doi:10.1038/nature11404

Supplementary Materials

Supplementary Methods S1: Biospecimen collection and clinical data:
- Table S1.1 Specimen and assay summary
- Table S1.2 Clinical data summary
- Figure S1.1 Pathology review

Supplementary Methods S2: Copy number analysis - SNP array and array CGH:
- Figure S2.1 Somatic copy number alterations
- Figure S2.2 Agilent aCGH 415K GISTIC amplifications
- Figure S2.3 Agilent aCGH 415K GISTIC deletions
- Figure S2.4 Copy number summary from Agilent 1M array
- Data File S2.1 GISTIC amplification peak annotations
- Data File S2.2 GISTIC deletion peak annotations

Supplementary Methods S3: DNA sequencing - exome and genome:
- Table S3.1 Significantly mutated genes
- Figure S3.1 Capture validation
- Figure S3.2 *KEAP1/NFE2L2* mutated samples compared to wild type
- Figure S3.3 CIRCOS plots of somatic alterations in whole genome sequencing data
- Data File S3.1 Whole genome rearrangements
- Data File S3.2 Pathways enriched with gene mutations

Supplementary Methods S4: RNA sequencing:
- Figure S4.1 RNA processing flowchart
- Figure S4.2 Gene expression subtype detection
- Figure S4.3 RNA mutation confirmation rates and mutant allele fractions
- Figure S4.4 *CDKN2A* expression analysis
- Data File S4.1 RNA fusion confirmation of DNA breakpoints
- Data File S4.2 RNA fusion predictions
- Data File S4.3 RNA annotated sequence mutation file

Supplementary Methods S5: miRNA sequencing:
- Table S5.1 Annotation priorities
- Figure S5.1 miRNA clustering

Supplementary Methods S6: DNA methylation:
- Figure S6.1 Methylation clustering
- Data File S6.1. Hypermethylated loci
- Data File S6.2. Hypomethylated loci

Supplementary Methods S7: Pathways and integrated analyses:
- Figure S7.A.1 Integrative clustering
- Figure S7.B.1 Pathway transcriptional signatures
- Figure S7.C.1 Shift flowchart
- Figure S7.C.2 Shift results for *NFE2L2* and *KEAP1*
- Figure S7.D.1 Validated and candidate therapeutic targets are commonly altered in lung SqCCs
- Figure S7.D.2 Mutations in the ERBB, FGFR, and JAK tyrosine kinases
- Figure S7.D.3 Oncoprint of therapeutically relevant alterations
- Figure S7.D.4 Pathway alterations in lung squamous cell carcinoma
- Data File S7.1 P16 Supplementary Data
- Data File S7.2: Data for Figure 5
- Data File S7.3: Data for Figure S7.D.1
- Data File S7.4: Data for Figure S7.D.1

- Data File S7.5: Clinical and genomic data table

Supplementary Methods S8: Batch effects analysis

Additional supporting material can be found at the TCGA Data Portal page for this study at:
https://tcga-data.nci.nih.gov/docs/publications/lusc_2012/

Supplementary Methods S1: Biospecimen collection and clinical data

Specimen and data acquisition

Surgically-extracted lung squamous cell carcinomas were collected following previously described TCGA collection protocols [1]. Inclusion criteria involved a minimum of tumor cellularity and confirmation of squamous cell carcinoma histology as determined by a pathology review. Samples included in this study were diagnosed as lung squamous cell carcinoma by a pathologist at the contributing center and passed specific quality control parameters for tumor cellularity and RNA quality (>60% cellularity, <20% necrosis and RIN >7. DNA and RNA were extracted and subjected to quality control measures from tissue specimens as previously described [1].

References

[1] Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature, 474(7353), 609-615. doi:10.1038/nature10166

Table S1.1 Specimen and assay summary.

Assay	Number of lung SqCC patient specimens
Exome sequencing	178 pairs
Whole genome sequencing	19 pairs
Validation capture sequencing	178 pairs
RNA sequencing	178
Gene expression microarray (Agilent 244K)	121
miRNA sequencing	158
DNA methylation (Infinium HM450)	74
DNA methylation (Infinium HM27)	104
DNA methylation (MethyLight)	178
DNA copy number (Affymetrix SNP6)	178 pairs
DNA copy number (Agilent aCGH 415K)	178 pairs
DNA copy number (Agilent aCGH 1M)	122 pairs

Table S1.2 Clinical data summary.

Age at surgery (median; range)	68.0 (40.0 – 85.0)
Gender	
Male	131
Female	47
Smoking Status	
Never smoker	7
Former smoker > 15 years	50
Former smoker ≤ 15 years	87
Current smoker	28
NA	6
Months of follow up (median; range)	15.8 (0.0 – 176.5)
Tumor stage	
I	97
II	38
III	38
IV	3
NA	2
T stage	
T1	35
T2	116
T3	14
T4	12
NA	1
N stage	
N0	117
N1	39
N2	17
N3	5

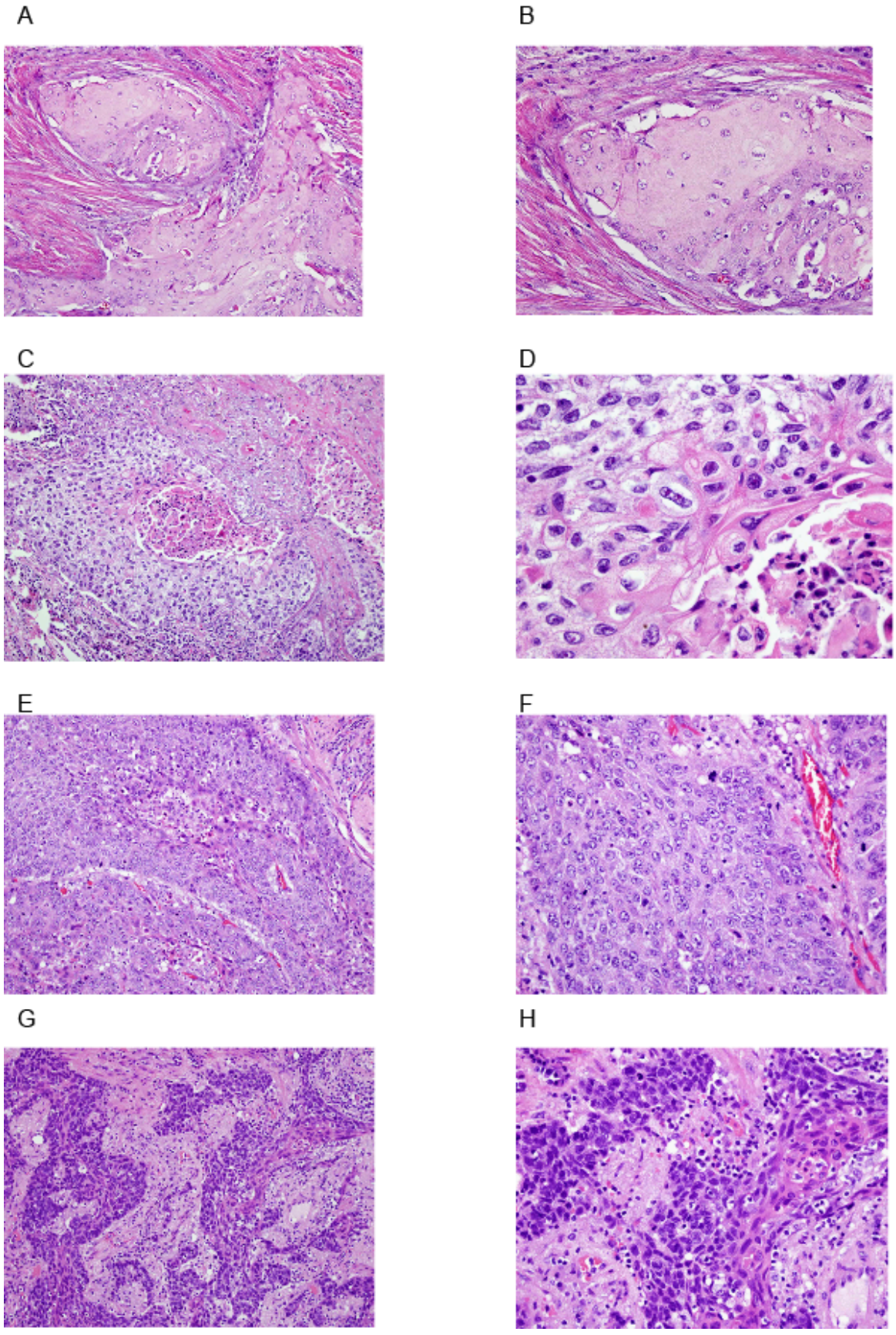
Pathology review

213 tumors were submitted as squamous cell carcinomas to the TCGA for analysis. An expert pathology committee was formed (William D. Travis, Memorial Sloan Kettering Cancer Center (MSKCC), NY, NY; Natasha Rekhtman, MSKCC, NY, NY; Joanne Yi, Mayo Clinic, Rochester MN; Marie C. Aubry, Mayo Clinic, Rochester, MN; Richard Cheney, Univ of Buffalo, Buffalo, NY; Sanja Dacic, Univ of Pittsburgh, Pittsburgh, PA; Douglas Flieder, Fox Chase Cancer Center, Philadelphia, PA; William Funkhouser, Univ North Carolina, Chapel Hill, NC; Peter Illei, Johns Hopkins Univ., Baltimore, MD; Jerome Myers, Penrose-St. Francis Health System, Colorado Springs, CO; Ming Sound Tsao, Princess Margaret Hospital, Toronto, Ontario) to review the histology of these tumors to determine whether the tumors were squamous cell carcinoma or other lung cancer histologies. Squamous cell carcinoma or probable squamous cell carcinoma cases were subclassified into keratinizing, nonkeratinizing and basaloid subtypes. Immunostaining for TTF-1, p63 and neuroendocrine markers helped reclassify some problem cases.

178 squamous cell carcinomas were separated from 17% of cases reclassified as other lung cancer histologies including adenocarcinoma (3), non-small cell carcinoma not otherwise specified (NOS) (28), large cell neuroendocrine carcinoma (3), large cell carcinoma (1), and carcinoid (1). 16% of the squamous cell carcinomas were keratinizing, 69% nonkeratinizing, and 15% were basaloid (Figure S1.1).

Figure S1.1 Pathology review.

A: Keratinizing squamous cell carcinoma. This tumor is composed of nests of cells within a dense fibrous stroma. B: Keratinizing squamous cell carcinoma. The tumor cells have abundant eosinophilic cytoplasm reflecting keratinization and squamous differentiation. Intercellular bridges are focally present. C: Focally keratinizing squamous cell carcinoma. This tumor is composed of focal areas with prominent keratinization but most of the tumor cells are small to medium in size and show little keratinization. D: Focally keratinizing squamous cell carcinoma. In the keratinized area, there is abundant eosinophilic cytoplasm and focal small pearl formation. E: Nonkeratinizing squamous cell carcinoma. This tumor is primarily composed primarily of poorly differentiated cells lacking clear keratinization, pearls or intercellular bridges which were present only very focally in this tumor. F: Nonkeratinizing squamous cell carcinoma. These tumor cells have moderate cytoplasm, lacking keratinization and the nuclei have vesicular chromatin and frequent prominent nucleoli. G: Basaloid squamous cell carcinoma. This tumor has extensive small tumor cells with scant cytoplasm and there are focal areas of keratinization (right). H: Basaloid squamous cell carcinoma. The tumor cells have a high nuclear to cytoplasmic ratio with hyperchromatic nuclei. The focal keratinized area on the right shows tumor cells with more abundant dense eosinophilic cytoplasm reflecting keratinization.


A
B
C
D
E
F
G
H

Supplementary Methods S2: Copy number analysis - SNP array and array CGH

Copy number analysis: Affymetrix SNP 6.0

From each sample of tumor or germline-derived DNA, samples were hybridized to the Affymetrix SNP 6.0 arrays using manufacturers' protocols at the Genome Analysis Platform of the Broad Institute. Data are subsequently processed from the raw CEL files using Birdseed to infer a preliminary copy-number at each probe locus. For each tumor, genome-wide copy number estimates are refined using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that is most similar to the tumor [1,2]. 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. The individual copy-number estimates then undergo segmentation using Circular Binary Segmentation [3]. As part of this process of copy-number assessment and segmentation, we remove regions corresponding to germline copy-number alterations by elimination of regions of copy-number alteration identified from either previously annotated copy-number variant filters created using the TCGA germline samples for the ovarian cancer analysis or from additional removal of regions identified in the germline genomes of 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]. Significant focal copy number alterations were identified from segmented data using GISTIC 2.0 [4] (Figure S2.1; Data Files S2.1 and S2.2).

Copy number analysis: Agilent aCGH 415K and Agilent aCGH 1M.

DNA copy number was estimated using Agilent 415K microarrays by Harvard Medical School following previously described methods [1]. GISTIC analysis of 415K copy number displayed highly similar peaks to those estimated by Affymetrix SNP6.0 (Figure S2.2 and S2.3). 26 of 33 415K amplification peaks and 35 of 41 deletion peaks were in common with the SNP6.0 peaks.

DNA copy number was estimated using Agilent 1M microarrays by Memorial Sloan Kettering Cancer Center following previously described methods [1]. Amplification and deletion peaks from this platform also showed high correspondence with Agilent 415K and Affymetrix SNP6.0 peaks (Figure S2.4).

References

1. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474:609-615 (2011).
2. Tabak B and Beroukhim R. Manuscript in preparation
3. Olshen AB, Venkatraman ES, Lucito R, and Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5:557-472 (2004).
4. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, and Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.*Genome Biol* 12: R41 (2011).

Figure S2.1 Somatic copy number alterations.

a, Comparison of SNP array copy number profiles of lung SqCCs with those of lung adenocarcinomas (both datasets from TCGA). Copy number increases (red) and decreases (blue) are plotted as a function of distance along the normal genome (vertical axis, divided into chromosomes). b, Statistically significant, focally amplified (red) and deleted (blue) regions plotted along the genome for the lung SqCCs. Annotations include all areas of significant copy number alteration identified by GISTIC with an FDR value <0.05. Peaks that include 25 or fewer genes are labeled with the number of genes in the peak in parenthesis. For peaks that contain a putative oncogene or tumor suppressor, the gene is noted.

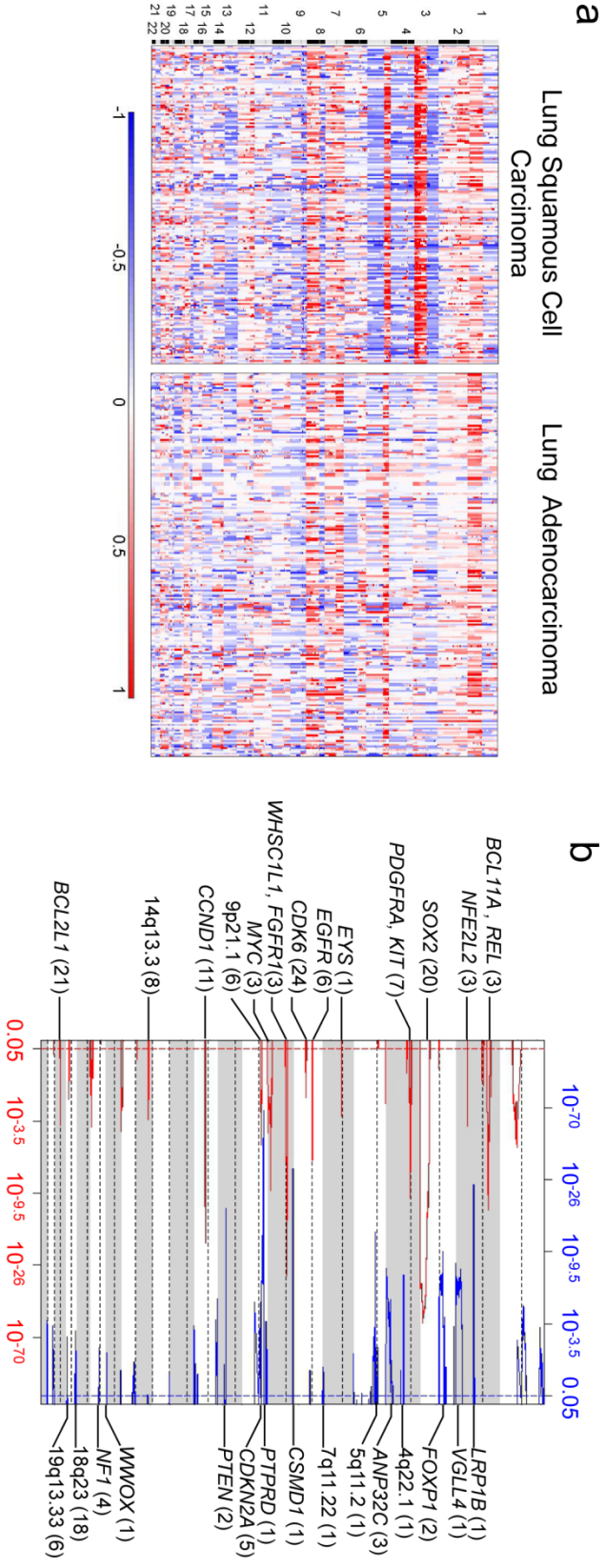

a
Lung Squamous Cell Carcinoma
Lung Adenocarcinoma
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
-1
-0.5
0
0.5
1
b
BCL11A , REL (3)
NFE2L2 (3)
SOX2 (20)
PDGFRA, KIT (7)
EYS (1)
EGFR (6)
CDK6 (24)
WHSC1L1, FGFR1(3)
MYC (3)
9p21.1 (6)
CCND1 (11)
14q13.3 (8)
BCL2L1 (21)
LRP1B (1)
VGLL4 (1)
FOXP1 (2)
4q22.1 (1)
ANP32C (3)
5q11.2 (1)
7q11.22 (1)
CSMD1 (1)
PTPRD (1)
CDKN2A (5)
PTEN (2)
WWOX (1)
NF1 (4)
18q23 (18)
19q13.33 (6)
10-70
10-26
10-9.5
10-3.5
0.05
0.05
10-3.5
10-9.5
10-26
10-70

Figure S2.2 Agilent aCGH 415K GISTIC amplifications.

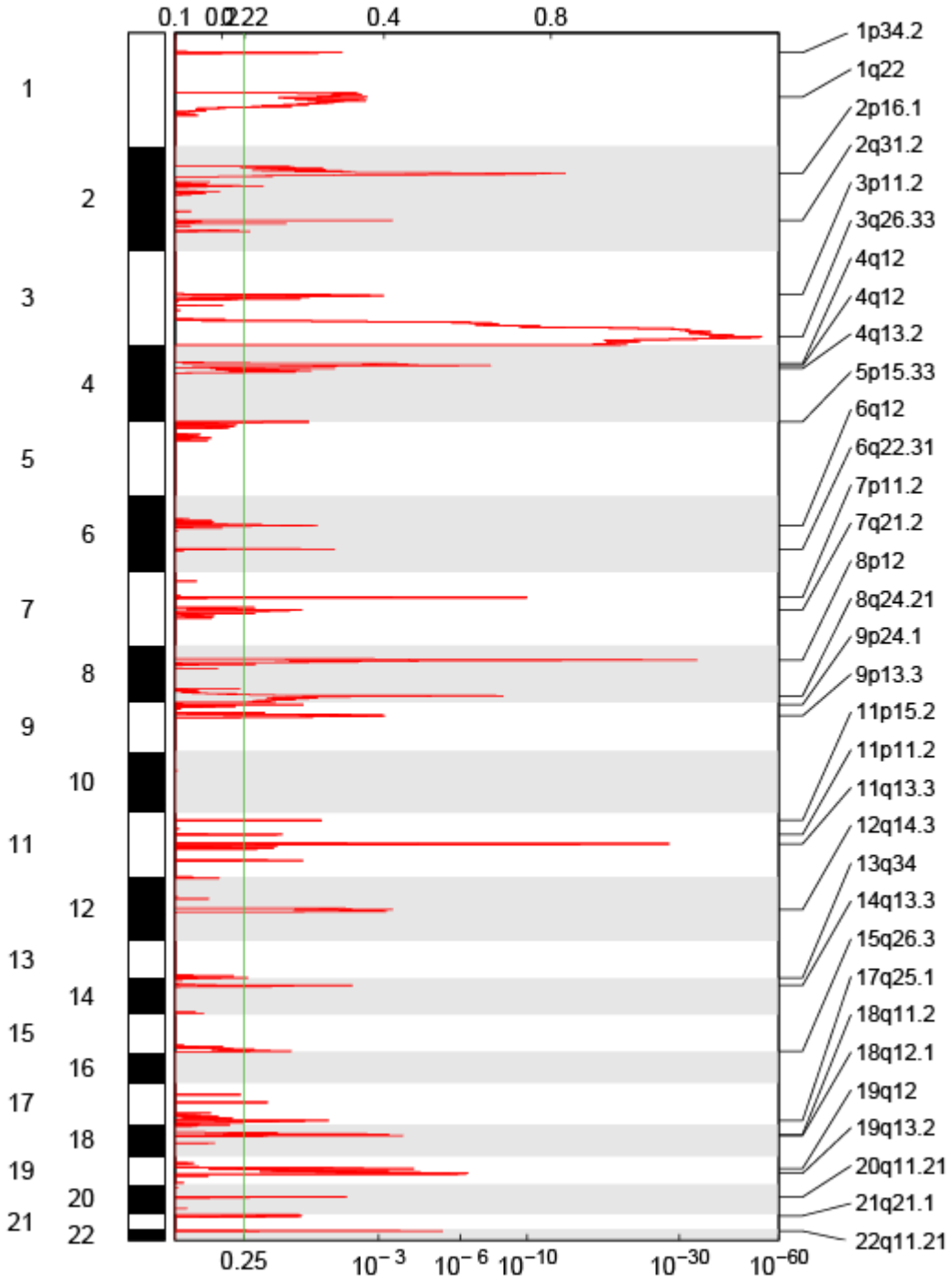


Figure S2.3 Agilent aCGH 415K GISTIC deletions.

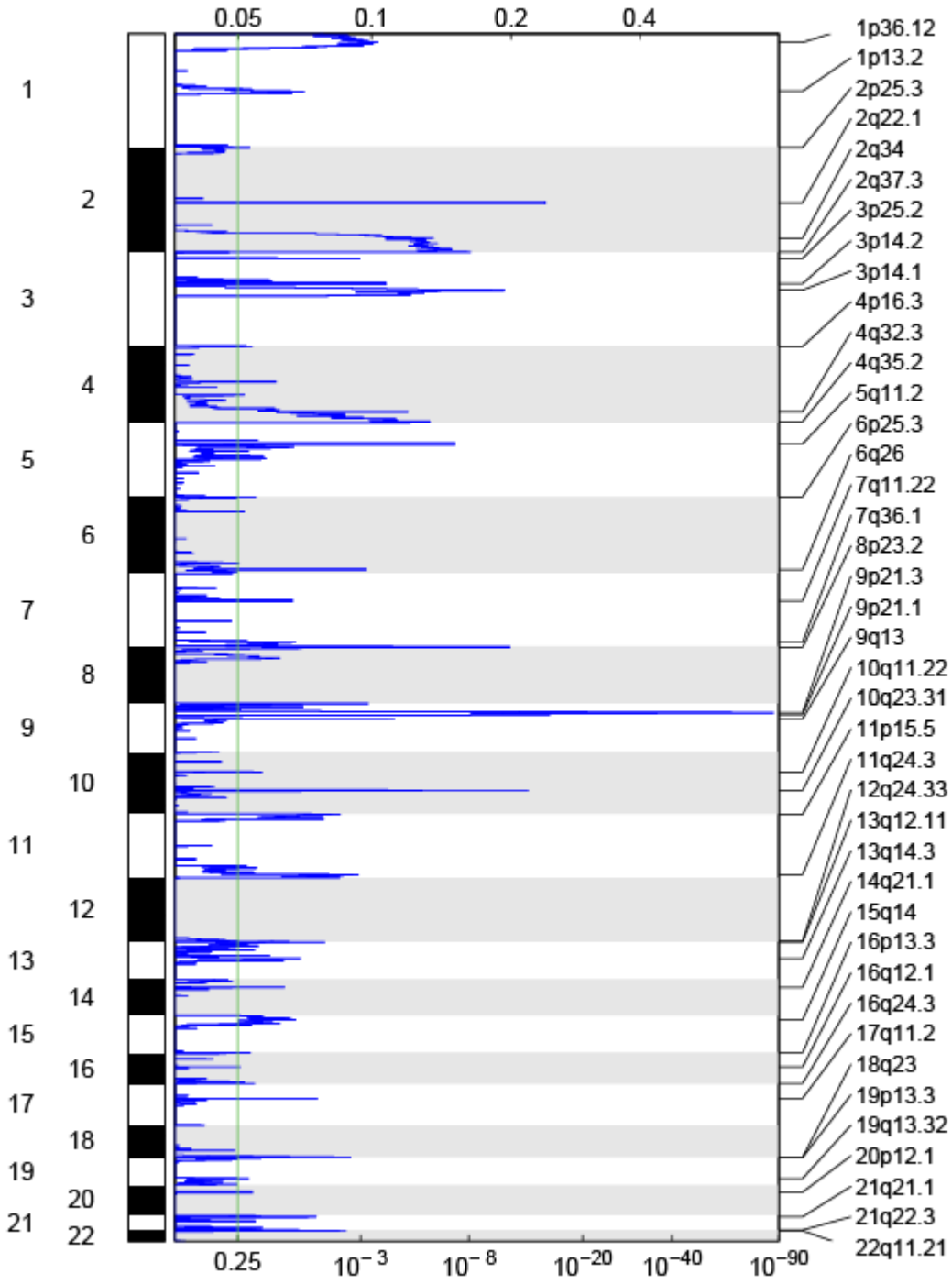
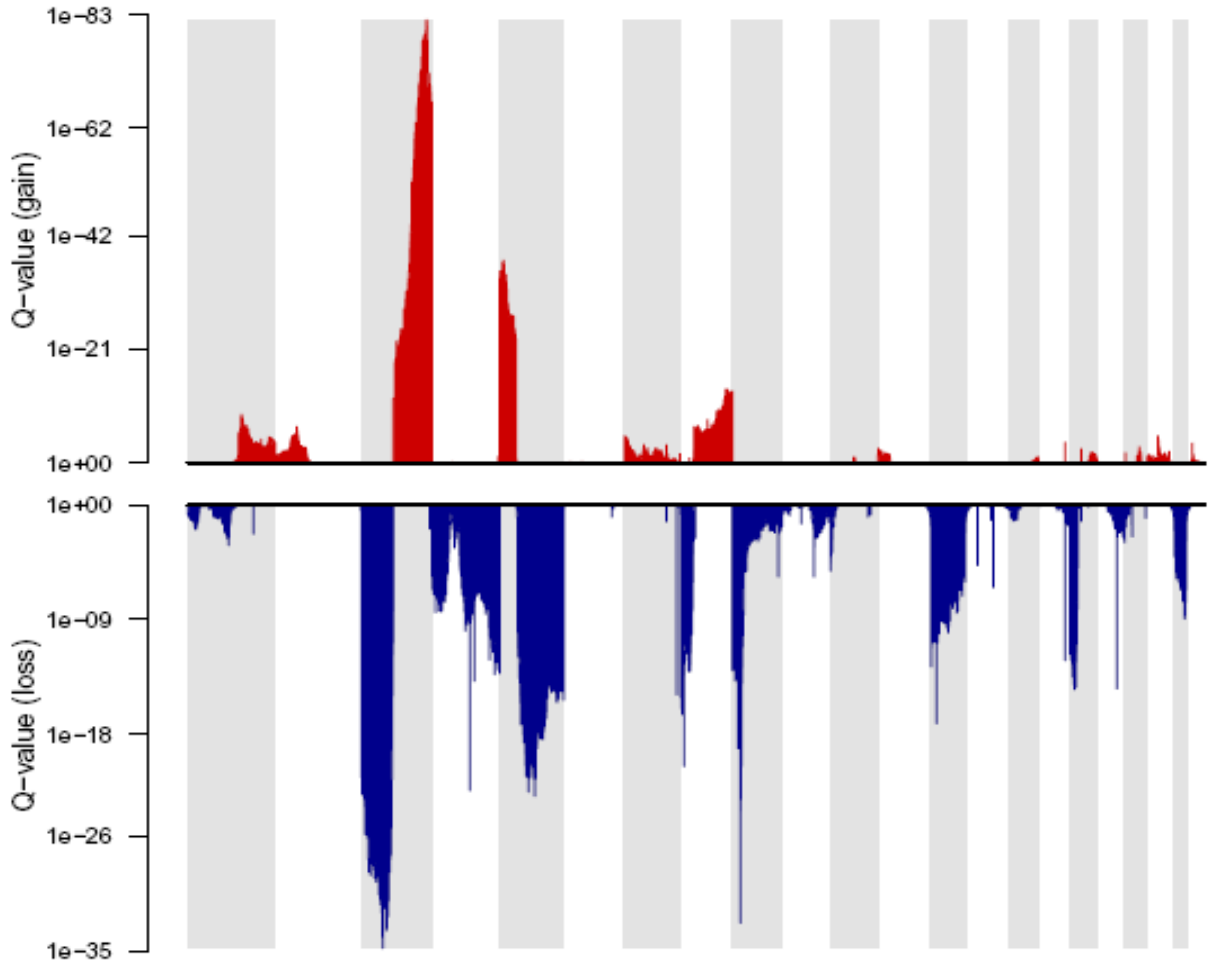


Figure S2.4 Copy number summary from Agilent 1M array. RAE analysis of gains and losses from Agilent 1M probe arrays.



