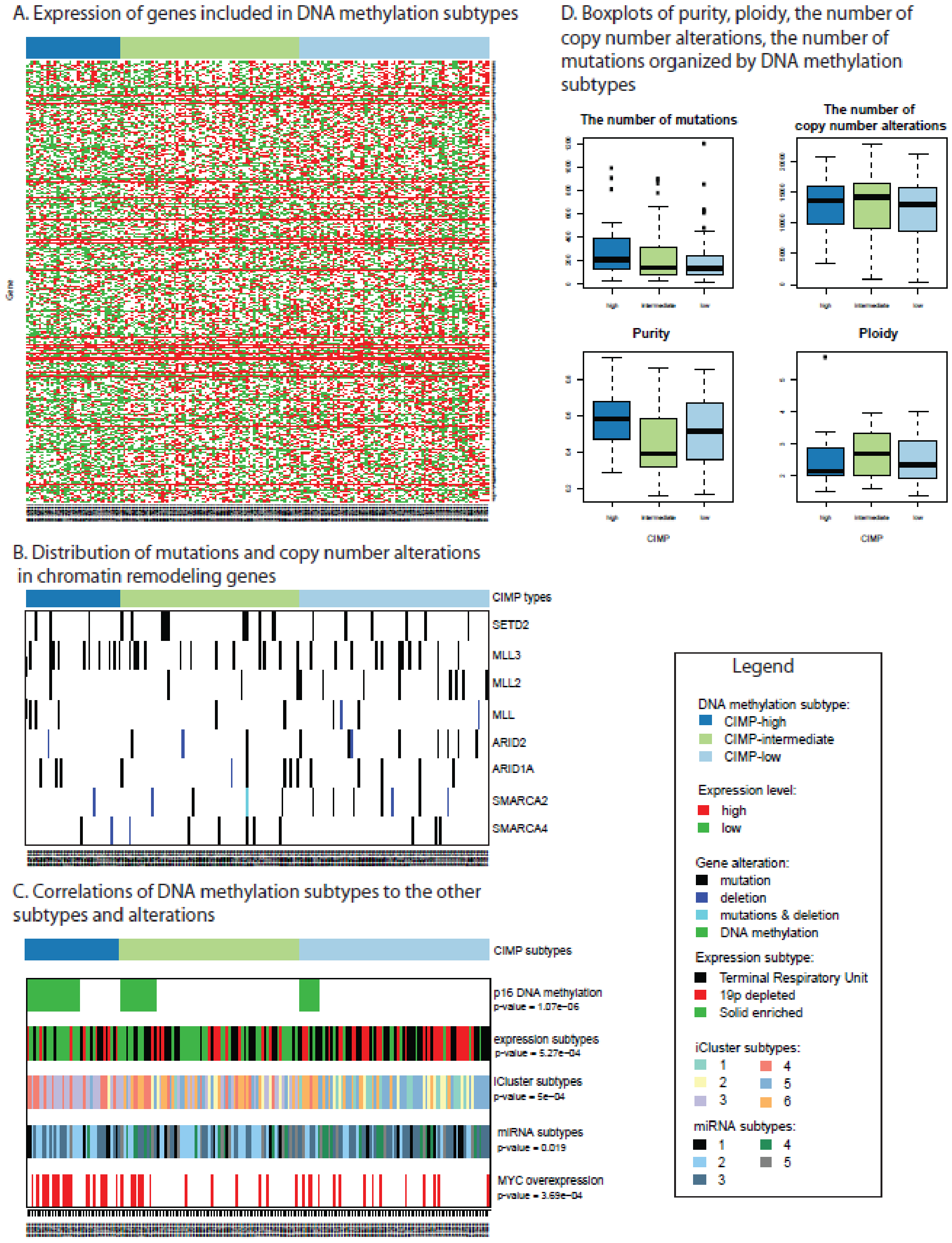
RESULTS

Correlative analysis of DNA methylation subtypes.
DNA methylation is just one of many factors regulating mRNA expression, but in the genes used to define the CIMP phenotype, expression tends to be inversely correlated to DNA methylation levels, with the lowest expression seen in CIMP-H samples (Supplementary Figure 14A).