Supplementary Data File S2.1: GISTIC amplification peak annotations.

Listed are regions of focal copy number amplification or deletion as determined by GISTIC [4] from the segmented copy number data obtained using the SNP 6.0 array platform. Peaks are ranked from left to right by q value with a maximum of 0.25. For each genomic region with focal copy number gain or loss the genes identified within the peak are noted.

data.file.S1.1.gistic.amplifications.txt

Supplementary Data File S2.2: GISTIC deletion peak annotations.

data.file.S2.2.gistic.deletions.txt

Supplementary Methods S3: DNA Sequencing - exome and genome

DNA sequencing and data processing

Exome capture was performed using Agilent SureSelect Human All Exon 50 Mb according to the manufacturers' instructions. Briefly, 0.5–3 micrograms of DNA from each sample were used to prepare the sequencing library through shearing of the DNA followed by ligation of sequencing adaptors. All whole exome (WES) and whole genome (WGS) sequencing was performed on the Illumina HiSeq platform.

Paired-end sequencing (2 x 101 bp for WGS and 2 x 76 bp for WE) was carried out using HiSeq sequencing instruments; the resulting data was analyzed with the current Illumina pipeline. Basic alignment and sequence QC was done on the Picard and Firehose pipelines at the Broad Institute [1]

Sequencing data were processed using two consecutive pipelines [1-3]:

(1) Sequencing data_processing pipeline – "Picard" - uses the reads and qualities produced by the Illumina software for all lanes and libraries generated for a single sample (either tumor or normal) and produces a single BAM file (http://samtools.sourceforge.net/SAM1.pdf) representing the sample. The final BAM file stores all reads and calibrated qualities along with their alignments to the genome.

(2) Cancer genome analysis pipeline – "Firehose" – takes the BAM files for the tumor and patient matched normal samples and performs analyses including qualitycontrol, local_realignment, mutation calling, small insertion and deletion identification, rearrangement detection, coverage calculations and others as described briefly below and more extensively in Stransky et al[1].

The Cancer Genome Analysis Pipeline ("Firehose")
The pipeline represents a set of tools for analyzing massively parallel sequencing data for both tumor DNA samples and their patient_matched normal DNA samples. Firehose uses GenePattern16 as its execution engine for pipelines and modules based on input files specified by Firehose. The pipeline contains the following steps [1-3].

1. Quality control – confirms identity if individual tumor and normal to avoid mix-ups between tumor and normal data for the same individual.

2. Local_realignment of reads – realigns sites potentially harboring small insertions or deletions in either the tumor or the matched normal to decrease the number of false positive single nucleotide variations caused by misaligned reads.

3. Identification of somatic single nucleotide variations (SSNVs) – Mutect algorithm – candidate SSNVs were detected using a statistical analysis of the bases and qualities in the tumor and normal BAMs.

4. Identification of somatic small insertions and deletions – Indelocator algorithm – putative somatic events were first identified within the tumor BAM file and then filtered out using the corresponding normal data.

5. Identification of inter_chromosomal and large intra_chromosomal structural rearrangements – dRanger algorithm – candidate rearrangements were identified as groups of paired-end reads which

connected genomic regions with an unexpected orientation and/or distance. When possible, breakpoints are mapped to basepair resolution using BreakPointer1.

6. Mutation rate calculation – we calculated base mutation rates using the detected mutations (SSNVs and indels) and the coverage statistics.

We observed a strong batch effect exhibited by samples with over-representation of CpG mutations associated with overall very high mutation rate. In addition, the CpG mutations in question occurred predominantly at low allelic fractions (below 0.2). Hence we employed the following filtering procedure: 1) for each sample we first calculated ratio r of CpG to non-CpG mutations as a function of allelic fraction; 2) we calculated the average, m, and standard deviation, s of all values of r *above* the 0.2 cutoff; 3) if the CpG-to-non-CpG ratio curve started above m+2*s at low AF, we removed from the data all the low-AF CpG mutations up to the point when r falls below the r+2*s. We queried all filtered mutations for validation in target DNA resequencing and in RNA sequencing data and observed very low validation rates (below 10%) further confirming that these events are artifactual.

Mutation significance analysis

Mutation significance analysis was carried out using the MutSig algorithm, as elaborated in several genome projects [1-7]. <http://confluence.broadinstitute.org/display/CGATools/MutSig> Several new refinements to the algorithm were developed in response to the unique challenges and opportunities of the squamous lung cancer sequencing project. The most important of these modifications was an improved systematic treatment of the genome-wide variation in the background mutation rate (BMR). The extremely high mutation rate of the lung tumors made it impossible to assume a uniform background rate across the genome, but it also made it possible to accurately map the structure of this variation. For very long genes, we directly estimated the local BMR from (a) synonymous mutations in the gene's coding sequence, and (b) noncoding mutations in the flanking UTR and intronic sequences, safely beyond splice sites. For shorter genes, where there was not enough data to confidently estimate the local BMR, we extended the binning approach first presented in Chapman et al., where genes were binned by estimated expression level, and then an average mutation rate was calculated for each bin, with the observation that mutation rate generally decreased with increasing expression. Here we augmented the expression data (lung squamous RNA-Seq data in our case) with other gene characteristics observed empirically to co-vary with mutation rate. These were GC content (measured on a 100Kb scale), local gene density in the gene's 1Mb neighborhood, local relative replication time [8] (and finally, open vs. closed chromatin status as estimated by "HiC" fine-scale mapping of the three-dimensional DNA contacts in the nucleus [9] Note that the last two of these covariates (replication time and chromatin state), despite not being measured in lung tissue, were nevertheless found to be excellent predictors of somatic mutation rate in these squamous lung cancer samples, consistent with these metrics being largely concordant across tumor types. We developed a general framework to encompass an arbitrary collection of such covariates. Briefly, each gene is placed in a high-dimensional covariate space, and the gene's nearest neighbors are identified. A set of nearest neighbors surrounding the gene of interest (which we term a "bagel" of genes to reflect the fact that the gene itself is not excluded and thus the set has a hole at its center) is built up around the original gene, and the local BMR is re-evaluated summing across the genes in the "bagel", gradually improving the confidence of the estimate as the amount of territory increases. A stopping criterion is imposed to balance the increased precision with the decreased accuracy (i.e. increase bias) that results from expanding outward to increasingly distant neighbors. Finally, a measurement is made of the ratio of “signal”, i.e. the number of nonsynonymous coding mutations in the gene of interest, divided by “noise” (or “background”), i.e.

the number of synonymous and noncoding mutations in the gene plus bagel. Genes with a high signal-to-noise ratio are assigned a high significance. The full details of this method will be described in a forthcoming manuscript (Lawrence et al. in preparation).

To augment our power to identify genes with potentially significant roles in cancer we performed two independent significance analyses. The first, as described above, applied a novel method to identify significantly mutated genes among all coding genes based on a model of non-uniform background mutation rate across the genome. In a second analysis, we considered only mutations in genes annotated in the COSMIC database to enrich our power to detect significantly mutated genes among genes known to be mutated in cancer.

Exome mutation validation

For validating mutations called in exome samples in another dataset (WGS, targeted DNA sequencing, or RNA sequencing data for the same sample), we undertook two alternate approaches. First, we called mutations independently in those datasets (using the same conservative settings for the mutation caller tuned for high-confidence calls as was applied to the exome data). Mutations re-called independently in another sequencing dataset for the same sample (that is, in both exome and genome sequencing, or exome and RNA sequencing) were considered validated. While extremely conservative and useful for confirming a particular mutation, this procedure is subject to a strong bias and grossly underestimates the overall validation rate. Namely, it is possible that a mutation in exome data is a true positive call that was not made in the other dataset for a variety of reasons (primarily lack of coverage, or lack of power to distinguish a true mutation from a sequencing artifact); in other words it is a *false negative* in the second dataset. To estimate a true accuracy of exome mutation calls, we considered only sites where there was enough power in the validation data to make a judgment. However, since coverage and other read alignment statistics may vary significantly both along the genome and across sequencing samples, it is difficult to accurately define the sites that are "powered" for validating (or invalidating) an exome-derived mutation call in the dataset used for validation. In order to address this problem, we used the following simple, semi-quantitative procedure. First we noted that an original call was made in a high-confidence mode, i.e. there already existed strong prior evidence for the mutation at the site. In this situation, observing a few reads carrying an alternative allele at the same site in the data used for validation may provide sufficient *additional* evidence to confirm the call, even if the validation data alone did not carry enough evidence for an *independent* call. We further confirmed that fraction of alternative alleles at the same sites in exome and validation datasets (DNA, targeted resequencing, as well as RNA sequencing) are strongly correlated. Hence, setting a fixed threshold, N, for the required number of alternative allele observations in the validation data will not introduce significant number of questionable validation calls with a strong discrepancy between allelic fractions observed in exome and validation data. Having made these observations, we 1) required at least N=2 observations of the alternative allele in the validation sample at a site of any given exome mutation call in order to consider the call to be validated, and 2) we considered only mutations at sites "powered" in validation data, defined as sites where there is at least p=0.8 probability to observe at least N=2 reads if the mutation call is true, given the allelic fraction of that mutation observed in the exome data. It is easy to see that when using this procedure even in the case when all the original mutation calls are true, the estimate of the accuracy of the original calls is only guaranteed to be above 80% (if the coverage in the validation data is such that each site just achieves p=0.8 power). Hence observed validation rates of 80% and above should be considered as indication of high quality of exome mutation calls. In reality, as it can be seen from Fig. S3.1, the validation rates determined using this power calculation are well above 90% for many samples, which suggests that most selected sites with at least p=0.8 power are actually powered more strongly (higher coverage).

References

1. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, Kryukov GV, Lawrence MS, Sougnez C, McKenna A, Shefler E, Ramos AH, Stojanov P, Carter SL, Voet D, Cortés ML, Auclair D, Berger MF, Saksena G, Guiducci C, Onofrio RC, Parkin M, Romkes M, Weissfeld JL, Seethala RR, Wang L, Rangel-Escareño C, Fernandez-Lopez JC, Hidalgo-Miranda A, Melendez-Zajgla J, Winckler W, Ardlie K, Gabriel SB, Meyerson M, Lander ES, Getz G, Golub TR, Garraway LA, Grandis JR. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011 Aug 26;333(6046):1157-60. Epub 2011 Jul 28.

2. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, Sougnez C, Voet D, Saksena G, Sivachenko A, Jing R, Parkin M, Pugh T, Verhaak RG, Stransky N, Boutin AT, Barretina J, Solit DB, Vakiani E, Shao W, Mishina Y, Warmuth M, Jimenez J, Chiang DY, Signoretti S, Kaelin WG, Spardy N, Hahn WC, Hoshida Y, Ogino S, Depinho RA, Chin L, Garraway LA, Fuchs CS, Baselga J, Tabernero J, Gabriel S, Lander ES, Getz G, Meyerson M. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 translocation. Nat Genet. 2011 Sep 4;43(10):964-8. doi: 10.1038/ng.936.

3. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, Anderson KC, Ardlie KG, Auclair D, Baker A, Bergsagel PL, Bernstein BE, Drier Y, Fonseca R, Gabriel SB, Hofmeister CC, Jagannath S, Jakubowiak AJ, Krishnan A, Levy J, Liefeld T, Lonial S, Mahan S, Mfuko B, Monti S, Perkins LM, Onofrio R, Pugh TJ, Rajkumar SV, Ramos AH, Siegel DS, Sivachenko A, Stewart AK, Trudel S, Vij R, Voet D, Winckler W, Zimmerman T, Carpten J, Trent J, Hahn WC, Garraway LA, Meyerson M, Lander ES, Getz G, Golub TR. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011 Mar 24;471(7339):467-72.

4. Cancer Genome Atlas Research Network. Comprehensive genomic charcterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061-8. Epub 2008 Sep 4.

5. Cancer Genome Atlas Research Network. Integrated genomic analysis of ovarian carcinoma. Nature. 2011 Jun 29;474(7353):609-15. doi: 10.1038/nature10166.

6. Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, Werner L, Sivachenko A, DeLuca DS, Zhang L, Zhang W, Vartanov AR, Fernandes SM, Goldstein NR, Folco EG, Cibulskis K, Tesar B, Sievers QL, Shefler E, Gabriel S, Hacohen N, Reed R, Meyerson M, Golub TR, Lander ES, Neuberg D, Brown JR, Getz G, Wu CJ. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. N Engl J Med. 2011 Dec 29;365(26):2497-506. Epub 2011 Dec 12.

7. Getz G, Höfling H, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES. Comment on “The consensus coding sequences of human breast and colorectal cancers” Science. 2007 Sep 14;317(5844):1500.

8. Chen CL, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, Audit B, d'Aubenton-Carafa Y, Arneodo A, Hyrien O, Thermes C. Impact of replication timing on CpG and non-CpG substitution rates in mammalian genomes. Genome Res. 2010 Apr;20(4):447-57. Epub 2010 Jan 26.

9. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of of long-range interactions reveals folding principles of the human genome. Science. 2009 Oct 9;326(5950):289-93.

Table S3.1 Significantly mutated genes.

Gene lists are shown, ordered by q value to a maximum of 0.1, of significantly mutated genes identified in 178 lung SCCs. The first table displays genes found to be significantly mutated when considering all genes (A), the second table only genes annotated in COSMIC (B). nsil is the number of silent mutations, nnon the number of nonsilent and nnull the number of frameshift or nonsense mutations.

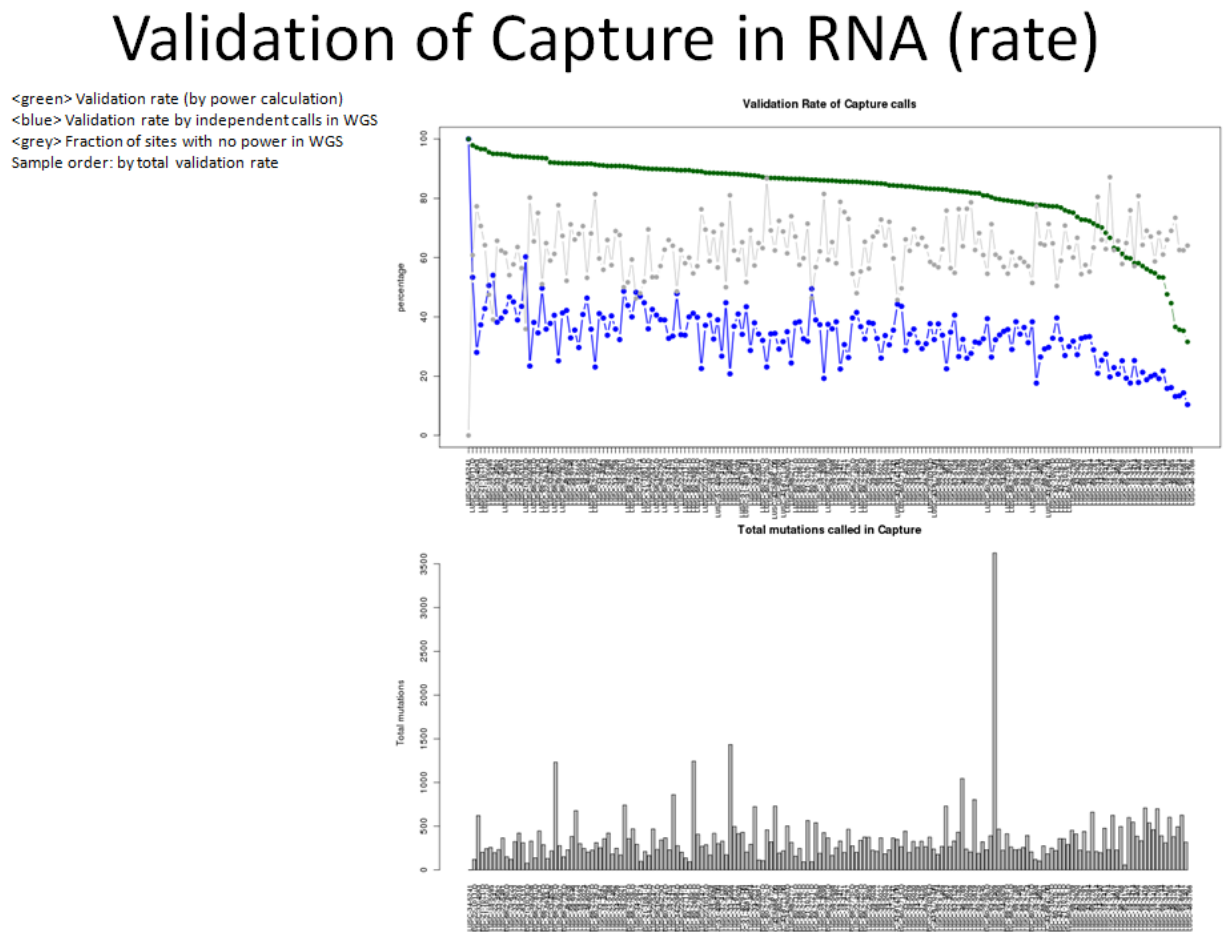
A.

name	nsil	nnon	nnull	p	q
TP53	7	148	47	3.38E-52	6.18E-48
CDKN2A	1	26	13	8.59E-33	7.84E-29
PTEN	0	16	8	1.19E-14	7.27E-11
PIK3CA	1	30	0	1.04E-13	4.73E-10
KEAP1	0	24	2	5.99E-13	2.19E-09
MLL2	7	40	18	4.46E-10	1.36E-06
HLA-A	0	7	6	9.14E-10	2.38E-06
NFE2L2	0	28	1	7.19E-07	1.64E-03
NOTCH1	3	17	8	3.29E-05	6.68E-02
RB1	0	12	8	5.07E-05	9.25E-02
PDYN	1	11	0	6.10E-05	1.01E-01

B.

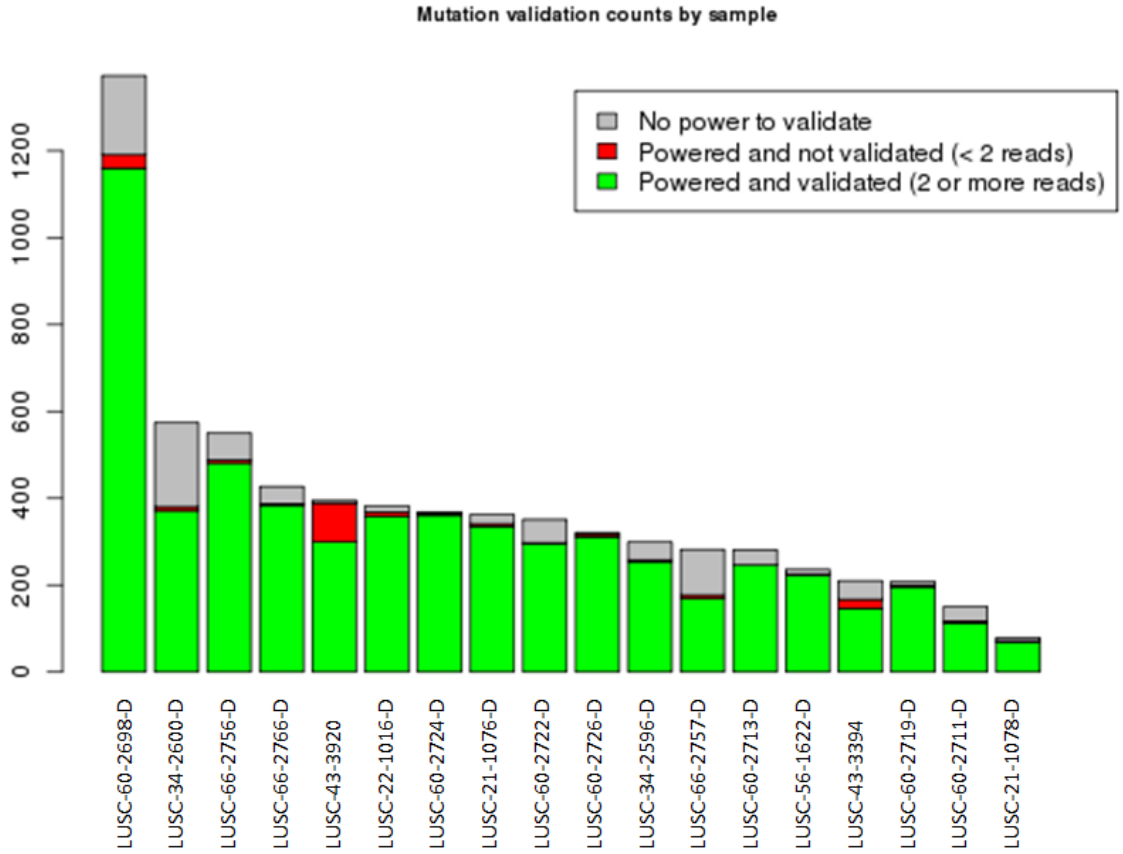
name	p	q
CDKN2A	0	0
PIK3CA	0	0
PTEN	0	0
RB1	0	0
FAM123B	0	0
HRAS	0	0
FBXW7	3.28E-237	8.47E-234
SMARCA4	1.07E-202	2.42E-199
NF1	6.87E-184	1.38E-180
SMAD4	2.98E-179	5.39E-176
EGFR	2.32E-149	3.82E-146
APC	7.21E-142	1.08E-138
TSC1	7.33E-86	1.02E-82
BRAF	1.36E-60	1.76E-57
NOTCH1	3.10E-53	3.73E-50
TP53	7.29E-23	8.24E-20
TNFAIP3	5.52E-17	5.86E-14
CREBBP	2.24E-06	2.25E-03

Figure S3.1 Capture validation.



Validation of mutations identified in exome sequencing by RNA sequencing. Samples ordered left to right by decreasing rate of validation of exome sequencing calls in RNA sequencing. The green line shows the proportion of validated mutation calls at powered sites and the gray line the fraction of unpowered sites (see text). The blue line depicts the proportion of mutations called independently (using the same parameters) in RNA sequencing. (Bottom): Number of total mutations by sample in exome sequencing ordered as in the top panel.

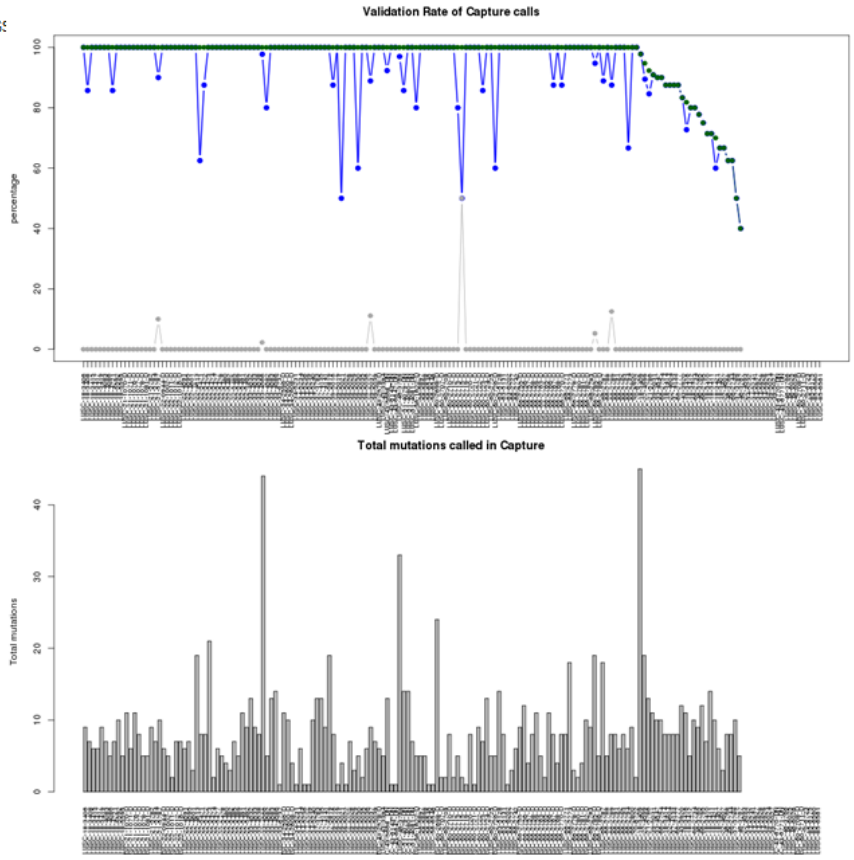
Validation of Capture in WGS (counts)



Validation of somatic mutations identified in exome sequencing by whole-genome sequencing in the same sample. Depicted are raw mutation counts (y-axis) and the number of sites at which a mutation was validated (green) or not validated with power to detect (red) or where there was insufficient power to detect (gray).