A number of chromatin remodeling genes with frequent somatic mutations in TCGA samples were evaluated as possible drivers of the CIMP phenotype. These are shown in Supplementary Figure 14B, along with any copy number alterations identified in the same genes. No statistically significant correlations were identified.

Sample cluster assignments based on mRNA, miRNA and icluster analyses were mapped to the DNA methylation clusters, as are two individual molecular features, P16 methylation and c-MYC expression, all of which are significantly associated with CIMP. The CIMP-H samples have higher mutation counts than the other methylation groups, and tend to have slightly higher cellularity as well. Gene set analysis revealed a general increase in the expression of DNA repair genes in CIMP-H tumors compared to the rest (p-value =0.003) suggesting that the machinery is at work, if not working effectively. There do not appear to be significant differences in the number of CNAs.

Supplementary Figure 14. Correlative analysis of DNA methylation subtypes. Panel A) Expression of CIMP genes. Expression levels are shown for the highly variable CpG island probes used to define the CIMP phenotypes in Figure 4B. Three expression levels are shown for each gene, the most highly expressed 25% of samples in red, and the least highly expressed 25% in green with intermediate samples in white. **Panel B) Frequently mutated chromatin remodeling genes.** A number of chromatin remodeling genes with frequent somatic mutations in TCGA samples are shown, along with any copy number alterations identified in the same genes. **Panel C) Correlation to other subtypes.** Sample cluster assignments based on mRNA, miRNA and icluster analyses are mapped to the DNA methylation clusters for ready reference, as are two individual molecular features, P16 methylation and c-MYC expression. **Panel D) Additional correlates to CIMP-H.** The CIMP-H samples have higher mutation counts than the other methylation groups, and tend to have slightly higher purity as well. There do not appear to be significant differences in the number of CNAs.


A. Expression of genes included in DNA methylation subtypes

B. Distribution of mutations and copy number alterations in chromatin remodeling genes

C. Correlations of DNA methylation subtypes to the other subtypes and alterations

D. Boxplots of purity, ploidy, the number of copy number alterations, the number of mutations organized by DNA methylation subtypes

REFERENCES

1. Bibikova, M.; Barnes, B.; Tsan, C.; Ho, V.; Klotzle, B.; Le, J. M.; Delano, D.; Zhang, L.; Schroth, G. P.; Gunderson, K. L.; Fan, J.-B. & Shen, R. (2011), 'High density DNA methylation array with single CpG site resolution.', *Genomics* **98**(4), 288--295.
2. Bibikova, M. & Fan, J.-B. (2010), 'Genome-wide DNA methylation profiling.', *Wiley Interdiscip Rev Syst Biol Med* **2**(2), 210--223.
3. Campan, M.; Weisenberger, D. J.; Trinh, B. & Laird, P. W. (2009), 'MethyLight.',

Methods Mol Biol **507**, 325--337.

4. Weisenberger, D. J.; Siegmund, K. D.; Campan, M.; Young, J.; Long, T. I.; Faasse, M. A.; Kang, G. H.; Widschwendter, M.; Weener, D.; Buchanan, D.; Koh, H.; Simms, L.; Barker, M.; Leggett, B.; Levine, J.; Kim, M.; French, A. J.; Thibodeau, S. N.; Jass, J.; Haile, R. & Laird, P. W. (2006), 'CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer.', *Nat Genet* **38**(7), 787--793.
5. Eads, C. A.; Lord, R. V.; Wickramasinghe, K.; Long, T. I.; Kurumboor, S. K.; Bernstein, L.; Peters, J. H.; DeMeester, S. R.; DeMeester, T. R.; Skinner, K. A. & Laird, P. W. (2001), 'Epigenetic patterns in the progression of esophageal adenocarcinoma.', *Cancer Res* **61**(8), 3410--3418.
6. Weisenberger, D. J.; Campan, M.; Long, T. I.; Kim, M.; Woods, C.; Fiala, E.; Ehrlich, M. & Laird, P. W. (2005), 'Analysis of repetitive element DNA methylation by MethyLight.', *Nucleic Acids Res* **33**(21), 6823--6836.
7. Sean Davis, Pan Du, Sven Bilke, Tim Triche, Jr., Moiz Bootwalla (2012). methylumi: Handle Illumina methylation data. R package version 2.0.4
8. Matt Wilkerson (2011). ConsensusClusterPlus: ConsensusClusterPlus. R package version 1.5.1.
9. Hansen, K. D.; Timp, W.; Bravo, H. C.; Sabunciyan, S.; Langmead, B.; McDonald, O. G.; Wen, B.; Wu, H.; Liu, Y.; Diep, D.; Briem, E.; Zhang, K.; Irizarry, R. A. & Feinberg, A. P. (2011), 'Increased methylation variation in epigenetic domains across cancer types.', *Nat Genet* **43**(8), 768--775.

Supplemental material: iCluster analysis

Data processing

Data processing methods are similar to those described in [1] and are briefly described here. Copy number segmented data (germline CNV removed) based on Affymetrix SNP 6.0 array was used. Dimension reduction was performed to obtain non-redundant copy number regions as described in [1,2]. For mRNA-seq gene expression data, median absolute deviation was used to select the top 4,000 most variable genes to include for integrative clustering. For methylation data, we combined HumanMethylation27 and HumanMethylation450 platforms by taking the common probe set. Median absolute deviation was calculated on the beta values and used to select the top 4,000 most variable CpG sites to include for integrative clustering.

Integrative clustering using iCluster

Integrative clustering of DNA copy number, DNA methylation, and mRNA expression data was performed using iCluster+ [2-4]. The analysis is formulated as a joint multivariate regression of multiple data types with respect to a set of common latent variables that represent the underlying tumor subtypes. A penalized likelihood approach was used for parameter estimation. A Monte Carlo Newton–Raphson algorithm was implemented for maximizing the penalized log-likelihood.

Model selection

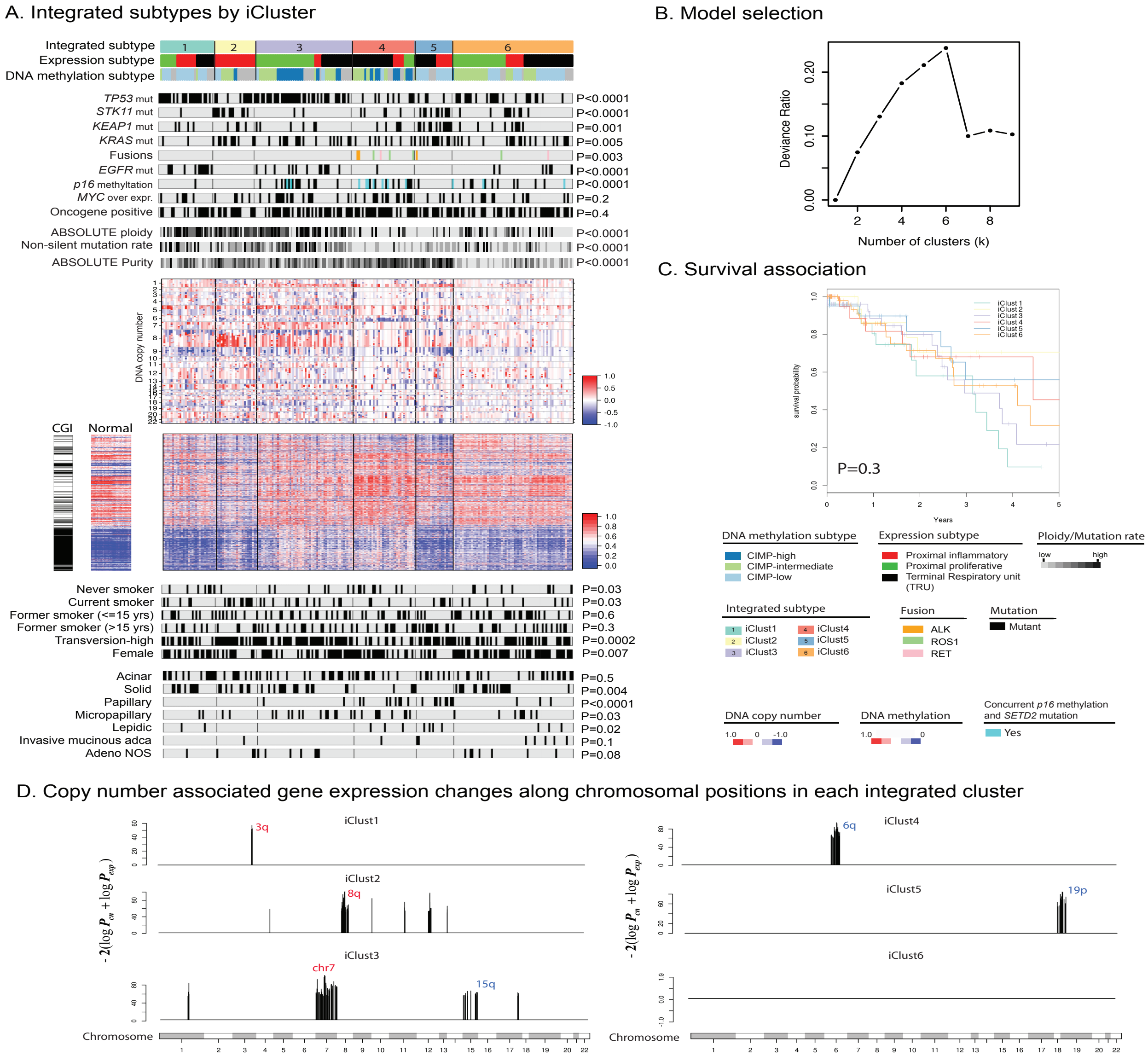
The number of clusters (K) is unknown and needs to be estimated. We compute a deviance ratio metric which can be interpreted as the percentage of variation explained by the current model, and K is chosen to maximize the deviance ratio. To determine the optimal combination of the penalty parameter values, a very large search space needs to be covered. We used an efficient sampling method that utilizes the uniform design (UD) [5]. At a given K, we determine the penalty parameter vector that minimizes a Bayesian information criterion. A theoretical advantage of the uniform design over an exhaustive grid search is the uniform space filling property that avoides wasteful computation at close-by points.