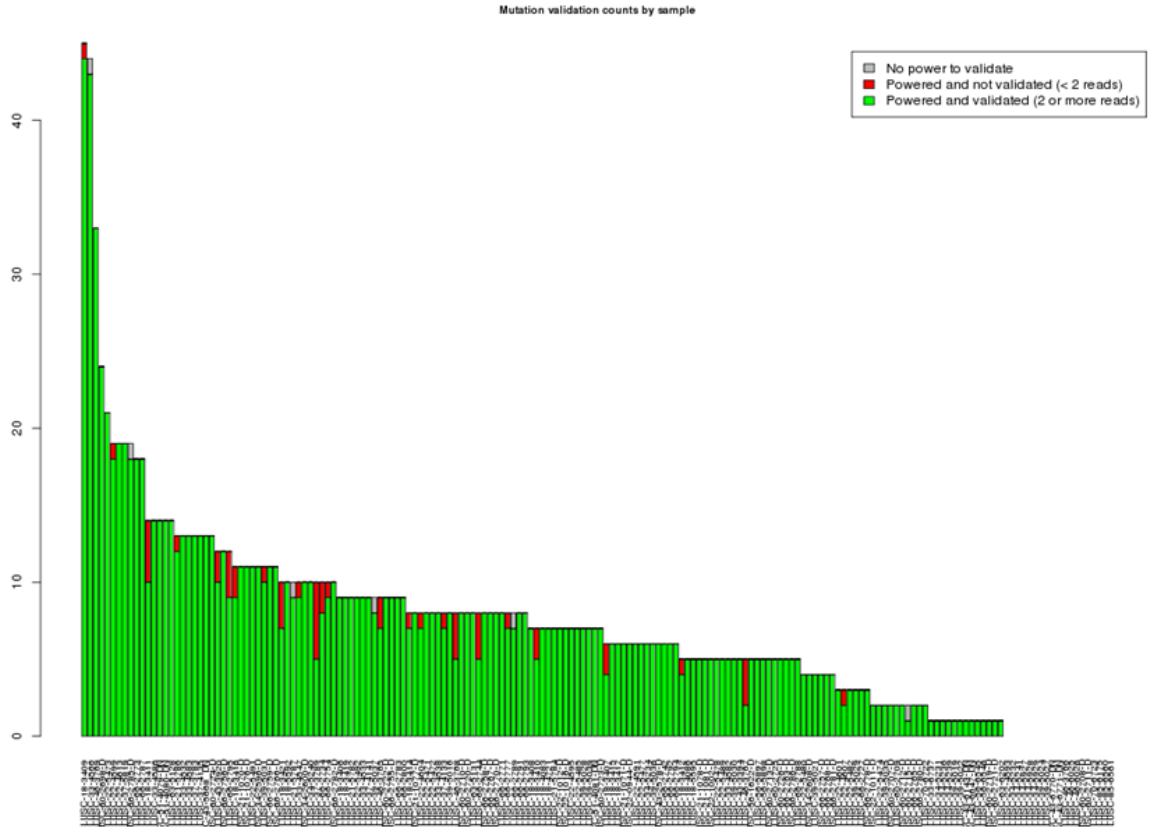
Validation of Capture in Targeted re-seq (rate)

<green> Validation rate (by power calculation)
<blue> Validation rate by independent calls in WGS
<grey> Fraction of sites with no power in WGS
Sample order: by total validation rate

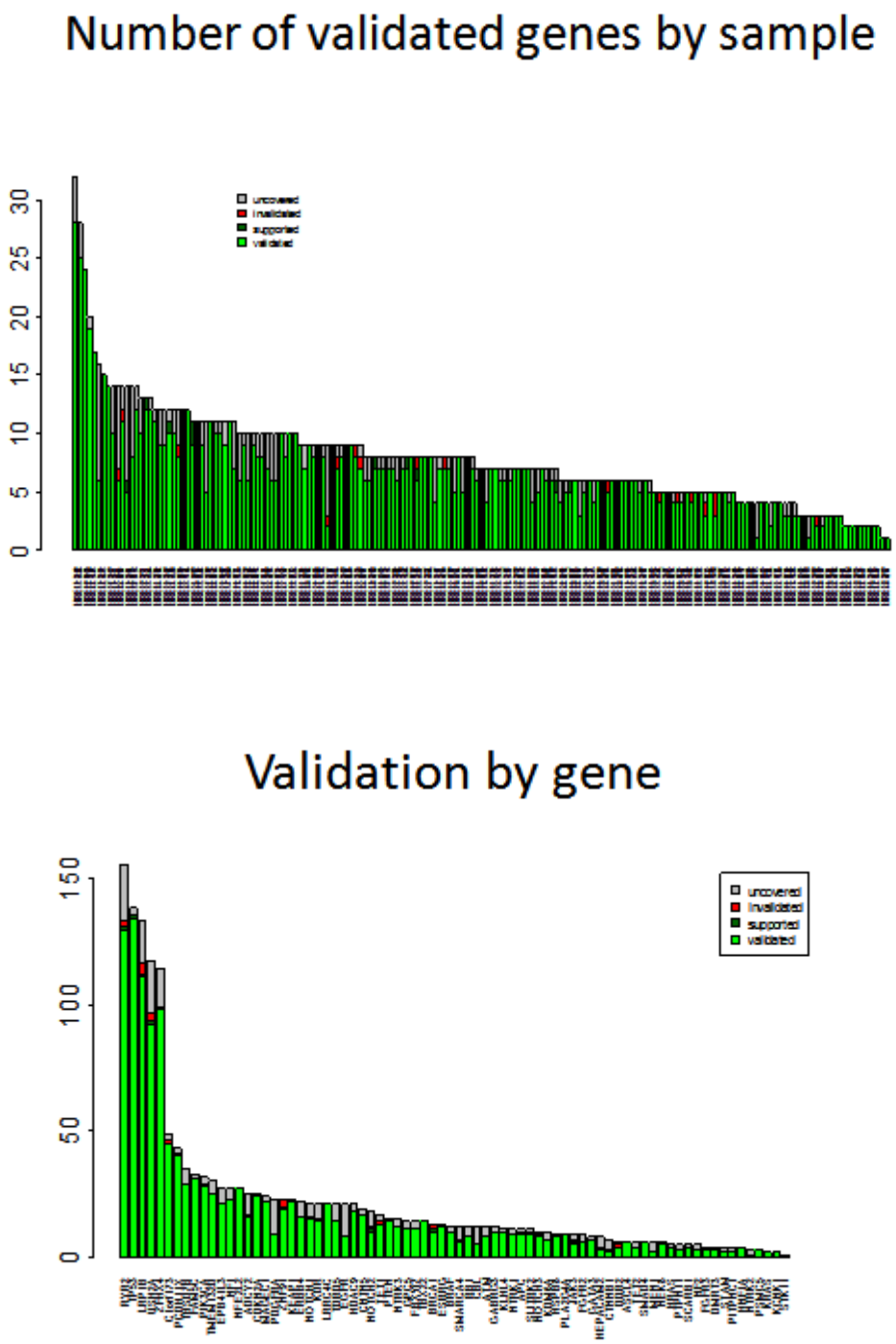


Validation of mutations identified in exome sequencing by targeted DNA re-sequencing. Samples ordered left to right by decreasing rate of validation of exome sequencing calls in targeted exome re-sequencing. The green line shows the proportion of validated mutation calls at powered sites and the gray line the fraction of unpowered sites (see text). The blue line depicts the proportion of mutations called independently (using the same parameters) in targeted re-sequencing. (Bottom): Number of total mutations by sample in exome sequencing ordered as in the top panel.

Validation of Capture in Targeted re-seq (counts)

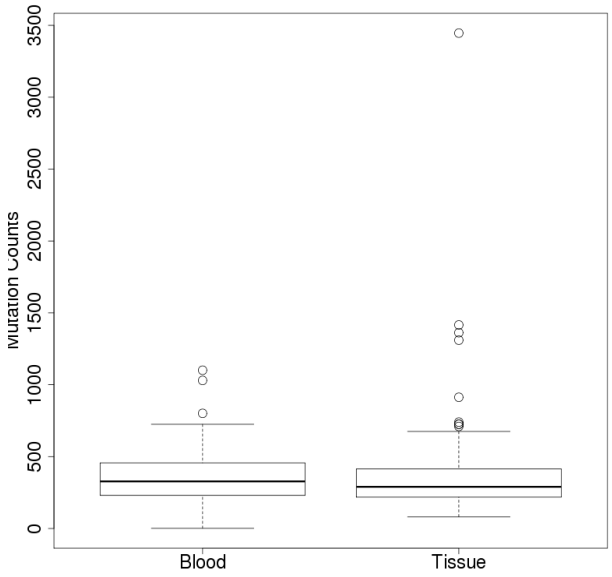


Validation of somatic mutations identified in exome sequencing by targeted re-sequencing in the same sample. Depicted are raw mutation counts (y-axis) and the number of sites at which a mutation was validated (green) or not validated with power to detect (red) or where there was insufficient power to detect (gray).



Tumors and genes exhibited high rates of validation by targeted re-sequencing across the cohort. In the top panel the number of genes in which a mutation was validated in a given sample is shown. In the bottom panel validation of mutations in each of the 76 selected genes is displayed. Validated: independently called in validation data applying the same automated calling used in the discovery set. Supported: not independently called but covered with 50 or more confirming reads in the tumor and

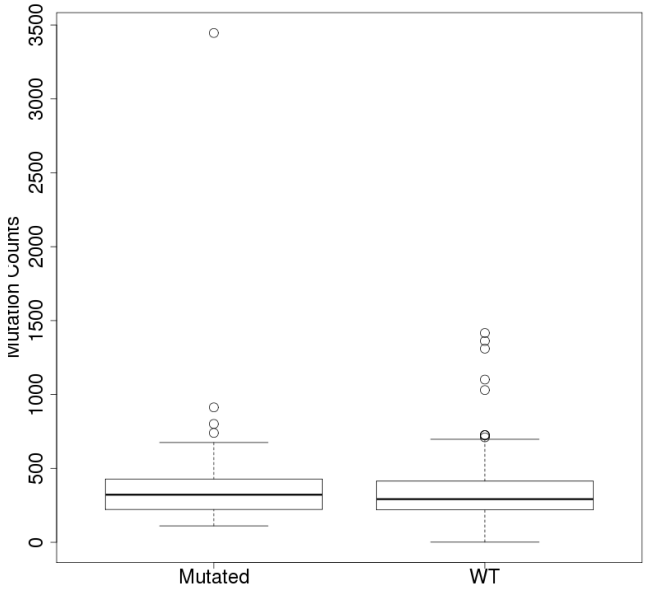
>5% of alternate reads in tumor and <3% in normal. Invalidated- adequately covered and either a different mutation called is independently or no call. Uncovered – Less than 50 reads in tumor and no mutation called.



Blood versus tissue normal mutation counts were not significantly different (*P*=0.74 by t-test).

Figure S3.2 *KEAP1/NFE2L2* mutated samples compared to wild type.

Mutation totals are not significantly different between mutated samples and wildtype (*P*=0.32 by t-test).



Mutation frequency by substitution mutation class frequency observed in a single tumor across the cohort. Mutation class rates are not appreciably different between mutant (orange) and wildtype (green) tumors.

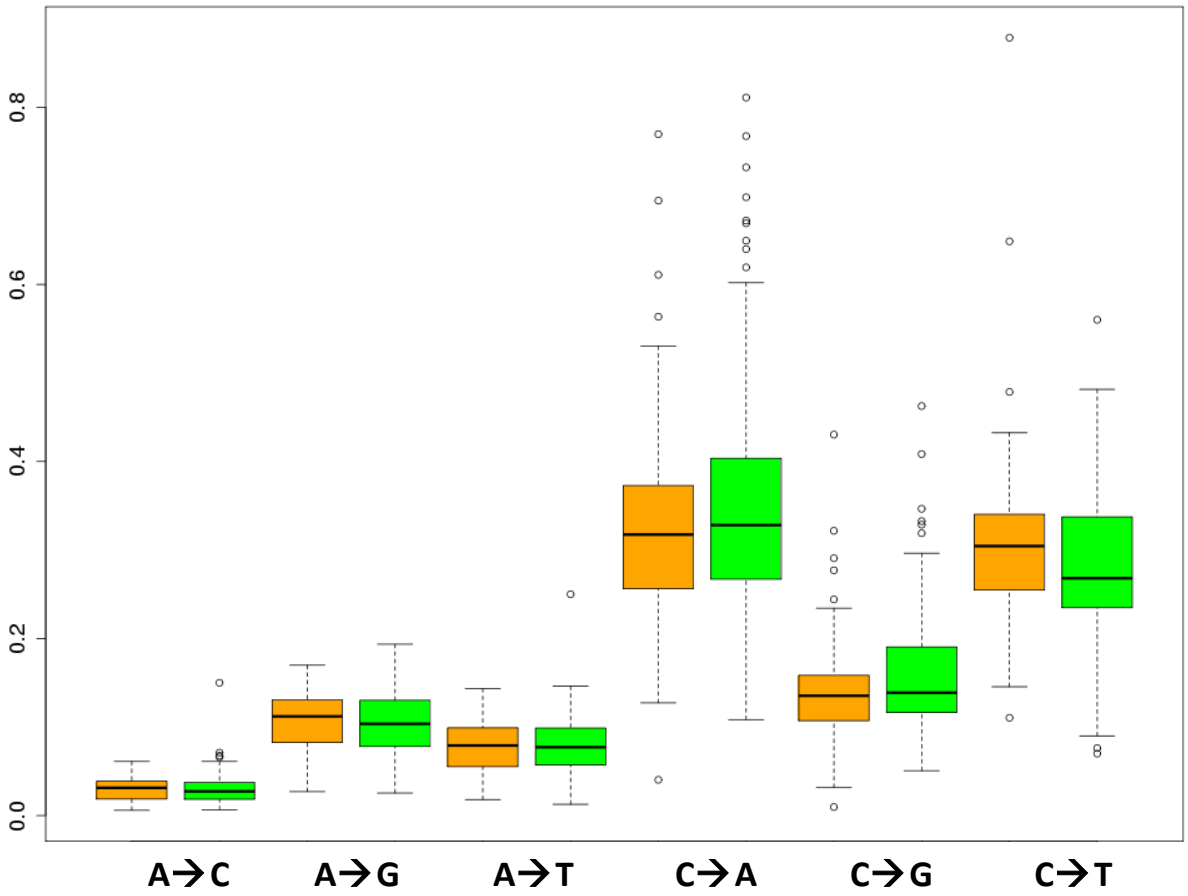
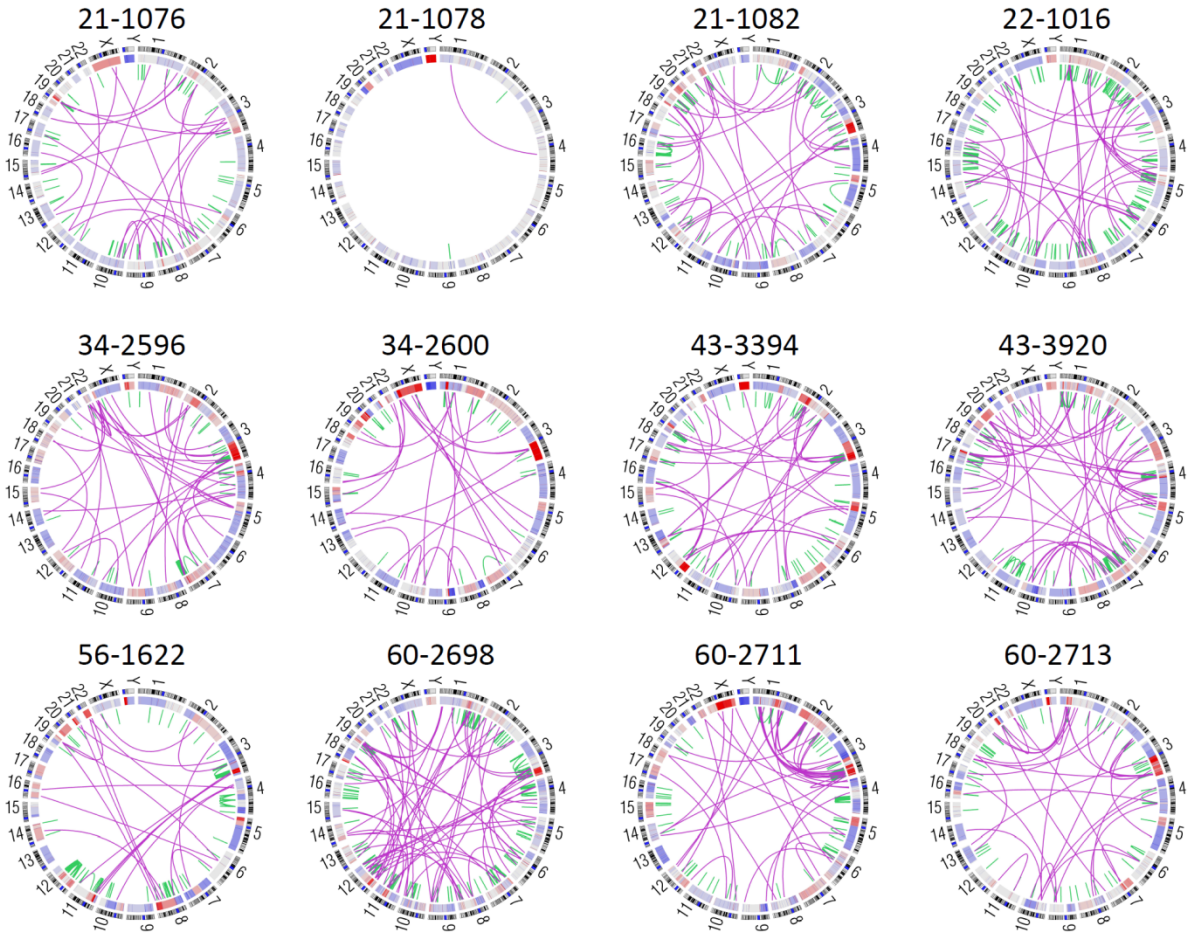
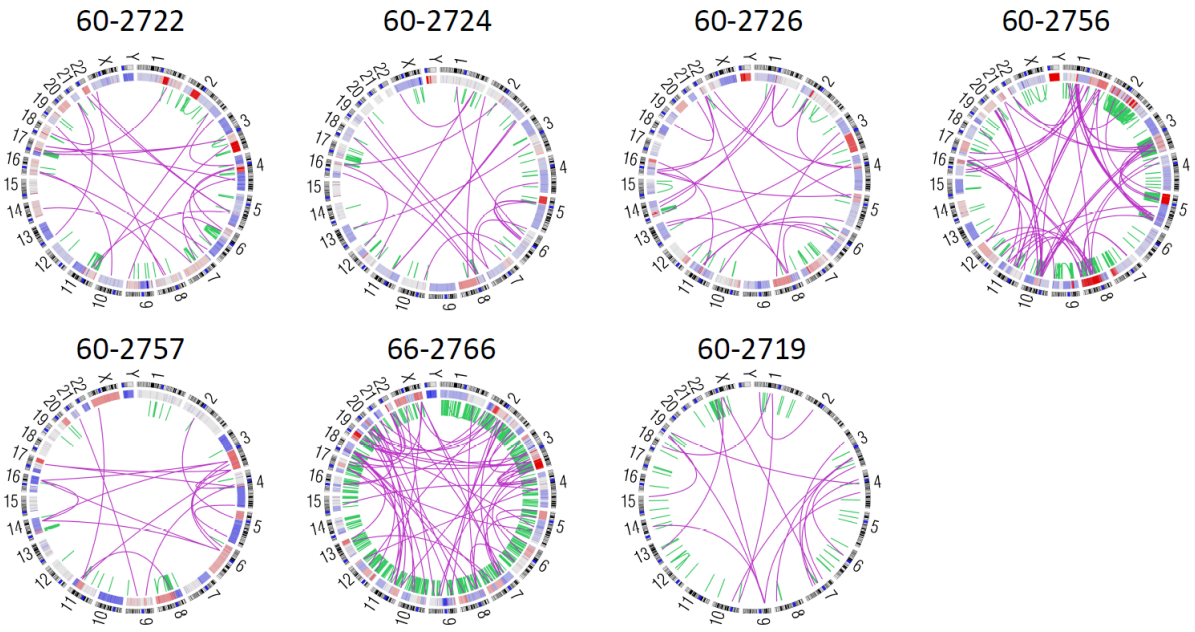


Fig S3.3 CIRCOS plots of somatic alterations in whole genome sequencing data.

DNA structural rearrangements and copy number alterations detected in the 19 lung SqCC tumors displayed as CIRCOS plots. Chromosomes are arranged circularly end-to-end with each chromosome's cytobands marked in the outer ring. The inner ring displays copy number data inferred from whole-genome sequencing with blue indicating losses and red indicating gains. Within the circle, rearrangements are shown as arcs with intrachromosomal events in green and interchromosomal translocations in purple.





Data File S3.1 Whole genome rearrangements.

data.file.S3.1.wgs.break.points.txt

Data File S3.2 Pathways enriched with gene mutations.

data.file.S3.2.pathways.enriched.with.mutations.txt

Supplementary Methods S4: RNA sequencing

RNA preparation, gene expression quantification.

RNA expression was quantified by two assays: Custom Agilent 244k microarrays and Hiseq 2000 RNAseq. For both assays, gene expression was quantified based on the gene models defined in the TCGA Gene Annotation File (GAF) [1].

Agilent 244k Microarrays: RNA was prepared, hybridized, quantified and processed to produce gene level expression estimates as previously described [2].

RNA sequencing (RNAseq): Total RNA for each sample was converted into a library of template molecules for sequencing on the Illumina Hiseq 2000 according to the protocol for the Illumina TruSeq Sample preparation kit. In brief, poly-A mRNA was purified from total RNA (1 ug) and purified using poly-T oligo-attached magnetic beads. The mRNA was then fragmented and the first strand of cDNA was synthesized from the cleaved RNA fragments using reverse transcriptase and random primers. Following the synthesis of the second strand of cDNA, end repair was performed on overhangs using T4 DNA polymerase and Klenow DNA polymerase, followed by adenylation of the 3' ends and ligation of sequencing adapters to the ends of the DNA fragments. The purified cDNA templates were enriched for 15 cycles of PCR amplification and validated using a BioAnalyzer to assess size, purity and concentration of the purified cDNA libraries.

The cDNA libraries were placed on an Illumina c-Bot for paired end cluster generation according to the protocol outlined in the Illumina Hiseq Analysis User Guide (Part# 11251649, RevA). The template cDNA libraries (1.5 µg) were hybridized to a flow cell, amplified and linearized and denatured to create a flow cell with ssDNA ready for sequencing. Each flow cell was sequenced on an Illumina Hiseq Genome Analyzer. Each sample underwent one lane of a paired end sequencing according to the protocol outlined in the Illumina Hiseq User Guide (Part# 11251649, RevA). After completion of the 50-cycle paired end sequencing run, bases and quality values were generated for each read by Bustard. The "illumina2srf" tool from the DNA Sequence Read Toolkit (http://sourceforge.net/projects/sequenceread/) was used to convert the raw sequencing data from the vendor-specific format to standard SRF (sequence read format), which is a compressed format for storing sequencing reads. Before data processing began, the "srf2fastq" tool from the Staden Package (http://staden.sourceforge.net/) was utilized to convert the data from SRF to the FASTQ format, using the "-C" parameter to simultaneously filter poor quality reads.

Spliced alignments of read sequences to the hg19 human genome assembly and fusion transcripts were calculated by the MapSplice program version 1.15.2 with the following options (-w 50, --pairend, -X 12, --fusion) [3]. This process detects RNA fusion transcripts in addition to contiguous spliced alignments, and is described in a flow chart (Fig. S4.1). WGS DNA breakpoints (Data File S3.1) were used as targets for RNA alignment and the resulting breakpoints are listed (Data File S4.1). RNA fusion predictions from Mapsplice from all tumors are listed (Data File S4.2). In parallel, lanes were assessed for a variety of pre- and post- alignment quality control measures as previously described [4].

Gene expression was quantified by counting the number of reads overlapping each gene model's exons using the Bioconductor packages GenomicFeatures and GenomicRanges. Gene read counts for each sample were converted to Reads Per Kilobase of exon model per Million mapped reads (RPKM) values by dividing by each sample's number of reads aligned in millions and by the transcribed gene length in kilobases.

Gene expression subtype detection

Previously-validated intrinsic gene expression subtypes from Wilkerson et al. [5] were detected in the TCGA cohort. These expression subtypes were named (Basal, Classical, Primitive and Secretory) based on gene expression characteristics and comparisons to lung model systems [5]. TCGA RNAseq RPKM data were converted to $\log_2$ values and were gene median centered. RPKM values of 0 (indicating there were no reads observed for a gene) were set to NA. TCGA Agilent microarray data were gene median centered. The Wilkerson et al. subtype predictor centroids [5, 6] were used to make subtype predictions for TCGA following a nearest centroid procedure. Specifically, Pearson correlations were calculated between the predictor centroids and TCGA specimens using the centroids' genes that were available on the respective platform (RNAseq: 203 genes, Agilent microarray: 207 genes). A TCGA tumor's subtype prediction was given by the centroid with the largest correlation value (Fig. S4.2).

To empirically evaluate the quality of the predictions in a manner similar to earlier studies [5, 7], the subtypes of the TCGA cohort assayed by RNAseq and by microarray were compared to the subtypes of the Wilkerson et al. cohort using the previously described subtype validation gene list (number of genes in common with validation list and both TCGA platforms: 1,091). Subtypes displayed highly concordant expression patterns between all three datasets (Wilkerson et al., TCGA RNAseq, TCGA microarray), indicating that the subtypes are robust among different patient cohorts and expression profiling technologies. The frequencies of the expression subtypes in the TCGA cohort (RNAseq: Basal 24%, Classical 37%, Primitive 15%, and Secretory 24%) were highly similar to the Wilkerson et al. cohort (Basal 21%, Classical 37%, Primitive 16%, and Secretory 26%). Between TCGA RNAseq and TCGA microarrays, subtype predictions were highly concordant, with greater than 94% of samples having the same subtype prediction. The expression subtypes in TCGA data were a statistically significant partition of gene expression space (SWISS [8] $P<0.01$, subtype versus random class labels). Taken together, these results indicate that the gene expression subtypes are highly robust across cohorts and expression platforms. Because a greater number of TCGA specimens were assayed by RNAseq than microarrays, the expression subtypes derived from RNAseq were used throughout this study.

RNA mutation confirmation and RNA mutant allele fraction

To estimate the degree at which DNA mutations might be detectable in RNA sequencing, each tumor's predicted DNA gene sequence mutations were interrogated in RNA sequencing by the program *UNCeqR* (Wilkerson et al. In preparation, http://cancer.unc.edu/nhayes/publications/unceqr/). The mutations interrogated were those predicted by exome sequencing contained in the MAF (SNP, DEL, INS, DNP, and TNP mutation types). RNA mutation confirmation consisted of having at least one read with the mutant nucleotide at the appropriate position. Across all 178 tumors and considering all exome mutations, the median tumor-wise RNA mutation confirmation rate was 45%. Considering bases were there was at least one RNA read covering the position, the median tumor-wise confirmation rate was 62% (Fig. S4.3A). In general given there was RNA read coverage indicating the region was transcribed, there was a high rate of observing the sequence mutation. This provided corroboratory evidence for the authenticity of these DNA-predicted-mutations because RNA sequencing is an independent observation and sequencing type. In addition, this process annotated which DNA mutations were expressed. The MAF file was annotated with RNA support as Data File S4.3. These analyses display the total rates of sample mutation confirmation, which was complemented by the confirmation power analysis of Fig. S3.1.

Because nonsense mediated decay could be a source of an inability to confirm DNA mutations that have nonsense protein coding changes, RNA mutant allele fractions were compared by protein coding impact.

In all DNA predicted mutations (Fig. S4.3B) and those confirmed by RNA (Fig. S4.3C), nonsense mutations have significantly reduced RNA mutant allele fraction compared to mutations with silent and missense mutations. This indicated that RNA sequencing mutation detection was affected by protein coding impact.

CDKN2A expression analysis

Expression of *CDKN2A* was evaluated for alternative transcript usage using SigFuGe. In brief, RNAseq coverage was obtained at each exonic nucleotide position of CDKN2A for all 178 samples. Principal Component Analysis (PCA) was performed and the scatter plot of the first two principal components revealed three clusters. Each sample was manually assigned to one of three clusters. The third principal component revealed some strong outlier samples, which were assigned to their own class, resulting in four total classes. Average coverage was calculated and plotted at each exonic position for the four classes (Fig. S4.4)

References

1. http://tcga-data.nci.nih.gov/docs/GAF/GAF.hg19.June2011.bundle/outputs/TCGA.hg19.June2011.gaf
2. Cancer Genome Atlas Research Network. 2011). Integrated genomic analyses of ovarian carcinoma. Nature, 474(7353), 609-615. doi:10.1038/nature10166
3. Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., et al. (2010). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Research, 38(18), e178-. doi:10.1093/nar/gkq622
4. TCGA Colorectal study, manuscript in preparation
5. Wilkerson, M. D., Yin, X., Hoadley, K. A., Liu, Y., Hayward, M. C., Cabanski, C. R., Muldrew, K., et al. (2010). Lung Squamous Cell Carcinoma mRNA Expression Subtypes Are Reproducible, Clinically Important, and Correspond to Normal Cell Types. Clinical cancer research : an official journal of the American Association for Cancer Research, 16(8), 4864-4875. doi:10.1158/1078-0432.CCR-10-0199
6. http://cancer.unc.edu/nhayes/publications/scc/wilkerson.scc.tgz
7. Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer cell. Elsevier Ltd. doi:10.1016/j.ccr.2009.12.020
8. Cabanski, C. R., Qi, Y., Yin, X., Bair, E., Hayward, M. C., Fan, C., Li, J., et al. (2010). SWISS MADE: Standardized WithIn Class Sum of Squares to evaluate methodologies and dataset elements. PloS one, 5(3), e9905. doi:10.1371/journal.pone.0009905

Figure S4.1 RNA processing flowchart.

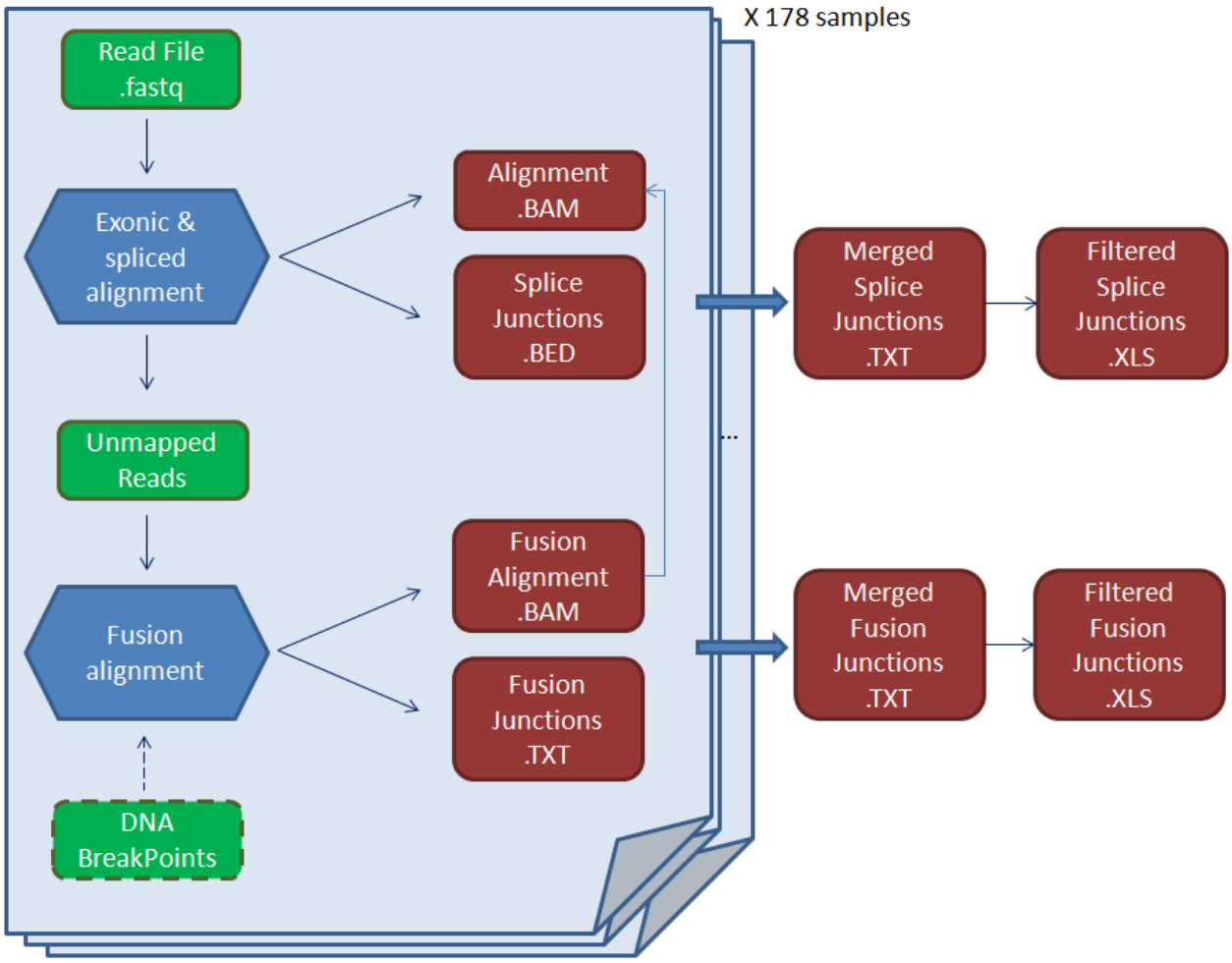


Figure S4.2 Gene expression subtype detection. Gene (mRNA) expression is displayed as heatmaps in which rows are genes and columns are tumor samples. Expression level is indicated by the shading given by the legend.

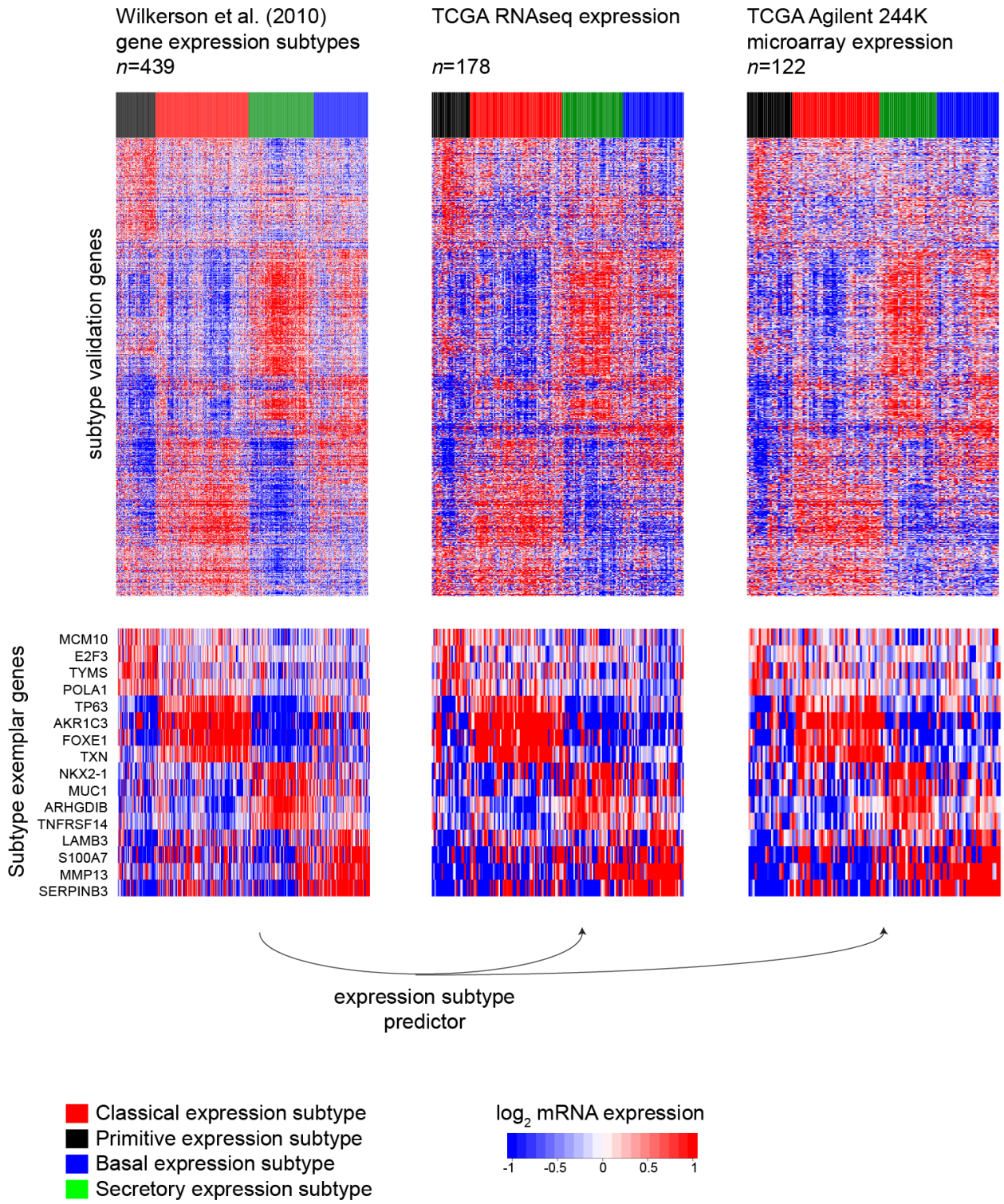


Figure S4.3. RNA mutation confirmation rates and mutant allele fractions. DNA mutations from exome sequencing were interrogated in RNA sequencing. Mutation regions with RNA coverage and with RNA confirmation (supporting the same DNA sequence mutation) are displayed by tumor sample (A). The sample-wise median percentage of mutations with coverage was 76% and confirmed was 45%. If there was any RNA coverage, the median sample-wise percentage of confirmed mutations was 62%. RNA mutation allele fractions are displayed for all DNA mutations (B) and for all DNA mutations with RNA confirmation (C) grouped by protein coding impact. Nonsense mutations had significantly reduced RNA mutant allele fraction (Kruskal-Wallis $P<0.05$).

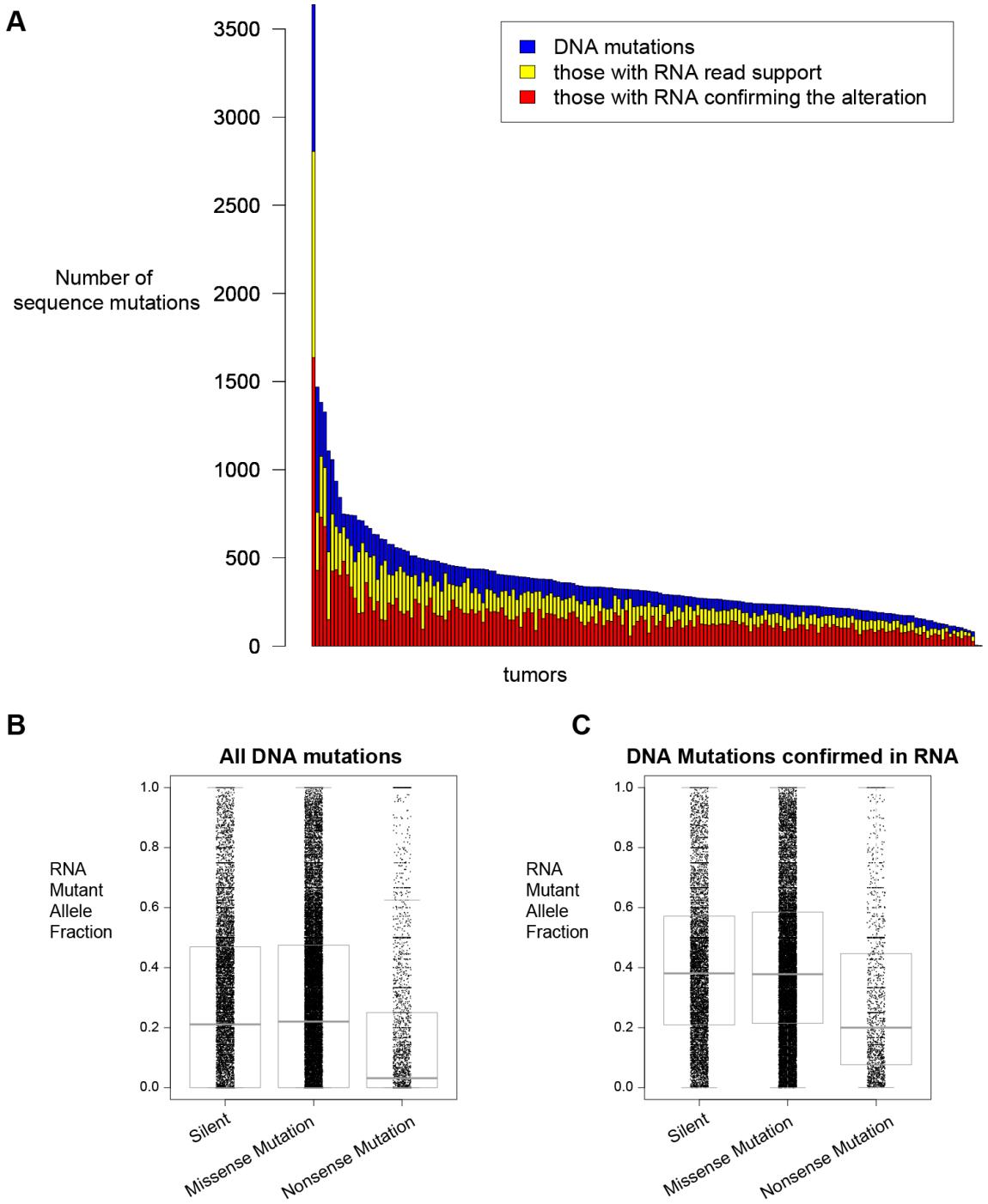
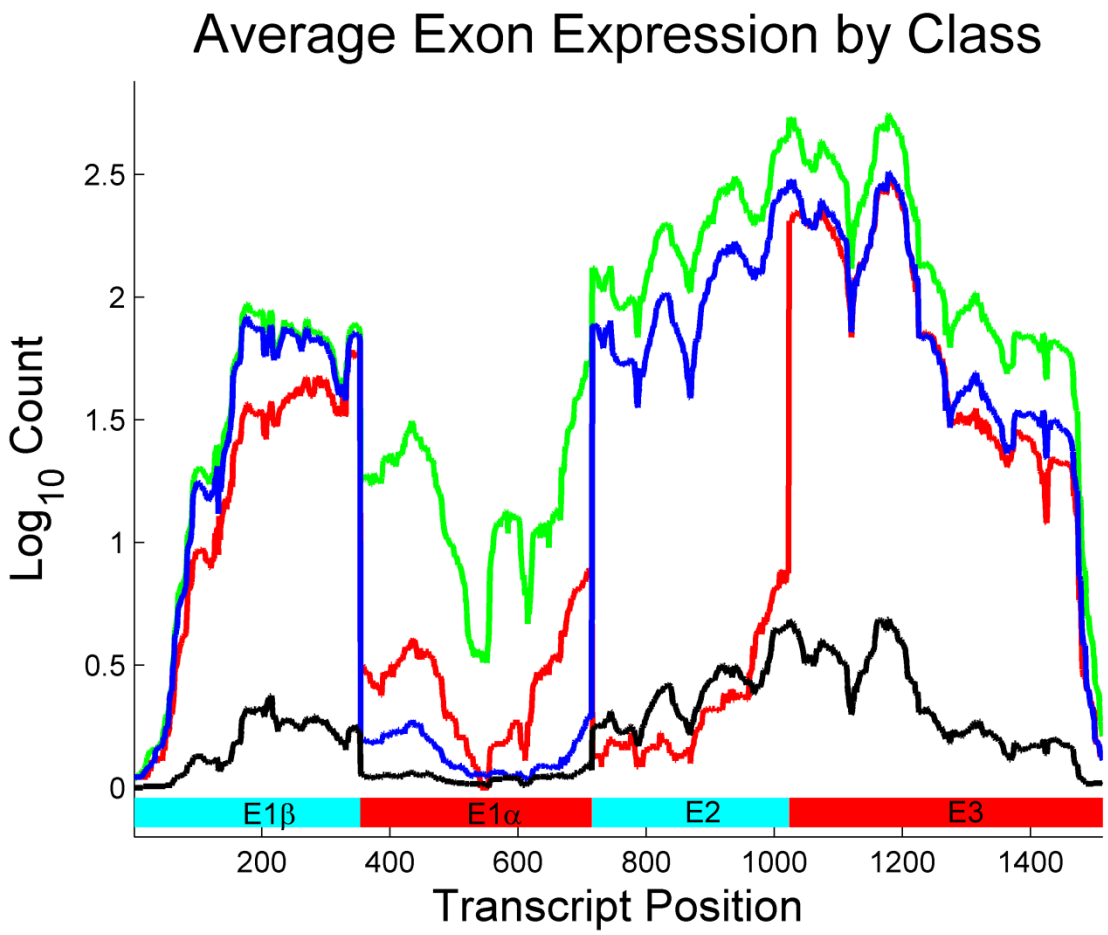


Figure S4.4 *CDKN2A* expression analysis.



Data File S4.1 RNA fusion confirmation of DNA breakpoints.

data.file.S4.1.RNA.confirmation.DNA.breakpoint.xlsx

Data File S4.2 RNA fusion predictions.

data.file.S4.2.RNA.fusion.junctions.xlsx

Data File S4.3 RNA annotated sequence mutation file.

This file is the exome mutation MAF with RNA confirmation added. 'RNAconfirm' indicates that a mutation is confirmed when it has a value of 1. 'RNA_refCnt' indicates the number of reads supporting the reference nucleotide. `RNA_varCount` indicates the number of reads supporting the tumor variant allele from exome sequencing. Positions were evaluated up to a depth of 10,000.

mafX.20120508_cleaned.csv (available at https://tcga-data.nci.nih.gov/docs/publications/lusc_2012/)

Supplementary Methods S5: miRNA sequencing

Library construction and sequencing

Two micrograms of total RNA per sample are arrayed into 96-well plates, with controls as described below. RNA entering library construction is required to have at least a minimum quality on the BCR submission documentation. Total RNA is mixed with oligo(dT) MicroBeads and loaded into a 96-well MACS column, which is then placed on a MultiMACS separator (Miltenyi Biotec, Germany). The separator's strong magnetic field allows beads to be captured during washes. From the flow-through, small RNAs, including miRNAs, are recovered by ethanol precipitation. Flow-through RNA quality is checked for a subset of 12 samples using an Agilent Bioanalyzer RNA Nano chip.

miRNA-Seq libraries are constructed using an plate-based protocol developed at the British Columbia Genome Sciences Centre (BCGSC). Negative controls are added at three stages: elution buffer is added to one well when the total RNA is 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 is 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 is then added, using a T4 RNA ligase (Ambion USA, cat. AM2141) and ATP, and is incubated at 37^{o}C for 1 hour. The sequence of the single strand RNA adapter is 5'GUUCAGAGUUCUACAGUCCGACGAUCUGGUCAA3'.

When ligation is completed, 1st strand cDNA is synthesized using Superscript II Reverse Transcriptase (Invitrogen, cat.18064 014) and RT primer (5'-CAAGCAGAAGACGGCATACGAGAT-3'). This is the template for the final library PCR, into which we introduce index sequences to enable libraries to be identified from a sequenced pool that contains multiple libraries. Briefly, a PCR brew mix is 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 is distributed evenly into a new 96-well plate. A Biomek FX (Beckman Coulter, USA) is used to transfer the PCR template (1st 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 is added to each well of the 96-well PCR brew plate. PCR is 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 is then checked across the whole plate using a Caliper LabChipGX DNA chip. PCR products are pooled, then are size selected to remove larger cDNA fragments and smaller adapter contaminants, using a 96-channel automated size selection robot that was developed at the BCGSC. After size selection, each pool is ethanol precipitated, quality checked using an Agilent Bioanalyzer DNA1000 chip and quantified using a Qubit fluorometer (Invitrogen, cat. Q32854). Each pool is then diluted to a target concentration for cluster generation and loaded into a single lane of an Illumina GAIIx or HiSeq 2000 flow cell. Clusters are generated, and lanes are sequenced with a 31-bp main read for the insert and a 7-bp read for the index.

Preprocessing, alignment and annotation

Briefly, the sequence data are separated into individual samples based on the index read sequences, and the reads undergo an initial QC assessment. Adapter sequence is then trimmed off, and the trimmed reads for each sample are aligned to the NCBI GRCh37-lite reference genome. Below we describe these steps in more detail.

Routine QC assesses a subset of raw sequences from each pooled lane for the abundance of reads from each indexed sample in the pool, the proportion of reads that possibly originate from adapter dimers (i.e. a 5' adapter joined to a 3' adapter with no intervening biological sequence) and for the proportion of reads that map to human miRNAs. Sequencing error is estimated by a method originally developed for SAGE (Khattra 2007).

Libraries that pass this QC stage are preprocessed for alignment. While the size-selected miRNAs vary somewhat in length, typically they are ~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 then 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 happens to match 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 was 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 containing the number of reads at each read length. Because the shortest mature miRNA in miRBase v16 is 15 bp, any trimmed read that is shorter than 15bp is discarded; remaining reads are submitted for alignment to the reference genome.

Alignment(s) for each read are checked with a series of three filters. A read with more than 3 alignments is discarded as too ambiguous. For TCGA quantification reports, only perfect alignments with no mismatches are used. 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 BWA CIGAR strings are discarded.

For reads retained after filtering, each coordinate for each read alignment is annotated using the reference databases (Table S5.1), 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. By considering

heuristically determined priorities (Table S5.1), we resolve the first issue by giving each alignment a single annotation. 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 S5.1 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, the read is flagged as cross-mapped (de Hoon 2010), and all of its miRNA annotations are preserved, while all of its non-miRNA annotations are discarded. This 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. 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 particular miRNAs are summed and normalized to a million miRNA-aligned reads to generate the quantification files that are submitted to the DCC. Quantification files include information on variable 5' and 3' read alignment locations, which can reflect isoforms, adapter trimming and RNA degradation.

NMF clustering of miRNA-seq data

Normalized read count data for 158 tumor samples were extracted from Level 3 data archives on the TCGA Data Portal website (http://tcga.cancer.gov/dataportal). The set of isoform.quantification.txt files, which give read counts at base pair resolution, was processed to sum up read counts at mature and star strand resolution (corresponding to miRBase v16 MIMAT accessions). Read counts were normalized to RPM, i.e. to reads per million reads aligned to miRBase mature or star strands. Mature and star strands were ranked by RPM variance across the samples, and the most variant 25% were used as the input to clustering. Unsupervised consensus clustering was done with the NMF v0.5.02 package (Gaujoux and Seoighe 2010) in R 2.12.0, using the default Brunet algorithm and 200 iterations, with the rank survey using 50 iterations. A four-cluster result was selected by considering profiles of cophenetic score and average silhouette width for clustering solutions having between 3 and 15 clusters; comparing silhouette plots for favorable solutions; comparing core/non-core cluster members with clinical covariate tracks (below); and preferring a result set with fewer, larger clusters. Silhouette results were generated from the consensus membership matrix using the R 'cluster' v1.14.1 package. Silhouette width profiles were generated for samples ordered as in the NMF heatmap, and atypical, or 'non-core' members in each cluster were identified using a silhouette width threshold set to a fraction (0.90) of the maximum width in each cluster. Asymptotic association p-values for covariate contingency tables were

calculated using R's chi-square test, for NMF sample groups against clinical covariates, sample groups representing expression subtypes, and groups generated by unsupervised clustering of DNA methylation data.

Table S5.1 Annotation priorities. Description of 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	mature strand	miRBase v13
2	star strand	
3	precursor miRNA	
4	stemloop, from 1 to 6 bases outside the mature strand, between the mature and star strands	
5	"unannotated", any region other than the mature strand in miRNAs where there is no star strand annotated	
6	snoRNA	UCSC small RNAs, RepeatMasker
7	tRNA	
8	rRNA	
9	snRNA	
10	scRNA	
11	srpRNA	
12	Other RNA repeats	
13	coding exons with zero annotated CDS region length	UCSC knownGenes
14	3' UTR	
15	5' UTR	
16	coding exon	
17	intron	
18	LINE	UCSC RepeatMasker
19	SINE	
20	LTR	
21	Satellite	
22	RepeatMasker DNA	
23	RepeatMasker Low complexity	
24	RepeatMasker Simple Repeat	
25	RepeatMasker Other	
26	RepeatMasker Unknown	

References

de Hoon MJ, Taft RJ, Hashimoto T, Kanamori-Katayama M, Kawaji H, Kawano M, Kishima M, Lassmann T, Faulkner GJ, Mattick JS, Daub CO, Carninci P, Kawai J, Suzuki H, Hayashizaki Y. Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. Genome Res. 2010;20(2):257-64.