Gene-centric identification of concordant copy number and expression alterations

Following a similar method proposed in [2], we performed a gene-centric integration per each cluster to highlight the copy number associated gene expression changes. For each gene, we applied independent two-sample t-tests on its copy number and on its mRNA expression between patient samples in cluster k versus the rest. We then use Fisher's method to combine the p-values as $-2(\log P_{cn} + \log P_{exp})$ which has a χ^2 distribution under the null with 2 degrees of freedom. A large χ^2 statistic (indicated by a black verticle line along chromosomal positions in Figure S15D) provides strong evidence for concordant events at that location.

Results:

Supplementary Figure 15 – Integrative Clustering. Integrative clustering of DNA copy number, DNA methylation, and mRNA expression using iCluster+ reveals six distinct molecular subgroups among the 230 lung adenocarcinomas and highlights significant interactions between molecular subtypes (A-C). Fisher's combined probability tests highlight significant copy number associated gene expression changes on 3q in cluster 1, 8q in cluster 2, chromosome 7 and 15q in cluster 3, 6q in cluster 4, and 19p in cluster 5 (D). Cluster 6 has no concordant alterations. Labels in red indicate amplification and over-expression. Labels in blue indicate loss and under-expression.



References

1. Hammerman PS, Hayes DN, Wilkerson MD, Schultz N, Bose R, Chu A, Collisson EA, Cope L, Creighton CJ, Getz G, Herman JG, Johnson BE, Kucherlapati R, Ladanyi M, Maher CA, Robertson G, Sander C,Shen R, Sinha R, Sivachenko A, Thomas RK, Travis WD, Tsao MS, Weinstein JN, Wigle DA, Baylin SB, Govindan R, Meyerson M. (2012) Comprehensive genomic characterization of squamous cell lung cancers. Cancer Genome Atlas Research Network. *Nature.* 27:519-25

2. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, and Shen R**.** (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*. 12:4245-4250

3. Shen, R., A.B. Olshen, and M. Ladanyi, Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis*.* *Bioinformatics,* 2009. **25**(22): p. 2906-12.

4. Shen R, Wang S, Mo Q. (2012) Sparse integrative clustering of multiple omics data sets. *Annals of Applied Statistics*. 7:269-294.

5. Fang, K.-t. and Y. Wang, *Number-theoretic methods in statistics*. 1st ed. Monographs on statistics and applied probability1994, London ; New York: Chapman & Hall. xii, 340 p.

6. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. (2013) Absolute quantification of somatic DNA alterations in human cancer. Nature Biotechnology 30(5):413-21