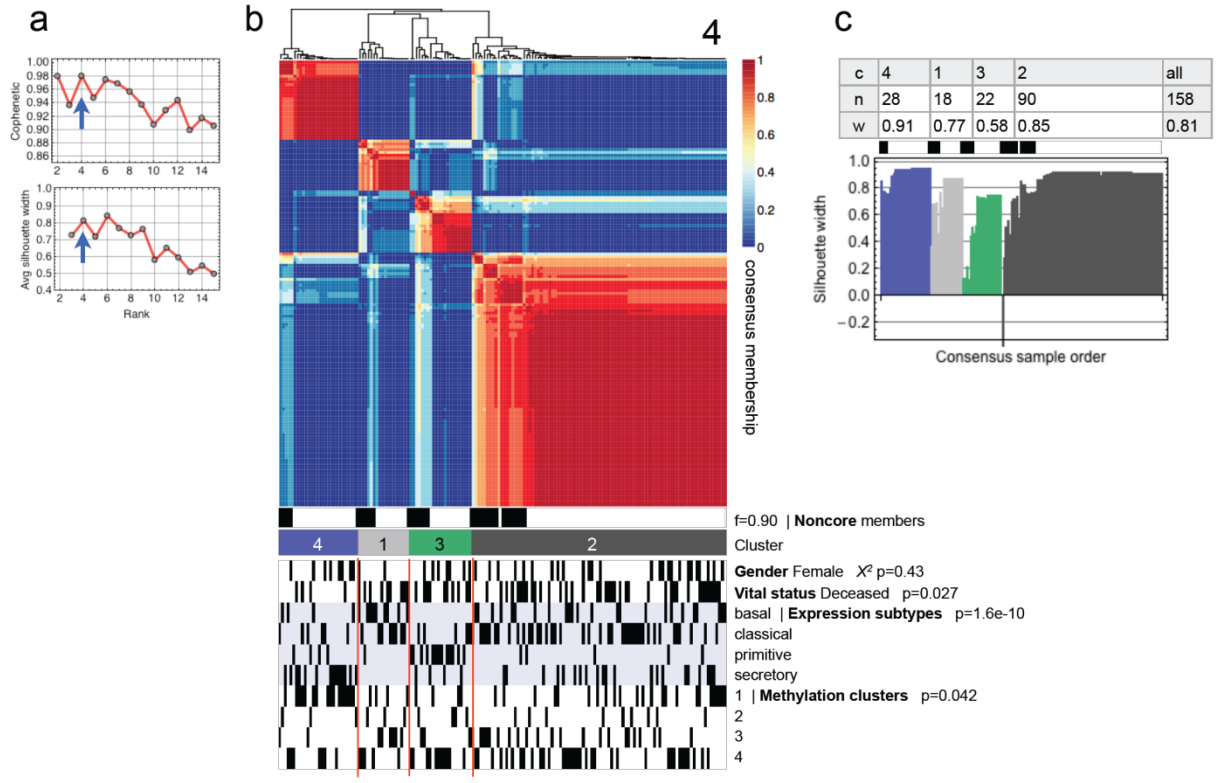
Khattra J, Marra MA. Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells and cell lines. Genome Res. 2007. 17(1):108-16 .

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009. 25(14):1754-60.

Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 2010 Jul 2;11:367.

Figure S5.1. miRNA clustering.

Sample groups identified by NMF consensus clustering of miRNA-seq abundance profiles for 158 tumor samples. **a)** Profiles of cophenetic correlation coefficient and average silhouette widths for ranks 3 to 15. **b)** Consensus membership heatmap for four clusters. Tracks below the heatmap show non-core or atypical members of each cluster (black), then clinical covariates and sample groups generated by other data types. **c)** Silhouette width profile for samples ordered as in the NMF heatmap. The table summarizes the number of samples in each cluster and silhouette scores of cluster members, as well as total number of samples and overall silhouette width.



c	4	1	3	2	all
n	28	18	22	90	158
w	0.91	0.77	0.58	0.85	0.81

Supplementary Methods S6: DNA methylation

Sample preparation and hybridization

Two high-throughput DNA platforms were used for TCGA LUSC 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 Infinium 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) sies. 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 using the MethyLight assay in TCGA LUSC samples from batches 23, 31, 39, 53, 60, 77, 101 and 140. In total we surveyed 215 TCGA LUSC tumors and 67 normal lung tissue samples in this analysis. MethyLight assays were performed as previously described [4]. The MethyLight assay primers and probe for *CDKN2A* (CDKN2A-M2; HB-081) have been 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)$_{sample}$ / (CDKN2A-M2 methylated reaction / *ALU* control reaction)$_{M.SssI\text{-}Reference}$, 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 Infinium HM27 arrays, mean non-background corrected M and U signal intensities for each locus were extracted from Illumina BeadStudio (or GenomeStudio) software. The beta value DNA methylation scores for each sample and locus were calculated as (M/(M+U)).

Detection p-values were calculated by comparing the set of analytical probe replicates for each locus to the set of 16 negative control probes. The negative controls are modeled by normal distribution. The detection p-value for the probe with intensity I_{probe}, is calculated as: $1\text{-}Z((|I_{probe} - \mu_{neg}|)/\sigma_{neg})$. In this formula, μ_{neg} and σ_{neg} are the average and the standard deviation of the signals from the negative controls, and Z is the one-sided tail probability of the standard normal distribution.

Infinium HM450 array raw data were imported into the R computing language (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.

Data Used

The data levels and the files contained in each data level package are described below and are available at the TCGA Data Portal website (http://tcga.cancer.gov/dataportal).

LEVEL 1: Level 1 HM27 data contain the non-background corrected signal intensities of the methylated (M) and unmethylated (U) probes and the mean negative control cy5 (red) and cy3 (green) signal intensities. A detection p-value for each data point, the number of replicate beads for methylated and unmethylated bead types as well as the standard error of methylated and unmethylated signal intensities are also provided for each sample and probe. Similar values are also provided for the negative control probes. Level 1 HM450 data consist of binary intensity data (.idat) files.

LEVEL 2: Level 2 data files contain the beta value calculations for each probe and sample. Data points with detection p-values > 0.05 were deemed not significantly different from background, and were masked as NA.

LEVEL 3: Level 3 data contain beta value calculations, gene IDs and genomic coordinates for each probe on the array. Annotation is based on the As in Level 2 data files, data points with detection p-values > 0.05 were deemed not significantly different from background, and were masked as NA. In addition, for HM450 arrays, probes marked as having a SNP (dbSNP build 128) within 10 bases of the interrogated cytosine, as well as those containing repetitive element DNA sequences in more than 10 bp of each 50 bp probe sequence are masked across all samples.

The data packages used for the following analyses are listed below. Please note that with continuing

updates of genomic databases, data archive revisions become available at the TCGA data portal. The following data archives were used for the analyses described in this manuscript:

jhu-usc.edu_LUSC.HumanMethylation27.Level_3.1.4.0/	27-Aug-2010 18:11
jhu-usc.edu_LUSC.HumanMethylation27.Level_3.2.4.0/	31-Aug-2010 07:30
jhu-usc.edu_LUSC.HumanMethylation27.Level_3.3.0.0/	27-Aug-2010 18:10
jhu-usc.edu_LUSC.HumanMethylation27.Level_3.4.0.0/	24-Sep-2010 10:27
jhu-usc.edu_LUSC.HumanMethylation27.Level_3.5.0.0/	23-Nov-2010 16:13
jhu-usc.edu_LUSC.HumanMethylation450.Level_3.1.1.0/	04-Jan-2012 16:36
jhu-usc.edu_LUSC.HumanMethylation450.Level_3.2.1.0/	04-Jan-2012 16:33
jhu-usc.edu_LUSC.HumanMethylation450.Level_3.3.1.0/	04-Jan-2012 16:31
jhu-usc.edu_LUSC.HumanMethylation450.Level_3.4.1.0/	04-Jan-2012 16:37
jhu-usc.edu_LUSC.HumanMethylation450.Level_3.5.1.0/	04-Jan-2012 16:36 -

Joint clustering of 27k and 450k platforms

More than 90% of the CpG loci assayed in the Infinium HM27 platform are also found on the HM450 array. However, the HM27 array uses the Type I chemistry exclusively, while the majority of HM450 probes employ Type II chemistry. The change of chemistry produces effects that, though moderate at individual loci, are sufficiently widespread to dominate unsupervised cluster analyses. To minimize these effects we first performed empirical Bayes modified t-tests to identify and filter any probes showing significant platform effects. This reduced the number of eligible CpG sites from 23525 snp-free, autosomal loci common to the two platforms, to 8228 eligible CpGs, from which the most variable were used in an unsupervised clustering analysis. We used consensus clustering as implemented in the Bioconductor package ConsensusClusterPlus [8], with Euclidean distance and partitioning around medoids (pam), to derive clusters.

Calculation of rates of differential DNA methylation

Rates of differential DNA methylation were estimated by direct comparison of matched tumor and normal samples. In total, 68 sample pairs were used for those CpG loci available on both platforms and 37 sample pairs for those loci found only on the 450k array. Specifically, denoting the difference in DNA methylation in sample pair i, at locus l, as $D_{il} = T_{il}-N_{il}$, we assume that when there is no differential expression, the D_{il} are distributed as $N(\mu=0, \sigma^2_l)$, where the variance, σ^2_l is platform specific.

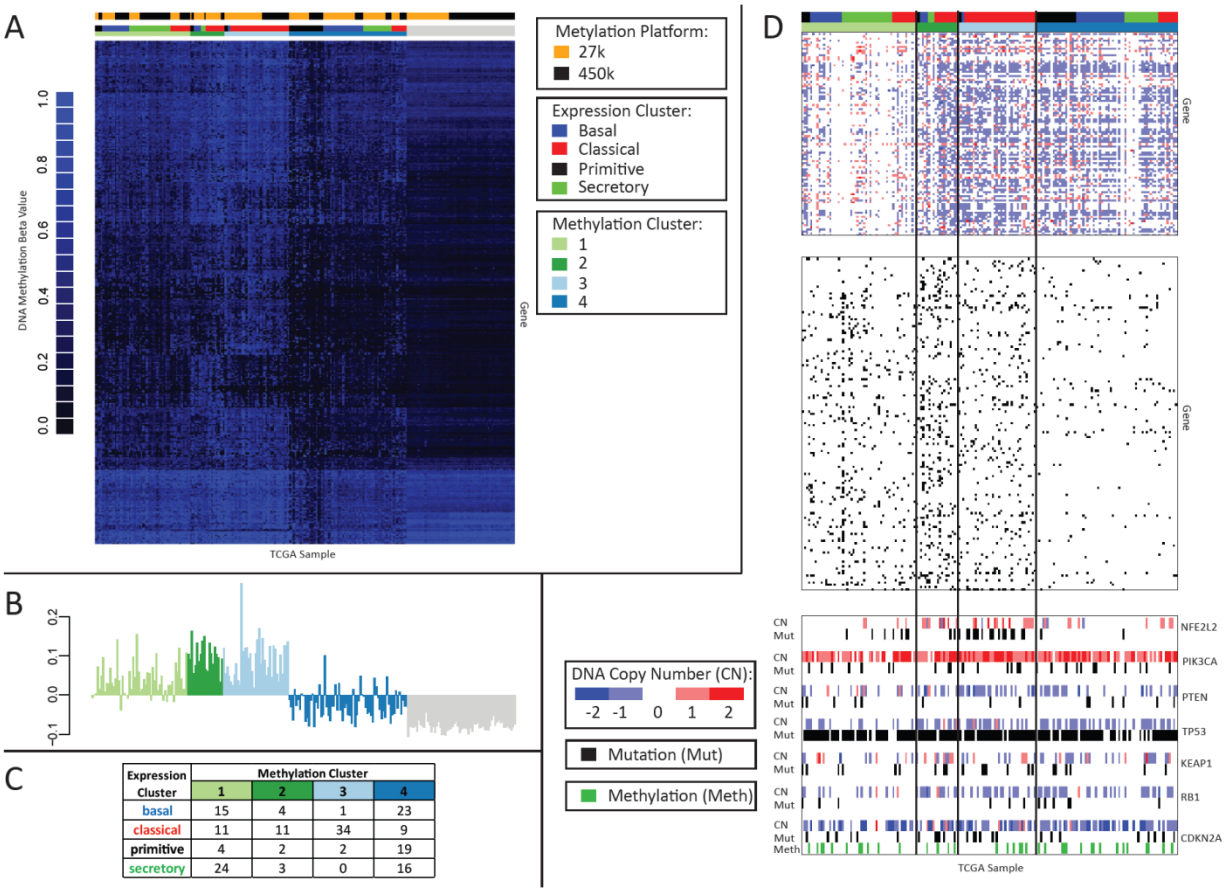
The variance of DNA methylation levels is higher in cancer than in normals, even in the absence of explicit differential methylation, [9] so we model the variance of D at each locus l as $\sigma^2_= \sigma^2_{Nl+} \sigma^2_{Tl=} \sigma^2_{Nl+} C\sigma^2_{Nl.}$ Using a simple linear model to describe the relationship between σ^2_{Nl} and σ^2_{Tl}, we estimated C to be 1.96, increasing it to 3.0 in order to conservatively shrink the estimates of rates of differential methylation slightly. Thus, hypermethylated samples satisfy $D > 3*q*\sqrt{2}*sd(N)$ while hypomethylated samples satisfy $D < -3*q*\sqrt{2}*sd(N)$, where q is 95^{th} percentile of the normal distribution. Probes exhibiting tumor specific DNA methylation changes are provided in Data Files S6.1 and S6.2, including estimates of the frequency of hyper- or hypo-methylation, confidence intervals for the same, and correlation with expression.

A DNA hyper-methylation score was calculated from methylation values at loci common to both platforms, for which the lower, 95% confidence bound on the frequency of DNA hyper-methylation was at least 0.05. In all 2097 loci covering 1377 distinct genes contribute to the score, which was calculated by first subtracting the mean methylation level at each locus, and then averaging the centered beta values within sample.

Figure S6.1 Methylation clustering

In Panel A, an unsupervised cluster analysis of DNA methylation levels at highly variable loci reveals several subgroups of LUSC tumors with distinctive methylation patterns. Expression based subtypes are annotated in color along the top, as are the two methylation platforms. In panel B, our DNA hypermethylation score, which summarizes methylation levels in frequently hypermethylated genes (see Data File 6.1 for details), is shown for the same samples. The table in panel C summarizes the overlap between expression-based subtypes, and those derived on the DNA methylation data. Finally, panel D depicts the distribution of major mutations across methylation subtypes.

Methylation cluster 4, the most normal-like of the tumor groups, shows little DNA hyper-methylation, and includes most of the samples in the primitive expression cluster. A number of secretory and basal samples also segregate to this cluster, though samples from the classical expression cluster are notably under-represented. Cluster 3 is predominantly made up of classical samples, and to a lesser extent cluster 2 is also classical rich. These two clusters show the highest levels of DNA hyypermethylation. Another interesting characteristic of Methylation Cluster 3 is evident in panel D. NFE2L2 mutations and CNVs are associated with the classical expression subtype, and more specifically with methylation cluster 3, which contains nearly all of the classical samples with NFE2L2 alterations. In cluster 1 which shows intermediate methylation levels, all expression groups except primitive are well represented.



Expression Cluster	Methylation Cluster 1	2	3	4
basal	15	4	1	23
classical	11	11	34	9
primitive	4	2	2	19
secretory	24	3	0	16

We estimated the rate of differential DNA methylation at each CpG locus, using tumor samples for which matched, adjacent normal tissue was available for comparison. The most frequently hypermethylated CpGs are described in Data File S6.1, while hypomethylated genes are listed in Data

File S6.2. A number of statistics are calculated for each locus, including a) the estimated frequency of differential methylation, b) an exact 95% confidence interval for the frequency, mean methylation in normal samples, mean methylation in altered tumor samples, and correlation to RNAseq expression levels. Each locus is also characterized by CpG Island status, and the distance to the transcription start site, when the CpG can be unambiguously associated with a specific gene. For CpGs that are found on both the HM27K and HM450K arrays, differential methylation rates are estimated from 61 samples with matched normals. Probes specific to the HM450K array are flagged to indicate that fewer tumor-normal sample pairs (n=37) were available for analysis.

Supplementary Data File S6.1. Hypermethylated loci.

data.file.S6.1.hypermethylated.loci.xlsx

Supplementary Data File S6.2. Hypomethylated loci.

data.file.S6.2.hypermethylated.loci.xlsx

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. Wilkerson, M. D., & Hayes, D. N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics (Oxford, England), 26(12), 1572-3. doi:10.1093/bioinformatics/btq170; 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.

Supplementary Methods S7: Pathways and integrated analyses

A. Integrative clustering

Data processing

- Somatic mutation: the mutation MAF file was used. A gene by sample matrix of binary values (1-mutated, 0-WT) was generated for clustering. MutSig genes with a q<=0.5 were included for clustering.
- Copy number: the copy number segmented data (CNV removed) based on Affymetrix SNP 6.0 array was used. Non-redundant copy number regions were first obtained by adapting a method described in [1]. Briefly, a region along a chromosome is defined by consecutive positions with a maximum Euclidean distance (based on copy number segmented values) between any adjacent two probes smaller than 0.01. This resulted in a total of 4,962 copy number regions. Each region was then represented by its medoid signature, reducing the dimension of the data to 4,962 rows and 178 columns.
- Gene expression: the gene expression RPKM data based on RNA-seq platform was used. Median absolute deviation was used to select the top 2,000 most variable genes for clustering. Data were log2 transformed and standardized.
- Methylation: the methylation Beta values were based on the combined data from the HumanMethylation27 and HumanMethylation450 platforms (see Methylation Supplemental Methods). Median absolute deviation was used to select the top 2,000 most variable CpG sites for clustering.

Integrative clustering using iCluster

Integrative clustering of somatic mutation, DNA copy number, DNA methylation, and mRNA expression data was performed using the iCluster framework originally described in Shen et al. [2]. The problem 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 with lasso [3] penalty terms for balancing the fitness and the complexity of the model. We applied an extended algorithm that generalizes the original method to encompass both discrete and continuous data types using the generalized linear model framework (Shen et al, manuscript in preparation). In brief, for the binary mutation data matrix, we assume each entry is a realization of a Bernoulli random variable x_{ij} associated with the *j*th gene in the *i*th sample, with marginal mutation probability π_{ij} and the logit function as its canonical link to equate to the latent variables. For data types that are on a continuous scale (copy number, methylation, and mRNA expression data), a Gaussian distribution with identity link function was used.

Model selection

The number of clusters (K) is unknown and needs to be estimated. We compute a deviance ratio metric that 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 lasso penalty parameter values, a very large search space needs to be covered. We used an efficient sampling method that utilizes the uniform design (UD) [4]. A theoretical advantage of the uniform design over an exhaustive grid search is the uniform space filling property that avoids computation at near-by points. For each sampled "experimental" point, we fit an iCluster model with the sampled parameter setting. The best

parameter setting was chosen that minimizes the Bayesian information criterion (BIC). The model selection procedure resulted in a three cluster (K=3) solution with 2 mutated genes (KEAP1 and NFE2L2), 1053 copy number regions, 1532 CpG sites, and 754 differentially expressed genes that are cluster-discriminant at a false discovery rate of 10% or less.

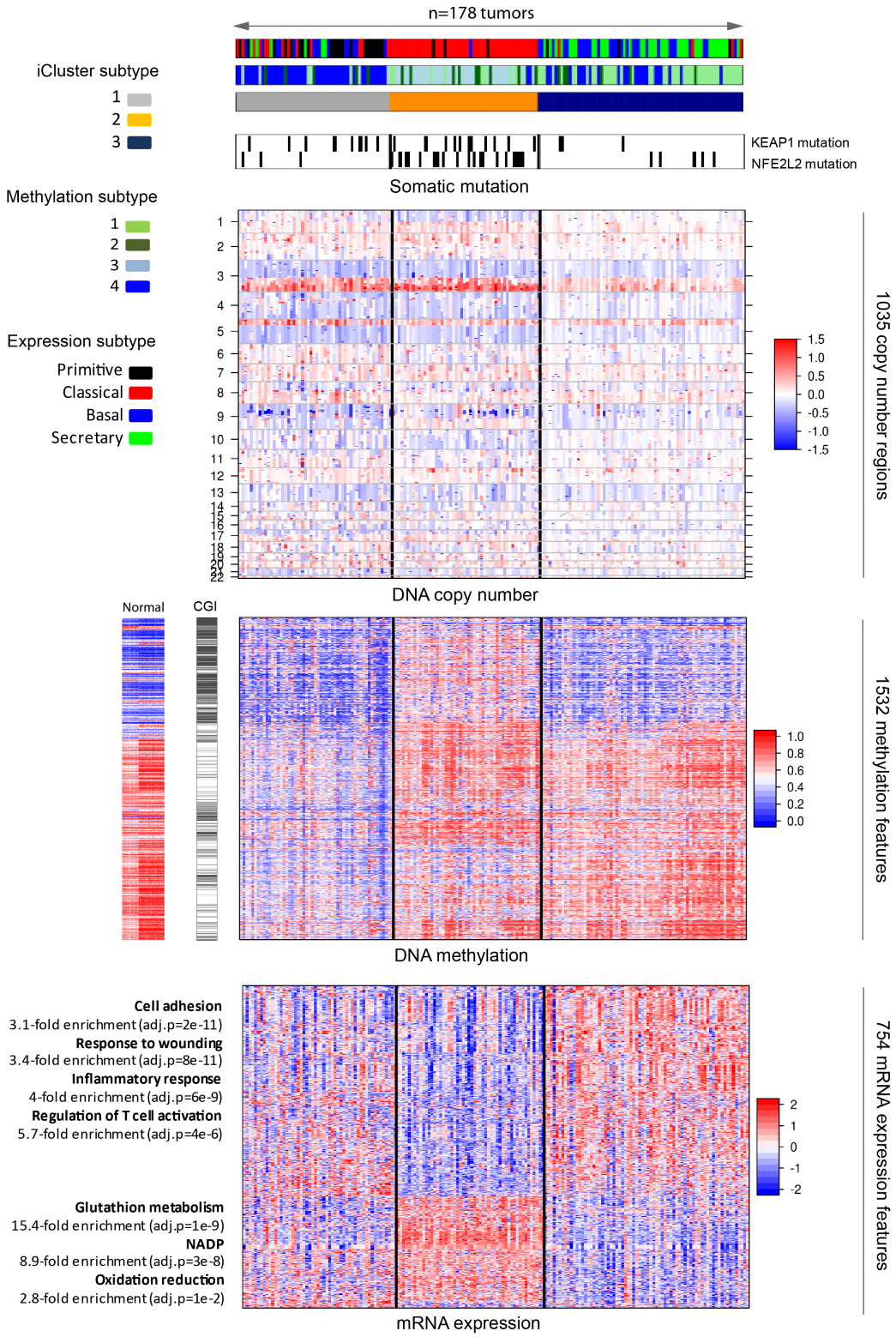
Gene-annotation enrichment analysis

Gene-annotation enrichment analysis was performed using DAVID (PMID: 19131956).

References

1. van de Wiel, M.A. and W.N. Wieringen, CGHregions: dimension reduction for array CGH data with minimal information loss. Cancer Inform, 2007. 3: p. 55-63.
2. 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.
3. Tibshirani, R., Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B-Methodological, 1996. 58(1): p. 267-288.
4. 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.

Figure S7.A.1 – Integrative Clustering. iCluster reveals three distinct molecular subgroups among the 178 lung squamous cell carcinomas. Heatmap displays coordinated patterns of alteration in somatic mutation, DNA copy number, DNA methylation, and mRNA expression.



B. Pathway Signatures

Gene transcription signature of pathways were defined as follows. P53 pathway: "IARC" signature, canonical bound and up-regulated p53 gene targets, as catalogued in the p53 IARC database [1]; "GSK" signature, from Glaxo-Smith-Kline (GSK) cell line database, coupled with "R14" p53 database of mutations in cell lines (N=248 cell lines with TP53 status), where a t-test of $P<0.01$ was used to determine genes higher in wt versus mutant cell lines; "Kannan" signature, from MSigDB ("UP" targets), [2]; "Troester" signature, list of genes reported repressed by TP53 knockdown in MCF7 cells (from Troester et al. [3]). RB pathway: "Lara" signature, from GEO database GSE9562, comparing mouse keratinocyte cultures with RB1 knockout versus wt ($P<0.01$, fold>1.5); "Chicas" signature, from GSE19864 profiles of RNAi-mediated suppression of RB in IMR90 cells (using $P<0.01$, fold>1.5). PI3K pathway: Gene signatures were described previously in Creighton et al. [4]; "Saal" PTEN loss signature, genes correlated with Pten protein levels in breast cancer; "CMap" PI3K/mTOR signature, genes modulated *in vitro* by inhibitors to PI3K or mTOR, according to CMap dataset ($P<0.01$, comparing PI3K/mTOR-inhibited cells with the rest of the Cmap profiles); "Majumder" Akt signature, genes modulated in a mouse model of inducible Akt ($P<0.001$). Nrf2 pathway: "Malhotra" signatures from ref [5], which combined expression profiling and Chip-seq of mouse embryonic fibroblasts (MEFs) with either constitutive nuclear accumulation (Keap1−/−) or depletion (Nrf2−/−) of Nrf2, including genes downregulated in Nrf2−/− versus WT AND Nrf2 bound (Supp Table 3 in the Malhotra paper), and genes upregulated in Keap1−/− versus WT AND Nrf2 bound (Supp Table 4); "GSE28230," from GEO dataset of A549 adenocarcinoma lung cancer cells with siRNA knockdown of NRF2 (used $P<0.01$, fold>1.5); "Osburn," from GSE11287 dataset of mouse liver with or without Keap1 knockout used ($P<0.01$, fold>1.5).

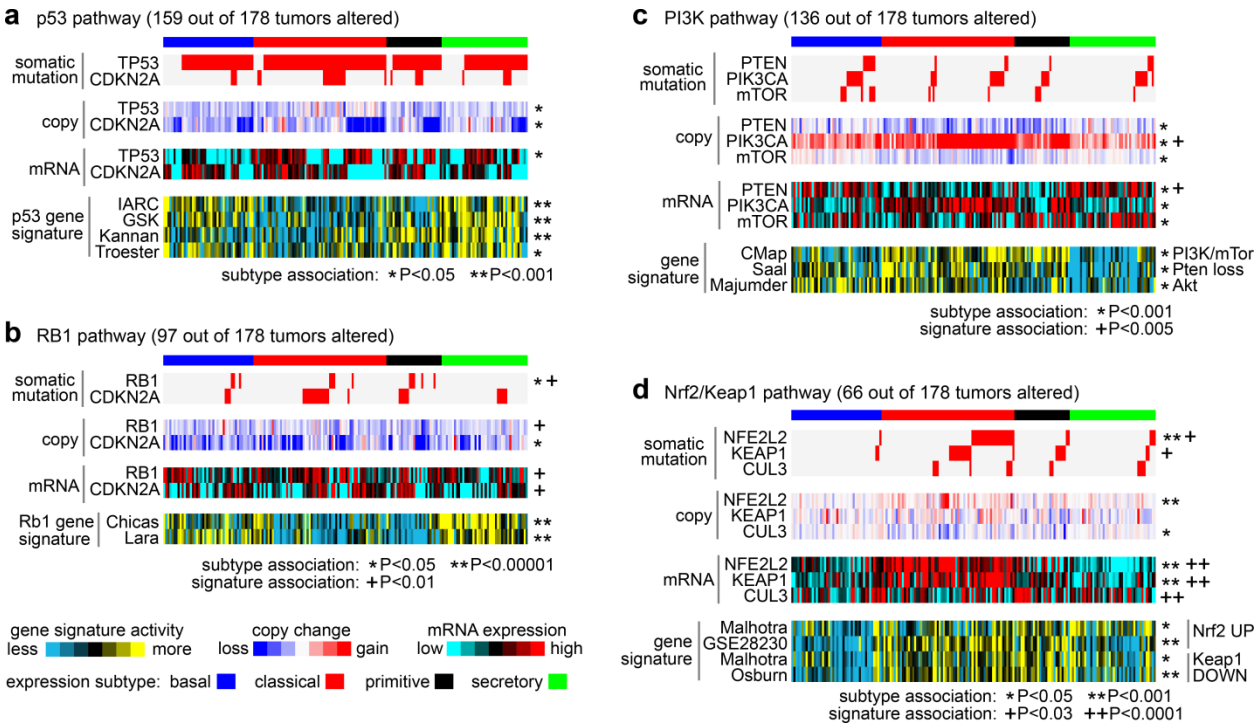
For a given gene signature, we extracted the expression values from the TCGA expression RNA-seq dataset. For each gene, we normalized expression values to standard deviations from the median across tumors. For signatures with genes moving in one direction (the p53 and Nrf2/Keap1 signatures), we computed the average normalized expression of the signature genes within each tumor. For signatures with "up" and "down" genes (the Rb1 and PI3K signatures), we computed our previously described "t-score" (as described in the TCGA Ovarian paper, ref [6]). Across the tumors, we normalized the gene signature scores to standard deviations from the median across tumors, and a "summary score" for each pathway was computed as the average of the individual normalized signature scores.

References

1. http://www-p53.iarc.fr/TargetGenes.html
2. http://www.broadinstitute.org/gsea/msigdb/cards/KANNAN_TP53_TARGETS_UP.html
3. Troester MA, Herschkowitz JI, Oh DS, He X, Hoadley KA, Barbier CS, Perou CM. Gene expression patterns associated with p53 status in breast cancer. BMC Cancer. 2006 Dec 6;6:276.
4. Creighton CJ, Fu X, Hennessy BT, Casa AJ, Zhang Y, Gonzalez-Angulo AM, Lluch A, Gray JW, Brown PH, Hilsenbeck SG, Osborne CK, Mills GB, Lee AV, Schiff R. Proteomic and transcriptomic profiling reveals a link between the PI3K pathway and lower estrogen-receptor (ER) levels and activity in ER+ breast cancer. Breast Cancer Res. 2010;12(3):R40.
5. Malhotra D, Portales-Casamar E, Singh A, Srivastava S, Arenillas D, Happel C, Shyr C, Wakabayashi N, Kensler TW, Wasserman WW, Biswal S. Global mapping of binding sites for Nrf2 identifies novel targets in cell survival response through ChIP-Seq profiling and network analysis. Nucleic Acids Res. 2010 Sep;38(17):5718-34
6. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011 Jun 29;474(7353):609-15

Figure S7.B.1 Pathway transcriptional signatures.

In squamous lung cancers, gene transcription signatures of pathway activity correlate with molecular alterations and expression-based subtype. For p53 (a), Rb (b), PI3K (c), and Nrf2/Keap1 (d) pathways, copy and mRNA levels and somatic mutation events are shown for key driver genes of interest. For each pathway, a number of gene transcription signatures were defined (using the literature or public databases, see Methods) and applied to the expression profiles, in order to score each tumor for relative inferred pathway activity (yellow: more active, blue: less active). The tumor ordering differs between panels a-d, though first-level ordering in each case is by expression subtype. For subtype associations, two-sided p-values are by ANOVA (except for mutation data, where p-values are by chi-squared); for signature associations, two-sided p-values are by Pearson's correlation of each feature versus the average of the individual signature scores. (a) The p53 pathway is altered in most all tumors, with relatively lower activity levels observed for the classical and primitive subtypes and higher activity for the secretory and basal subtypes. (b) The RB pathway appears more active within the secretory subtype. (c) The PI3K pathway appears more active within the classical subtype, which has more frequent high-level copy gain of *PIK3CA* and copy loss of mTOR. (d) The Nrf2 pathway appears more active within the classical subtype, which has more frequent *NFE2L2/KEAP1* mutations and *NFE2L2* gain/*CUL3* loss.



C. Pathway shift

Arguably, gene activities are never truly observed but instead are inferred from surrounding measurable consequences. Take for example a metabolic enzyme that acts on a substrate to produce a product. One could argue that the only definitive way to assess the activity of the enzyme is to measure the rate of substrate turnover. However, if a direct measure of the disappearance of the substrate (or appearance of the product) is not available than one could look at the activity of other enzymes that depend on the product as their substrate to infer the activity of the first enzyme. This leads to a recursive definition of the activity of the enzyme in terms of the activity of other neighboring enzymes. In the same way, we can infer the activity of a transcription factor or kinase based on the activities of other genes in their regulatory neighborhoods.

We derive a *Pathway Shift* (*P-Shift*) score based on the intuition of comparing the observed downstream consequences of a gene's activity to what is expected from its regulatory inputs. The *P-Shift* has the form:

$$PS(f) = \log\left(\frac{Observed(f)}{Expected(f)}\right), \quad (1)$$

where the observed activity of *f* is derived from the downstream targets and the expected activity of *f* is derived from the upstream regulators. The caveat of course is that we never get to truly "observe" gene *f*'s activity so we must infer it from the activity of downstream targets. As implied above, the estimation of such an activity is necessarily recursive, requiring us to first estimate the activity of the downstream targets of *f* before we can predict *f*'s own activity. The computation is naturally framed as an inference problem over a set of interdependent variables, some of which are hidden. In the last couple of decades, efficient procedures have been developed to compute the probabilities of a system of variables connected together in a probabilistic graphical model (see (Friedman 2004) for a review).

We previously described a factor-graph-based approach for integrating diverse types of omics data with genetic pathways called PARADIGM (Vaske, Benz et al. 2010). The approach assesses the activity of a gene in the context of a genetic pathway diagram ϕ by drawing inferences from a dataset of observations *D*. The dataset can include multiple different types of measurements for a patient sample such as gene expression and genomic copy number variation. Before supplying the data to PARADIGM, each dataset type is first transformed into rank-ratios. This is done by ranking all values across all samples from smallest to largest and then each rank *r* is transformed into the range *[0,1]* by the formula *(r-1)/(N*G-1)* where *N* is the number of samples and *G* is the number of genes measured. Briefly, PARADIGM then uses a belief-propagation algorithm on a factor graph derived from ϕ to combine gene expression, copy number, and genetic interactions to compute *inferred pathway levels* (IPLs) for each gene, complex, protein family, and cellular process. The IPL for a gene is a signed log-posterior odds of the state of the gene given the observed data. Positive IPLs reflect how much more likely the gene is active in the tumor, while negative IPLs reflect the negative log probability of how likely the gene is inactive in the tumor relative to normal.

In many cases the activity can be inferred from the direct and indirect regulatory influences that the gene participates. For example, PARADIGM's inferred activity of a transcription factor increases if several of its targets are overexpressed and amplifications in the genome cannot explain their up-regulation. By logical extension of this reasoning, PARADIGM may infer that a kinase is active if several

of its targets have already been inferred to be active (e.g. the targets may be transcription factors with over-expressed targets).

The core of our approach estimates the *P-Shift* in each tumor sample for each *focus gene* (FG) using two runs of the original PARADIGM algorithm. We refer to these two runs as the Regulators-only and the Targets-only runs (R-run and T-run for short). In the R-run, a neighborhood of upstream regulators is left connected to FG but all downstream targets are disconnected. All variables in PARADIGM are trinary, representing whether a feature is more active in the tumor relative to normal, x^a, more inactive in the tumor than normal, x^i, or the same level in tumor as in normal, x^0. The inferences derived from the R-run reflect the expected level of FG given the state of its regulators in a particular sample. In the T-run, FG is left connected to a neighborhood of its downstream targets while upstream regulators are disconnected. The P-Shift score then computes the difference between the inferred activities of FG determined in the T-run from those determined in the R-run.

To estimate pathway-neighborhood dependent inferences on focus gene *f*'s activity, we restrict our view of the data to subsets of features in ϕ and supply these as arguments to *D*. If $R \subset \phi$ is the set of regulators of *f* and $T \subset \phi$ is the set of targets, then *D(R), D(T),* and *D(f)* refer to the data observed for the regulators, targets, and the focus gene *f* respectively. Likewise, the interactions are restricted to the subset of features in the focus gene's neighborhood and denoted $\phi(T)$ to represent the pathway features and interactions involving only the targets to each other and to *f* itself. Similarly, $\phi(R)$ represents the same for the upstream regulators. With these definitions in hand we write the *P-Shift* score as the following log ratio of two constituent likelihood-ratios:

$$PS(f) = \log\left(\frac{LR\left(D(T) \mid x_f^a, \phi(T)\right)}{LR\left(D(R), D(f) \mid x_f^{\neg a}, \phi(R)\right)} \right), \quad (2)$$

where $LR(Y|x^a,Z)$ is defined as $P(Y|x^a,Z)/P(Y|x^{\neg a},Z)$, the likelihood ratio computed over one possible alternative value for *X*, x^a, compared to the probability of the other two possible values – less active in tumor x^i and similarly active in tumor x^0 – the combined event $x^{\neg a} = \{x^i, x^0\}$ is written for short. Note that only the expected term in the denominator contains the entry *D(f)*, which represents the actual data for the focus gene of interest. This reflects the assumption that the data on the focus gene provides evidence for the *cis*-regulatory state of the gene and so is included among the regulators of *f*. Note that if data on the direct activity of *f* were instead available, such as phosphorylation status or enzymatic activity, then that data could be considered for inclusion into the numerator term for the observed targets. The quantity in equation (2) reflects the degree to which the observed data for the targets is consistent with high activity of the focus gene relative to the observed data for the regulators and the gene in question.

Further expansion of the *P-Shift* score reveals the method by which it can be computed using the original PARADIGM algorithm. Application of Bayes Rule gives:

$$PS(f) = \log\left(\frac{P\left(D(T), x_f^a \mid \phi(T)\right)}{P\left(D(T), x_f^{\neg a} \mid \phi(T)\right)} \right) - \log\left(\frac{P\left(D(R), D(f), x_f^a \mid \phi(R)\right)}{P\left(D(R), D(f), x_f^{\neg a} \mid \phi(R)\right)} \right) - prior, \quad (3)$$

where *prior* is the log-prior-odds and has the same form as the first two terms in the equation except that all entries involving *D* are dropped. Another application of Bayes Rule would show that the first two joint probability ratio terms are equivalent to the LPO that the gene is active given either the state of the downstream targets (left-hand term) or the upstream regulators (right-hand term). The advantage of writing the joint probabilities in this form shows explicitly those terms of the form $P(D, x \mid \phi)$ that are each efficiently computed with a message-passing belief propagation procedure on the underlying factor graph encoded by ϕ. The message-passing procedure takes care of summing out all of the hidden variables present in ϕ – the states of complexes, cellular processes, and the activities of all other genes other than *f*. The computation implements an iterative form of the Expectation-Maximization procedure that sequentially updates all variables by forming a running average until either a convergence tolerance of 10^{-9} is reached or 10,000 maximum iterations are exceeded. The code is freely available through the libDAI C++ open source library (Mooij 2009). In the R-run version of PARADIGM, the LPO shown in Equation (3) and its corresponding log-prior odds are computed in two separate full factor graph convergence runs. Likewise, the T-run involves two separate EM runs to compute its two terms in Equation (3). Thus, in total, the computation time involved to compute the *P-shift* requires four EM convergence runs, but each task is run on a reduced pathway representation involving only the neighborhood of the focus gene. Thus, the computation time to calculate a *P-Shift* for an entire dataset requires *2k* PARADIGM runs where *k* is the number of mutated genes in the cohort.

In practice we use the inferred pathway levels (IPLs) from PARADIGM for the computation of the *P-Shift*. Specifically, we set $PS(f) = IPL_T(f) - IPL_R(f)$, where $IPL_{T|R}(f)$ is the IPL derived from the T- or R-run. The IPL is a signed LPO that always puts the highest probability state for *f* in the numerator. If the inactive state is in the numerator the IPL gives the negation of the LPO. This quantity is similar to Equation (3) except that the highest probability states determined in each run are contrasted. In the case where the active form of the gene is the most probable in each case the two formulas are equivalent. Finally, we found that a transformation of the P-Shift score to a Z-score provided better overall results. Each gene's local neighborhood could have a certain bias to lean toward either positive or negative scores. To account for this we constructed 100 random samples for each gene by shuffling data tuples around the SuperPathway. This effectively associated random data with each gene's neighborhood. *P-Shifts* were calculated for each of these 100 samples and each *P-Shift* was then normalized by subtracting the mean and dividing by the standard deviation determined from this simulation. We henceforth refer to these *P-Shift* Z-scores simply as the *P-Shifts.*

For computational efficiency, we use a local neighborhood around each gene rather than the entire network. Using the local neighborhood provides a good approximation of the full network as genes far away are expected to exert a diminishing influence on the inference of the focus gene. We tested including neighbors at distances 1, 2, and 3 and full for the positive controls. The empirical results indicate that the distinction degree increases only moderately after distances of two (data not shown). To build the neighborhood, we traverse the graph and include any pathway features (proteins, families, complexes, and processes) when there is at most one other intervening protein between the feature and FG. All interactions between the selected features were included in the neighborhood. If a protein was present in both the upstream and downstream neighborhoods, due to feedback circuitry, it was excluded from both the R- and T-runs. In addition, uninformative neighbors were excluded by removing any protein with a low level of expression variation across the cohort. A cutoff of 0.10 on the rank-ratio-transformed expression data was used for this filtering step.

Figure S7.C.1. Shift flowchart.

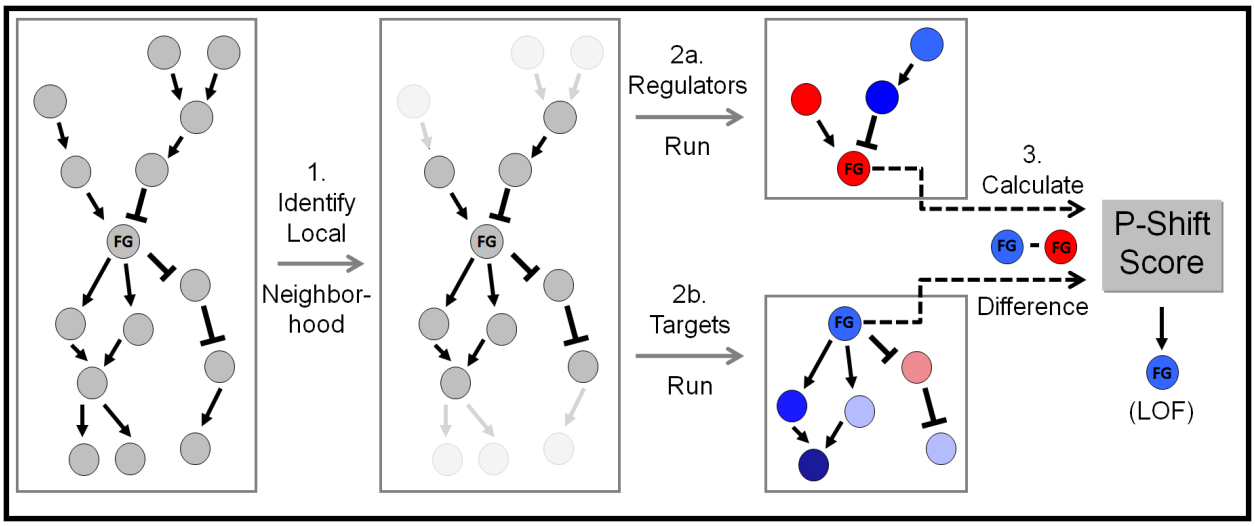
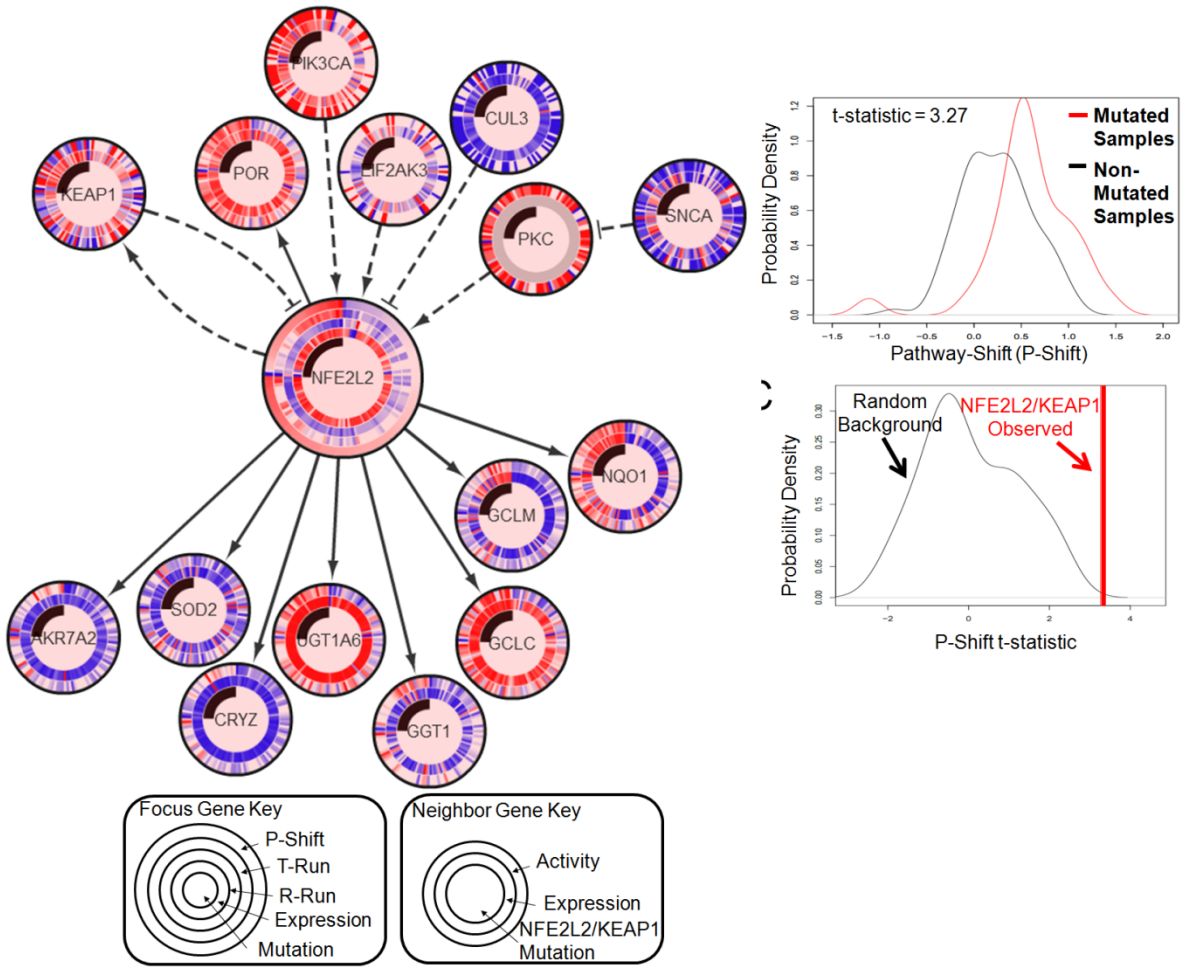


Figure S7.C.2 Shift results for *NFEL2L* and *KEAP1*.



D. Mutations with therapeutic relevancy.

Druggable genes were defined based on protein families as per Hopkins et al. and Russ et al [1, 2] and a publically-available list of 3361 druggable genes generated by Russ and coworkers was downloaded2. Mutated genes from the mutation annotation file (MAF) that overlapped with this druggable gene list were identified using an Oracle database. To determine the functional category of these potential therapeutic targets, protein domain information was downloaded from the Pfam database. The protein domains of each targetable gene were examined and tyrosine kinase, serine/threonine kinases, tyrosine phosphatases, PI 3-kinase catalytic and regulatory subunit genes, proteases, G-protein coupled receptors and nuclear hormone receptors were identified. Mutations from the MAF were then filtered based on RNA-Seq data and only mutations that were confirmed by RNA-Seq analysis were retained. Additionally, focal amplifications or deletions that were found to be significantly altered by the GISTIC algorithm were also considered as potential therapeutic targets. Five potential therapeutic targets were added from this analysis of deletions and amplifications (PTEN, FGFR1, EGFR, PDGFRA, and KIT).

References

1. Hopkins, A. L.; Groom, C. R., The druggable genome. Nat Rev Drug Discov 2002, 1, (9), 727-30.
2. Russ, A. P.; Lampel, S., The druggable genome: an update. Drug Discov Today 2005, 10, (23-24), 1607-10.

Figure S7.D.1 Validated and candidate therapeutic targets are commonly altered in lung SqCCs.
a, Heatmap representation of the number of somatic mutations validated by RNA sequencing within each protein class among all 178 lung SqCCs. Tumor samples are arranged horizontally with the number of mutations in each class of targetable alteration indicated by the color scheme at the bottom right (yellow=0 to red=5 or greater). The percentage of mutations in each class with a Mutation Assessor (MA) score of high or medium is noted on the right. b, Mutations and amplifications in selected genes across 114 patients are shown ranked into two classes. Class 1 genes are validated therapeutic targets being studied in clinical trials as of January 23, 2012; Class 2 genes are candidate therapeutic targets based on pre-clinical and early clinical studies. Mutational events are color-coded by predicted functional impact, as determined by Mutation Assessor score.

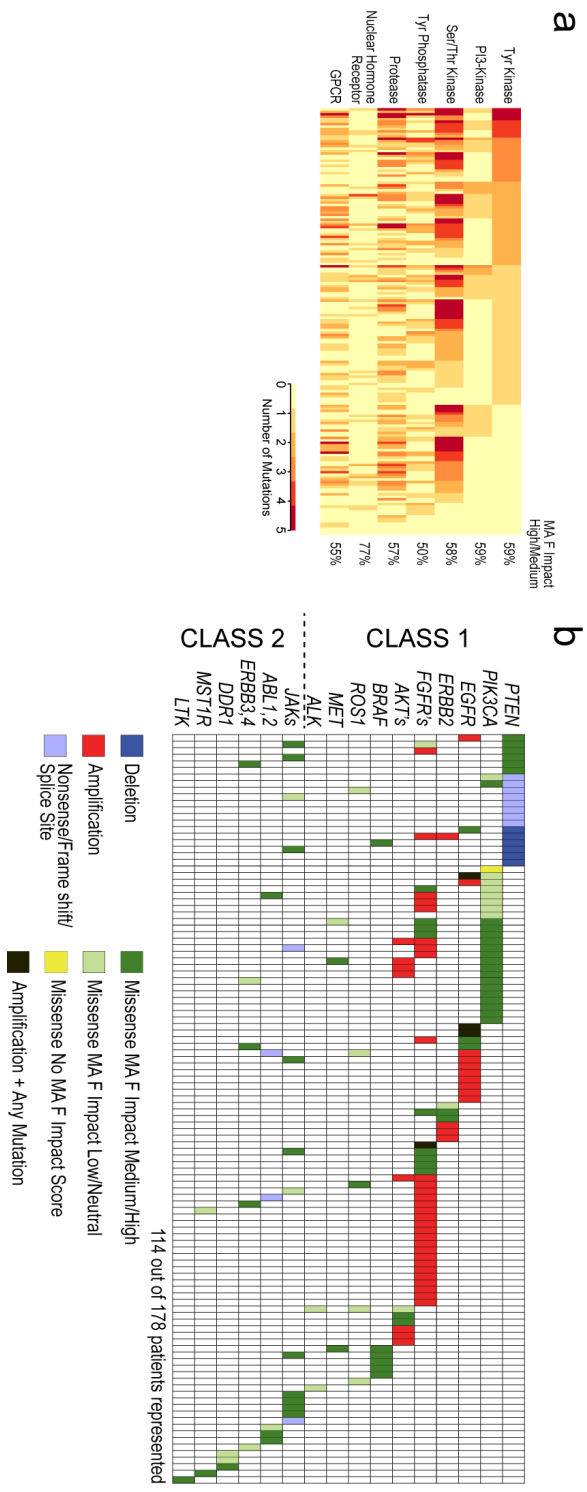

a
Tyr Kinase
PI3-Kinase
Ser/Thr Kinase
Tyr Phosphatase
Protease
Nuclear Hormone Receptor
GPCR
Number of Mutations
0 1 2 3 4 5
MA F Impact High/Medium
59%
59%
58%
50%
57%
77%
55%
b
CLASS 2
CLASS 1
PTEN
PIK3CA
EGFR
ERBB2
FGFRs
AKTs
BRAF
ROS1
MET
ALK
JAKs
ABL1,2
ERBB3,4
DDR1
MST1R
LTK
114 out of 178 patients represented
Deletion
Amplification
Nonsense/Frame shift/ Splice Site
Missense MA F Impact Medium/High
Missense MA F Impact Low/Neutral
Missense No MA F Impact Score
Amplification + Any Mutation

Figure S7.D.2 Mutations in the ERBB, FGFR, and JAK tyrosine kinases.

RNAseq confirmed mutations observed in the ERBB receptor tyrosine kinases (*EGFR*, *ERBB2*/Her2, *ERBB3*/Her3, *ERBB4*/Her4). B. RNAseq confirmed mutations observed in the FGFR receptor tyrosine kinases (*FGFR2*, *FGFR3*, and *FGFR4*). Three missense mutations were observed in *FGFR1* (P25Q, R445W, and W471L) but they were not verified in the RNA-Seq data. C. RNAseq confirmed mutations observed in the JAK tyrosine kinases (*JAK1*, *JAK2*, *JAK3*, and *TYK2*).

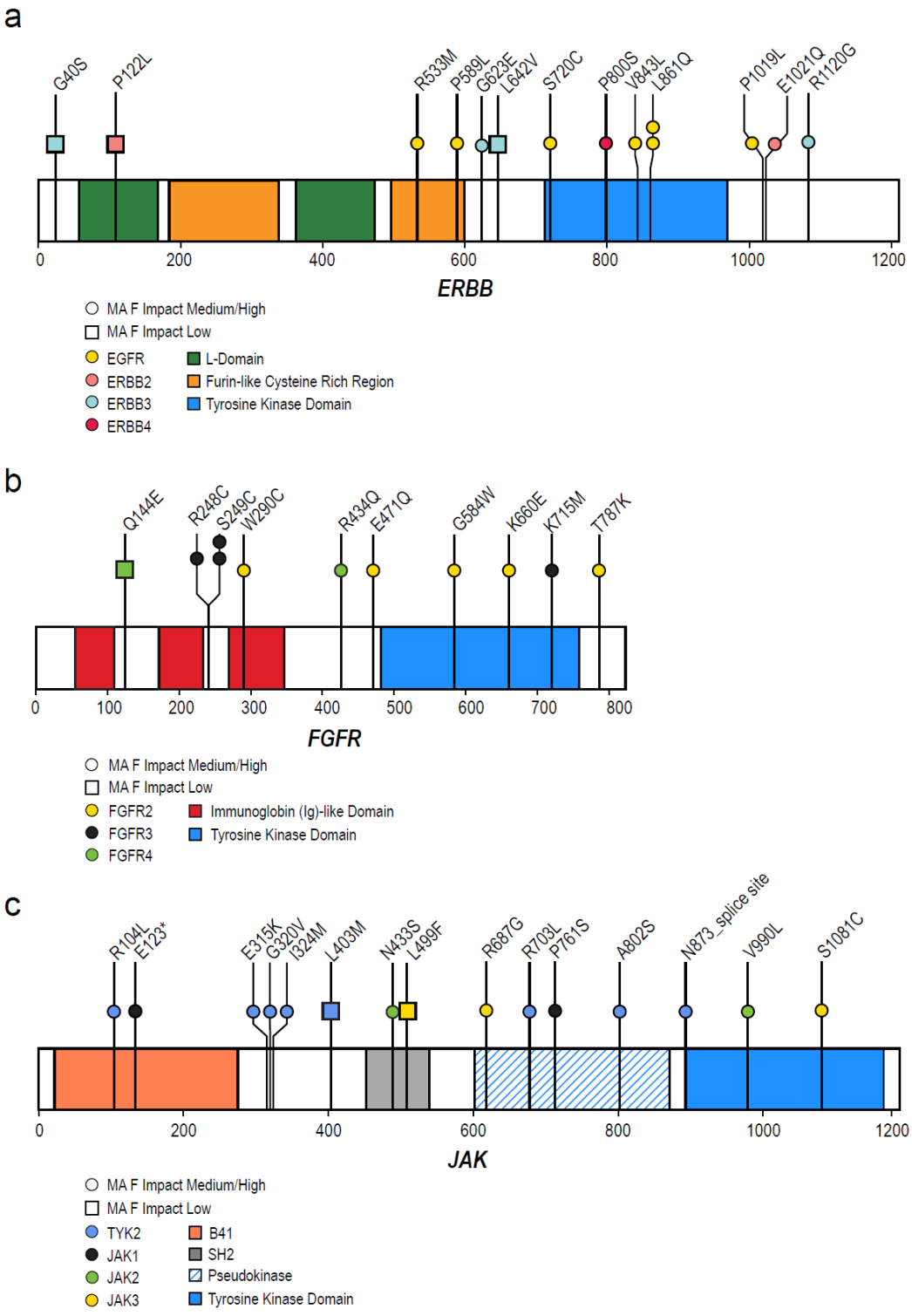

a
G40S
P122L
R533M
P589L
G623E
L642V
S720C
P800S
V843L
L861Q
P1019L
E1021Q
R1120G
0
200
400
600
800
1000
1200
ERBB
MA F Impact Medium/High
MA F Impact Low
EGFR
ERBB2
ERBB3
ERBB4
L-Domain
Furin-like Cysteine Rich Region
Tyrosine Kinase Domain
b
Q144E
R248C
S249C
W290C
R434Q
E471Q
G584W
K660E
K715M
T787K
0
100
200
300
400
500
600
700
800
FGFR
MA F Impact Medium/High
MA F Impact Low
FGFR2
FGFR3
FGFR4
Immunoglobin (Ig)-like Domain
Tyrosine Kinase Domain
c
R104L
E123*
E315K
G320V
I324M
L403M
N433S
L499F
R687G
R703L
P761S
A802S
N873_splice site
V990L
S1081C
0
200
400
600
800
1000
1200
JAK
MA F Impact Medium/High
MA F Impact Low
TYK2
JAK1
JAK2
JAK3
B41
SH2
Pseudokinase
Tyrosine Kinase Domain

Figure S7.D.3 Oncoprint of therapeutically relevant alterations.

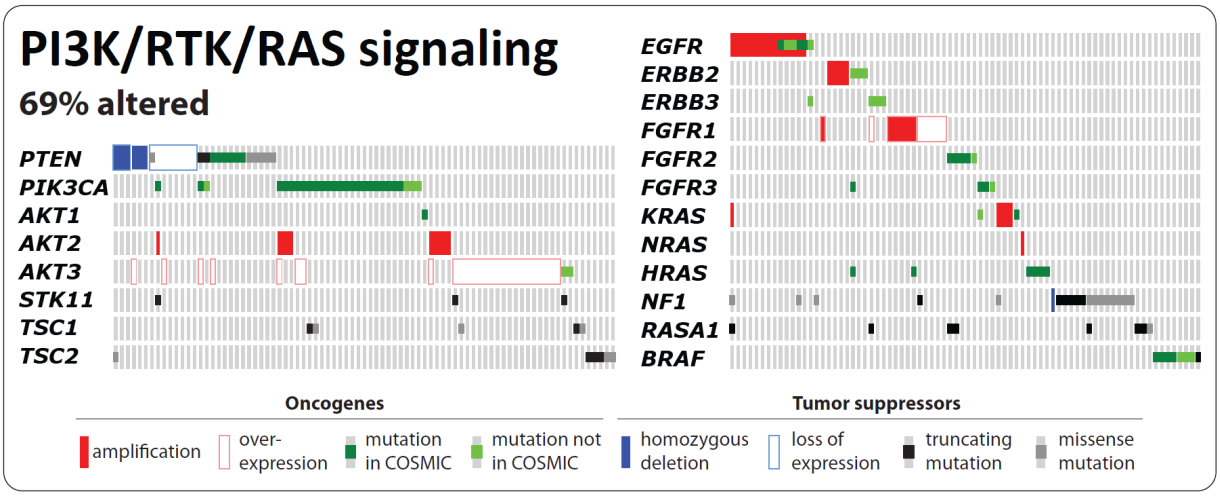


Figure S7.D.4 Pathway alterations in lung squamous cell carcinoma.

A. The MEMo algorithm [1] was run using mutations in all significantly mutated genes (MutSig), high-level amplifications and homozygous deletions of genes in GISTIC copy-number regions of interest with a positive correlation between copy-number and expression, as well as a couple of manually identified expression events (AKT3 over-expression, PTEN and RB1 down-regulation). The algorithm identified two gene modules with significantly mutually exclusive alterations: An RB-pathway module (top), and an EGFR/PI3K pathway module. Each column of the OncoPrint represents a tumor with at least one alteration in the module. Only altered tumors are shown.

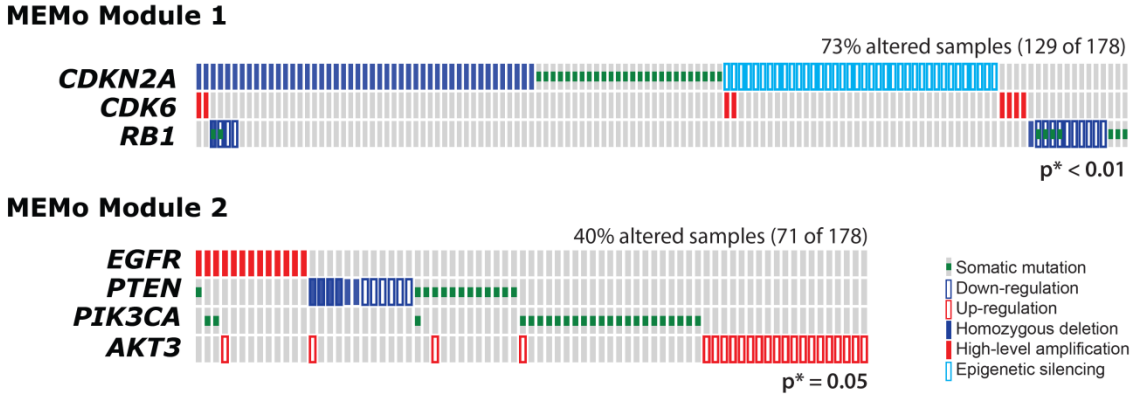
B. *RB1* and *PTEN* expression is well correlated with copy-number, but some tumors exhibit low expression in the absence of an explaining copy-number event. These down-regulated samples were included as potentially inactivating events in all pathway analyses. *AKT3* expression is not well correlated with DNA copy-number changes, but a subset of samples exhibits markedly higher expression. Samples are divided into five discrete copy-number categories, as defined by GISTIC.

C. The focal amplicon on 8p spans *FGFR1* and *WHSC1L1* and several other genes. While *FGFR1* mRNA expression is not well-correlated with copy-number (many amplified samples have low expression), *WHSC1L1* expression is strongly correlated.

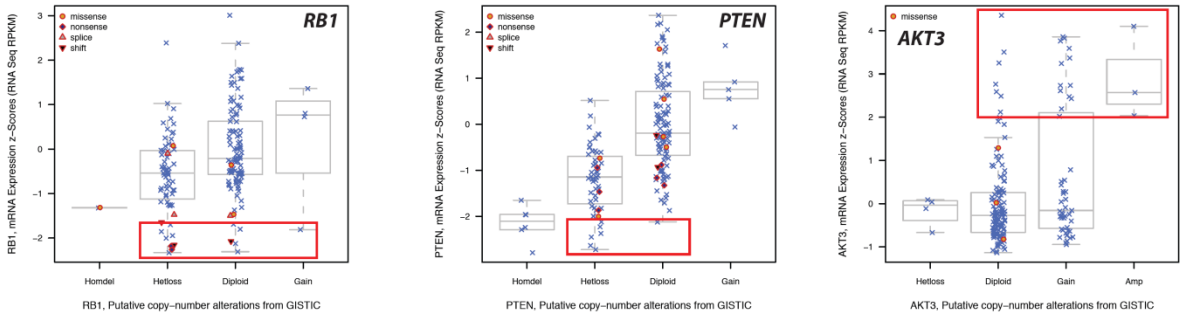
References

1. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res* **22**, 398-406, doi:gr.125567.111 [pii]10.1101/gr.125567.111 (2012).

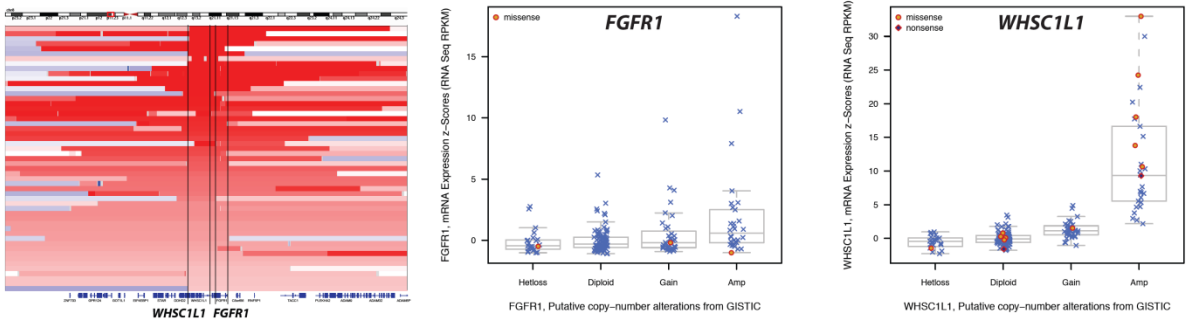
A.


MEMo Module 1
73% altered samples (129 of 178)
CDKN2A
CDK6
RB1
p* < 0.01
MEMo Module 2
40% altered samples (71 of 178)
EGFR
PTEN
PIK3CA
AKT3
p* = 0.05
Somatic mutation
Down-regulation
Up-regulation
Homozygous deletion
High-level amplification
Epigenetic silencing

B.


RB1
missense
nonsense
splice
shift
RB1, mRNA Expression z-Scores (RNA Seq RPKM)
Homdel
Hetloss
Diploid
Gain
RB1, Putative copy–number alterations from GISTIC
PTEN
PTEN, mRNA Expression z-Scores (RNA Seq RPKM)
PTEN, Putative copy–number alterations from GISTIC
AKT3
AKT3, mRNA Expression z-Scores (RNA Seq RPKM)
Amp
AKT3, Putative copy–number alterations from GISTIC

C.


WHSC1L1
FGFR1
FGFR1
FGFR1, mRNA Expression z-Scores (RNA Seq RPKM)
FGFR1, Putative copy–number alterations from GISTIC
WHSC1L1
WHSC1L1, mRNA Expression z-Scores (RNA Seq RPKM)
WHSC1L1, Putative copy–number alterations from GISTIC

Supplementary Data File S7.1 P16 Supplementary Data.

data.file.S7.1.p16.alterations.xlsx

Supplementary Data File S7.2: Data for Figure 5.

data.file.S7.2.druggable.genes.xls

Supplementary Data File S7.3: Data for Figure S7.D.1

Supplementary table to support figure 5A_Bose_02 23 2012.cleaned.xlsx

(available at https://tcga-data.nci.nih.gov/docs/publications/lusc_2012/)

Supplementary Data File S7.4: Data for Figure S7.D.1

data.file.S7.4.support.for.figure5B.xlsx

Supplementary Data File S7.5: Clinical and genomic data table.

data.file.S7.5.clinical.genomic.data.table.xlsx

Supplementary Methods S8: Batch effects analysis

We used hierarchical clustering and Principal Components Analysis (PCA) to assess batch effects in the TCGA lung squamous cell carcinoma (LUSC) data sets. Nine different data sets were analyzed: mRNA expression - genes (Agilent G4502A, and Affymetrix U133A microarrays), mRNA expression – exons (Affymetrix Human Exon 1.0 ST microarray), mRNA expression - sequencing (RNA-seq Illumina GA), miRNA expression sequencing (RNA-seq Illumina GA), DNA methylation (Infinium HM27 microarray), SNPs (GW SNP 6), and copy number data (Agilent HG CGH 415k, and CGH 1M). All of the data sets were at TCGA level 3, since that's the level on which most of the analyses in the paper are based. We assessed batch effects with respect to two variables; batch ID and Tissue Source Site (TSS).

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 at the bottom. Each color corresponded to a batch ID or a TSS. For PCA, we plotted the first four principal components, but only plots of the first two components are shown here. To make it easier to assess batch effects, we enhanced the traditional PCA plot with centroids. Points representing samples with the same batch ID (or TSS) 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 nine data sets follow. In addition, the results can be analyzed in more detail at the MD Anderson TCGA batch effects website at http://bioinformatics.mdanderson.org/tcgabatcheffects/

mRNA expression – genes (Agilent G4502A microarray)

Figures 1-3 show clustering and PCA plots for the Agilent G4502A mRNA expression platform. We noticed that batch 53 stood out slightly from the others in the PCA plot (light blue cluster in Fig. 2), and also that samples from the TSS "Cureline" stood out in the PCA plot as well (light blue cluster in Fig. 3). However, further investigation showed that batch 53 and the TSS Cureline stood out in other data sets as well, suggesting that perhaps the effect is biological in nature rather than technical. More discussions about that will follow.

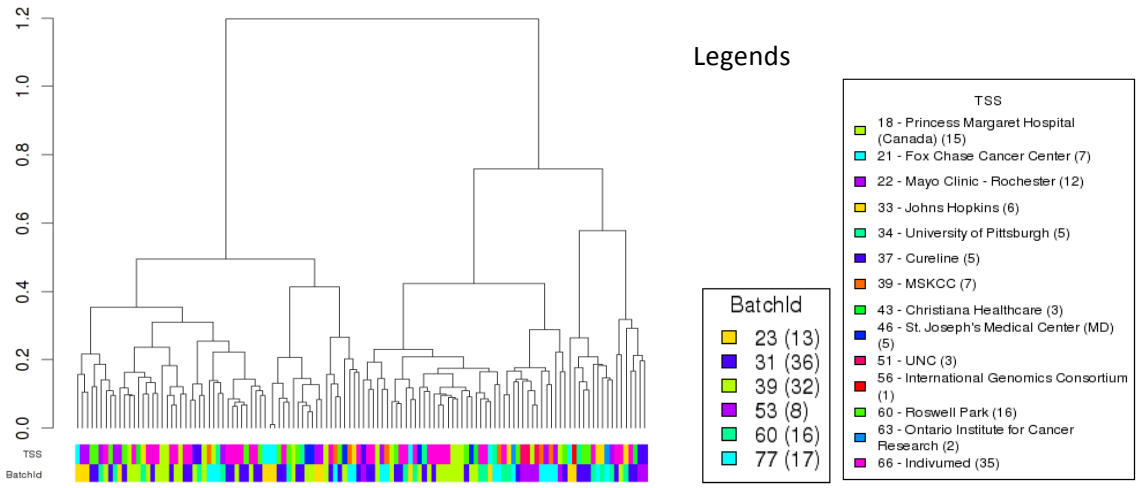


Fig. 1. Hierarchical clustering plot for mRNA expression (Agilent microarray)

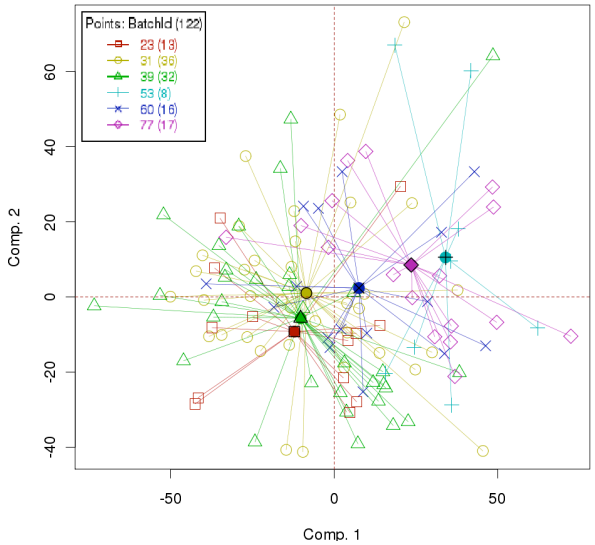


Fig. 2. PCA: First two principal components for mRNA expression (Agilent microarray), with samples connected by centroids according to batch ID.

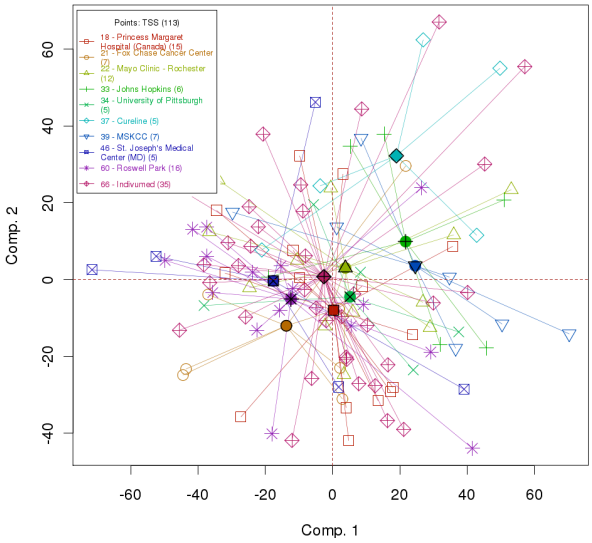


Fig. 3. PCA: First two principal components for mRNA expression (Agilent microarray), with samples connected by centroids according to TSS.

mRNA expression - genes (Affymetrix U133A)

Figures 4-6 show the clustering and PCA plots for the gene expression Affymetrix U133A microarray platform. Note once again that batch 53 in Fig. 5 and Cureline in Fig. 6 stand out like before.

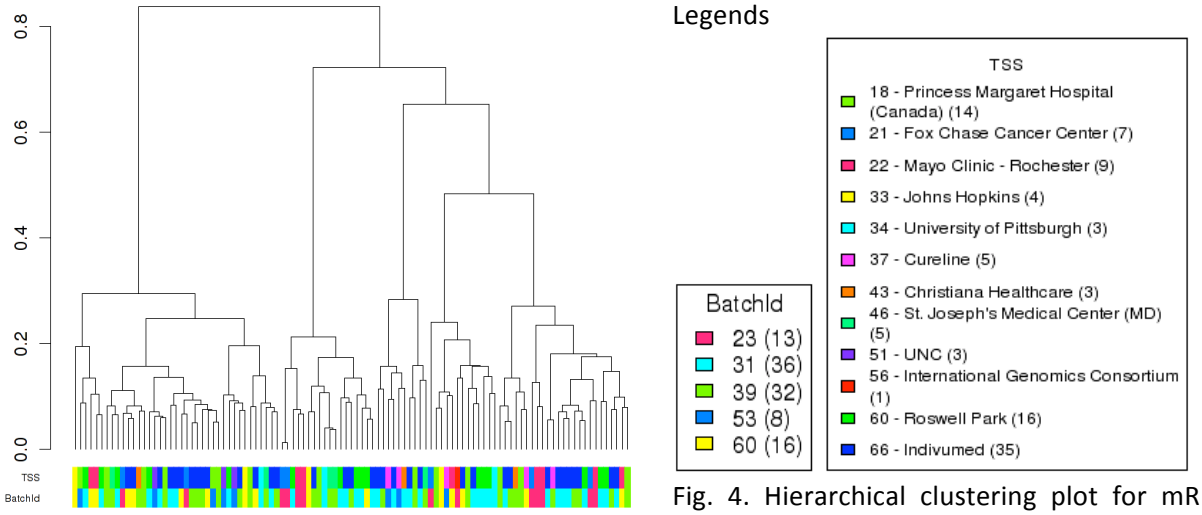


Fig. 4. Hierarchical clustering plot for mRNA expression (Affymetrix microarray)

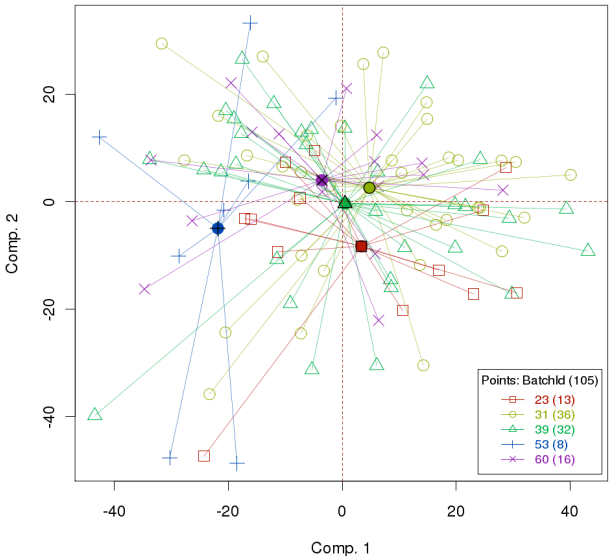


Fig. 5. PCA: First two principal components for mRNA expression (Affymetrix microarray), with samples connected by centroids according to batch ID.

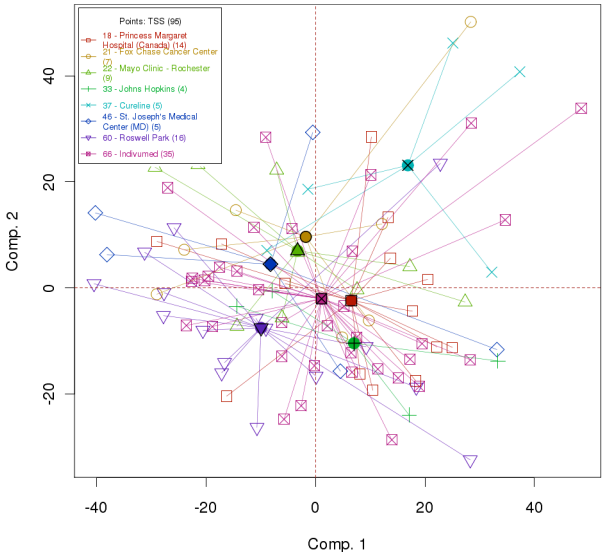


Fig. 6. PCA: First two principal components for mRNA expression (Affymetrix microarray), with samples connected by centroids according to TSS.

mRNA expression - exon (Affymetrix Human Exon 1.0 ST microarray)

Figures 7-9 show the clustering and PCA plots for the exon expression Affymetrix Human Exon 1.0 ST microarray platform. Again, batch 53 in Fig. 7 and Cureline in Fig. 9 stand out like before. Since all three expression platforms, which were processed by three different centers, show batch 53 and Cureline samples to be different, we have reason to suspect that the differences are biological in nature. Other platforms show similar results as well.

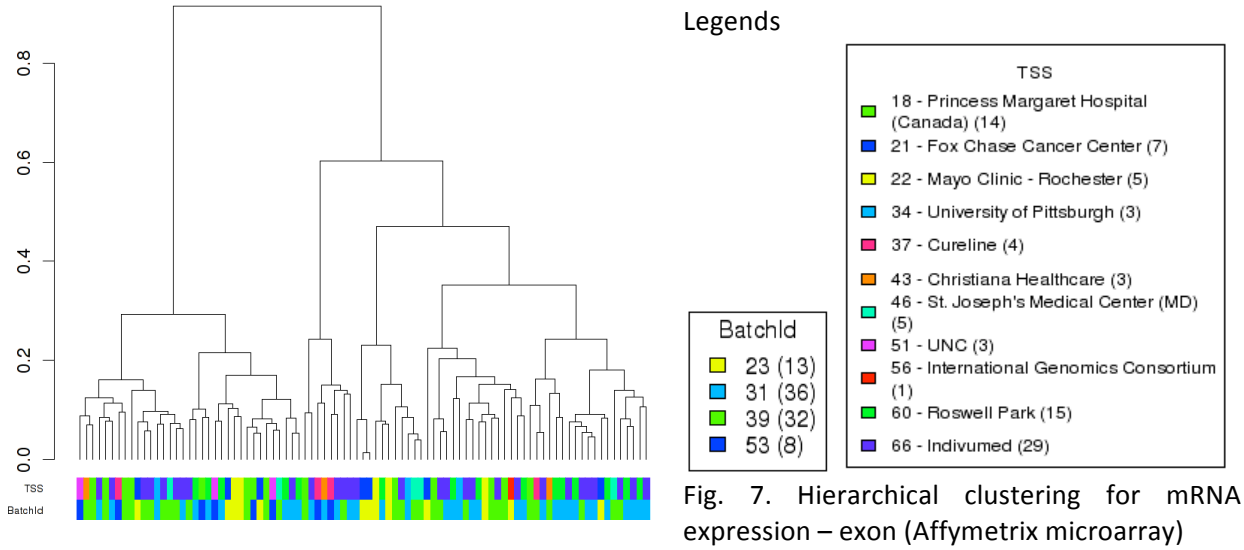


Fig. 7. Hierarchical clustering for mRNA expression – exon (Affymetrix microarray)

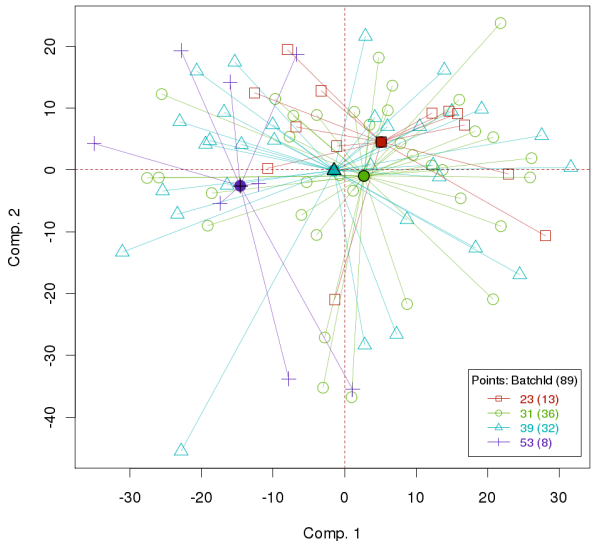


Fig. 8. PCA: First two principal components for mRNA exon expression, with samples connected by centroids according to batch ID.

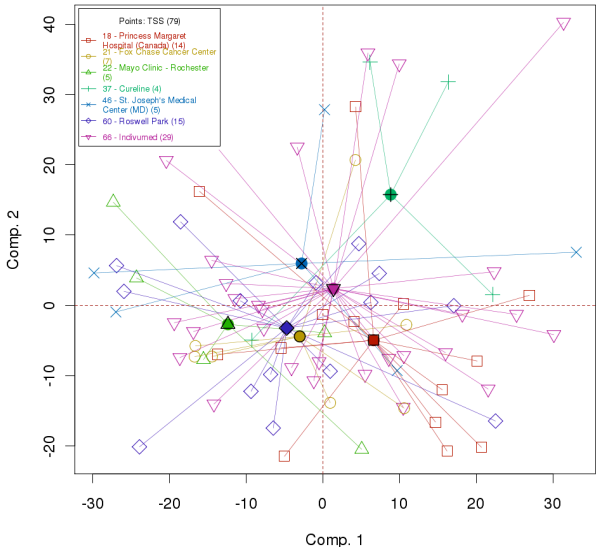


Fig. 9. PCA: First two principal components for mRNA exon expression, with samples connected by centroids according to TSS.

mRNA expression – sequencing (RNA-seq Illumina GA)

Figures 10-12 show clustering and PCA plots for the RNA-seq platform. Genes with zero values were removed and the RPKM values were $\log_2$-transformed before generating the figures. Once again, batch 53 and Cureline samples stand out in the PCA plots.

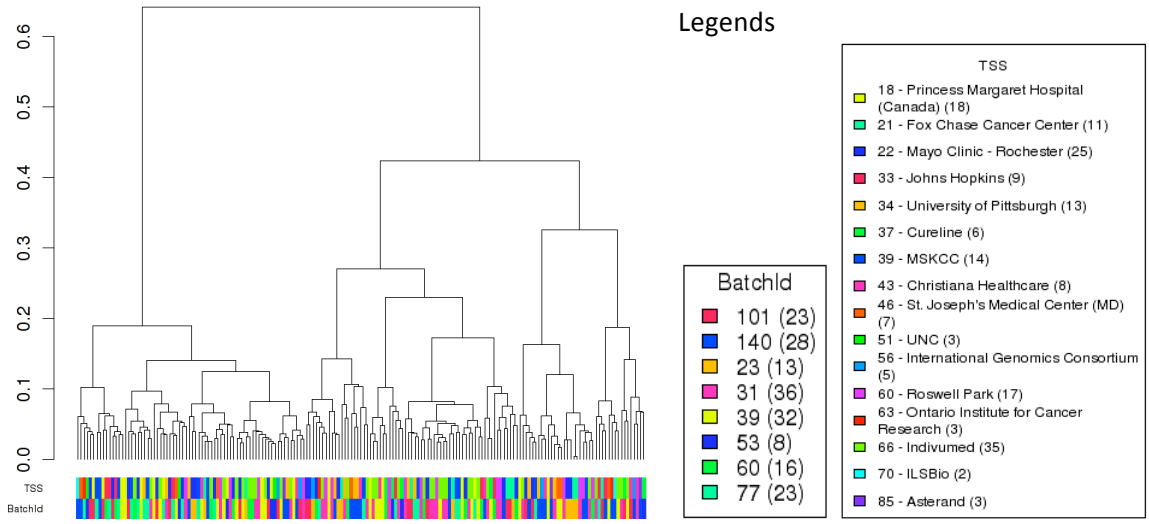


Fig. 10. Hierarchical clustering for mRNA expression from RNA-seq data

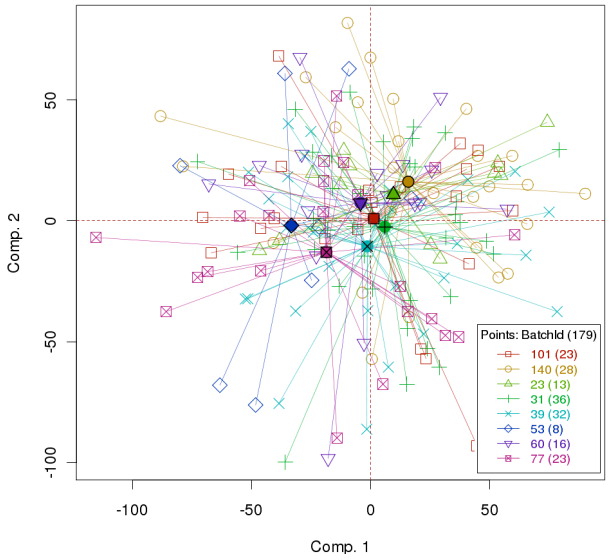


Fig. 11. PCA: First two principal components for RNA-seq, with samples connected by centroids according to batch ID.

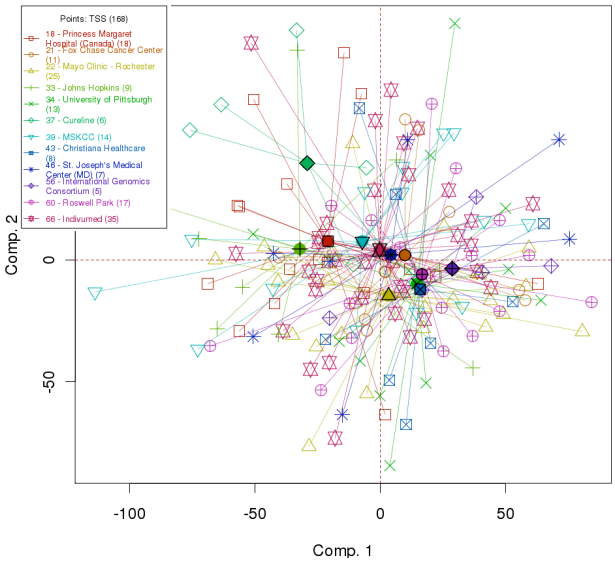


Fig. 12. PCA: First two principal components for RNA-seq, with samples connected by centroids according to TSS.

miRNA expression – sequencing (RNA-seq Illumina GA)

The following figures show clustering and PCA plots for RNA-seq miRNA data. Genes with zero values were removed and the read counts were $\log_2$-transformed before generating the figures. For miRNA, batch 53 and Cureline samples do appear distinct, but they don't stand out as much as on the other platforms. However, interestingly, the clustering plot (Fig. 13) shows that batch 39 (in light green) clusters several samples together. The PCA plot, nevertheless, fails to show batch 39 as being very different.

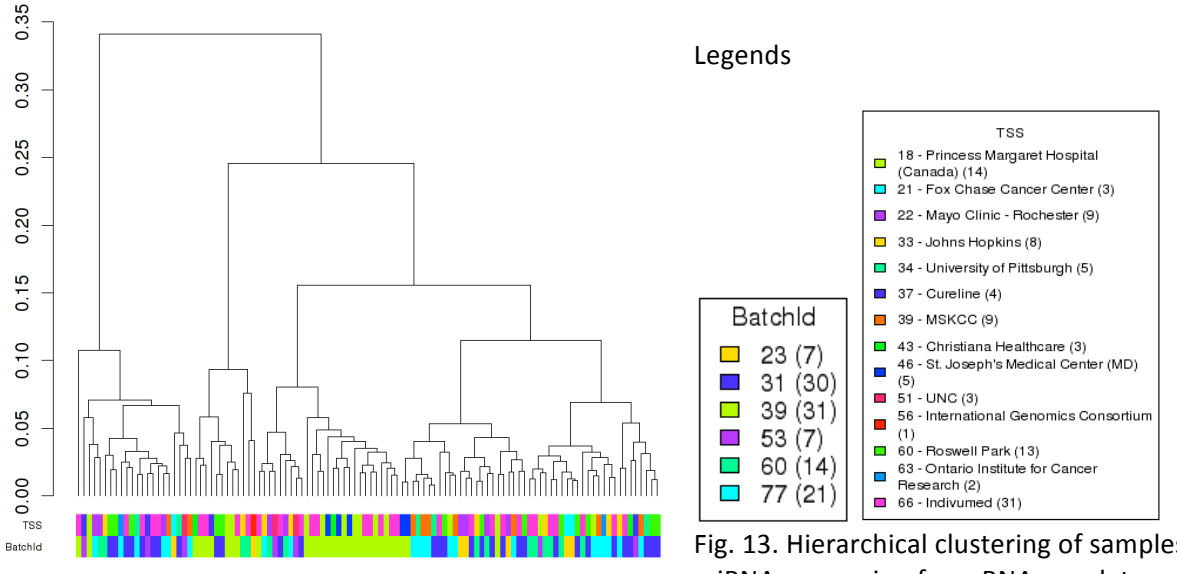


Fig. 13. Hierarchical clustering of samples for miRNA expression from RNA-seq data.

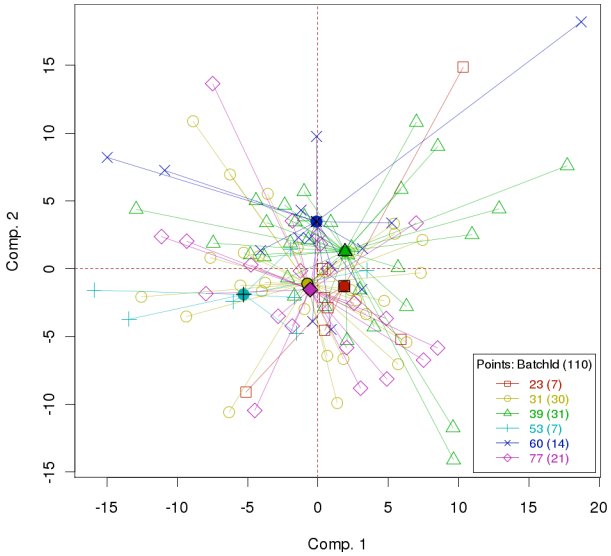


Fig. 14. PCA: First two principal components for miRNA expression from RNA-seq data, with samples connected by centroids according to batch ID.

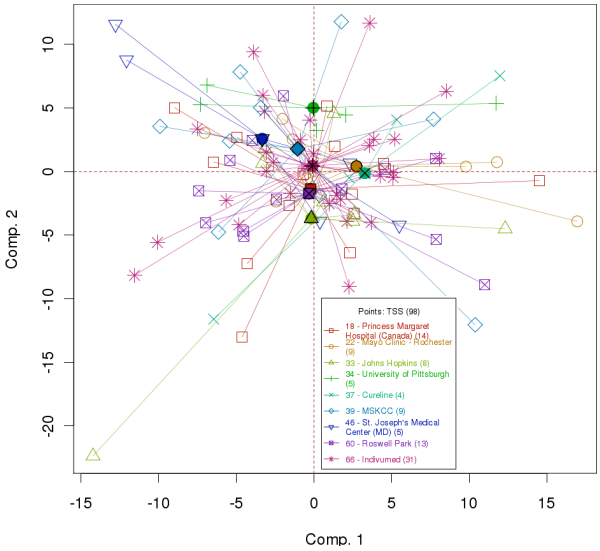


Fig. 15. PCA: First two principal components for miRNA expression from RNA-seq data, with samples connected by centroids according to TSS.

DNA Methylation (Infinium HM27 microarray)

The following figures show clustering and PCA plots for the Infinium DNA methylation HM 27 platform. Once again, batch 53 stands out from the rest, but Cureline does not stand out in the DNA methylation platform.

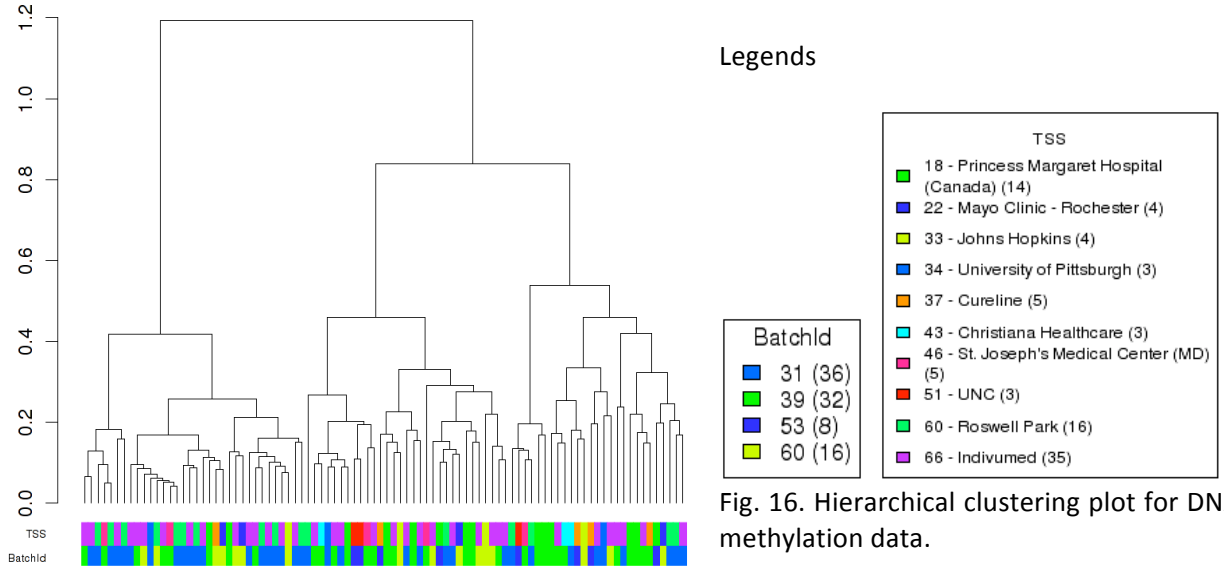


Fig. 16. Hierarchical clustering plot for DNA methylation data.

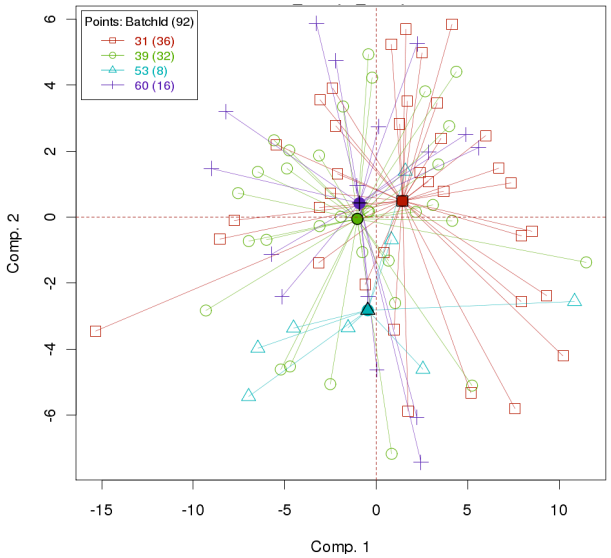


Fig. 17. PCA for DNA methylation, with samples connected by centroids according to batch ID.

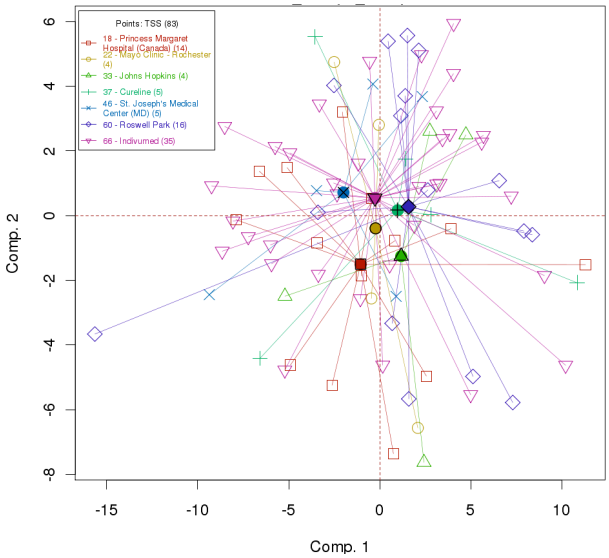


Fig. 18. PCA for DNA methylation, with samples connected by centroids according to TSS.

SNPs (GW SNP 6)

Figures 19-21 show clustering and PCA plots for the SNP platform. At level 3, the TCGA SNP data resemble copy number data when we use chromosomal segment counts (rather than actual SNPs). We mapped the chromosomal segments to genes (using build hg18) and then used them to construct the plots shown in the figures. Batch 53 only stands out slightly in Fig. 20.

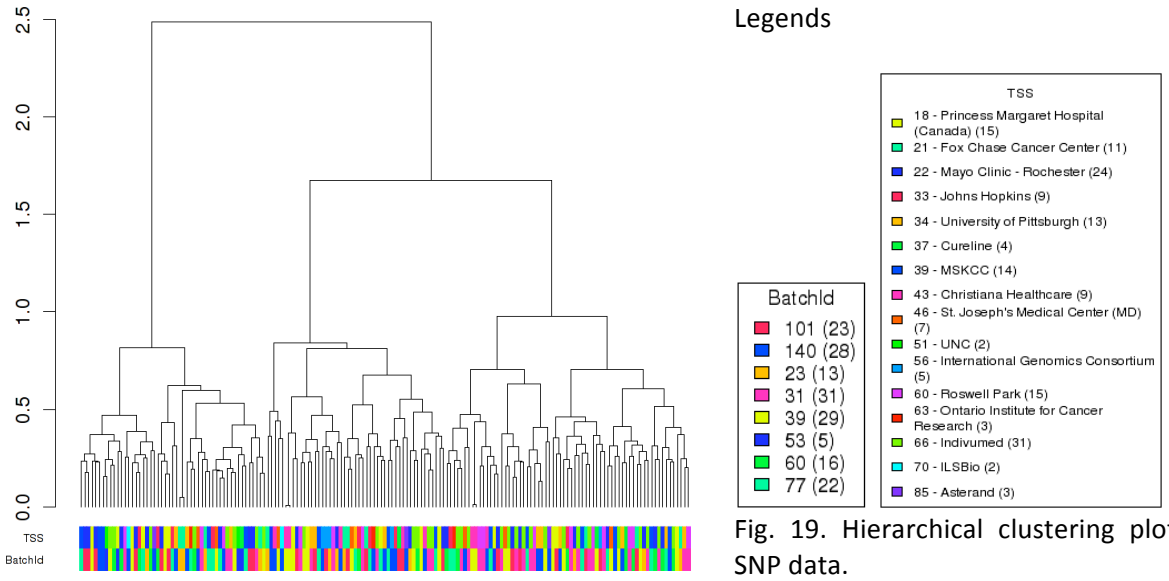


Fig. 19. Hierarchical clustering plot for SNP data.

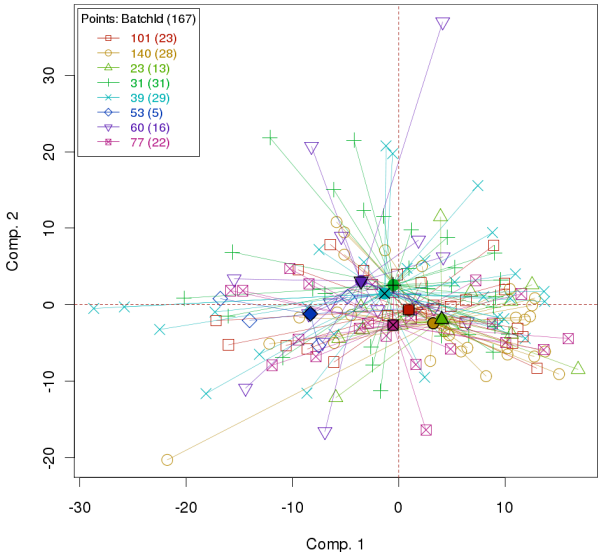


Fig. 20. PCA for SNPs, with samples connected by centroids according to batch ID.

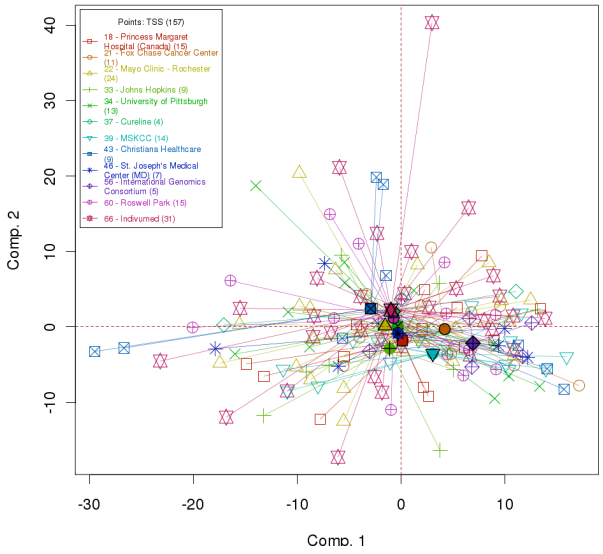


Fig. 21. PCA for SNPs, with samples connected by centroids according to TSS.

Copy Number (Agilent HG CGH 415k)

The following figures show clustering and PCA plots for the copy number Agilent HG CGH 415k platform. We mapped the chromosomal segments to genes (using build hg18) and then used them to construct the plots shown in Figs. 22-24. Batch 53 stands out a little in Fig. 23.

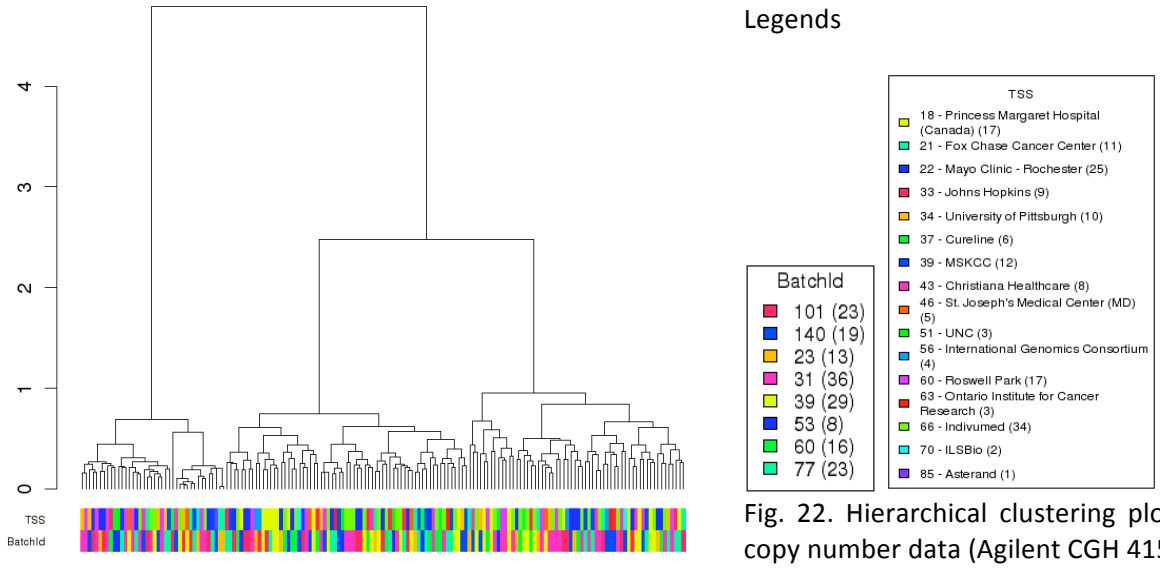


Fig. 22. Hierarchical clustering plot for copy number data (Agilent CGH 415k)

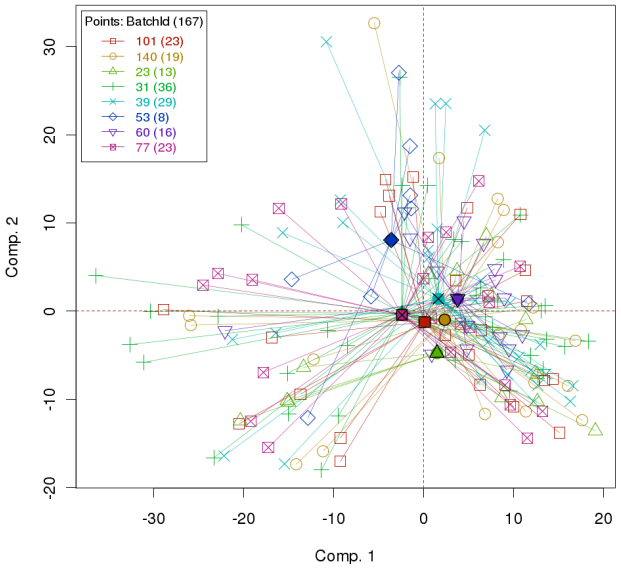


Fig. 23. PCA for copy number data (415k), with samples connected by centroids according to batch ID.

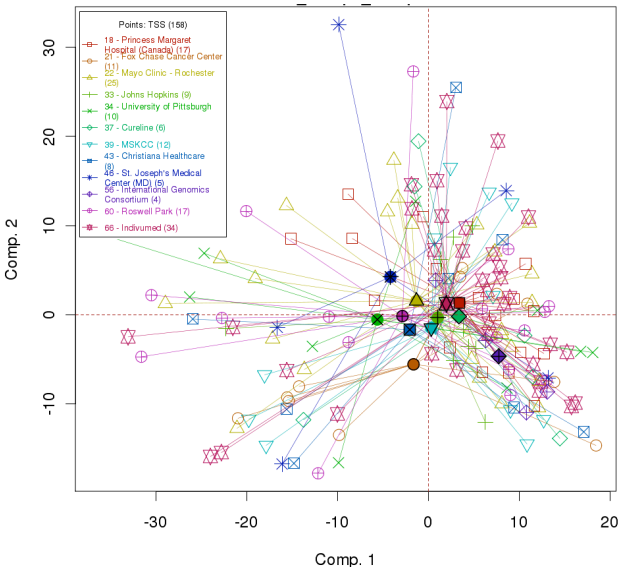
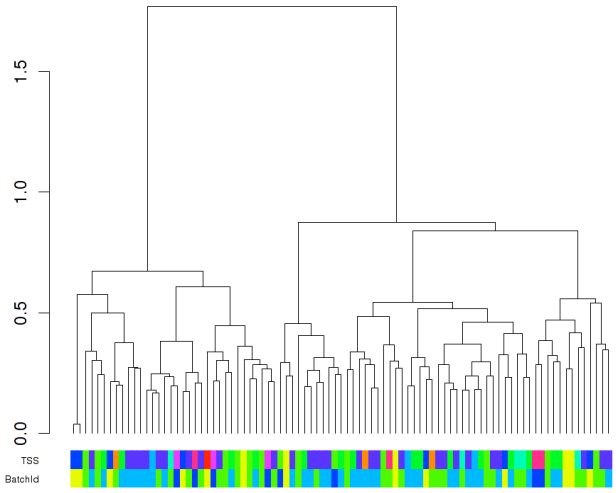


Fig. 24. PCA for copy number data (415k), with samples connected by centroids according to TSS.

Copy Number (Agilent CGH 1M)

Figures 25-27 show clustering and PCA plots for the copy number Agilent CGH 1M platform. We mapped the chromosomal segments to genes (using build hg18) and then used them to construct the plots shown in Figs. 25-27. Batch 53 stands out in Fig. 26 and Cureline stands out in Fig. 27, as in several other platforms before.



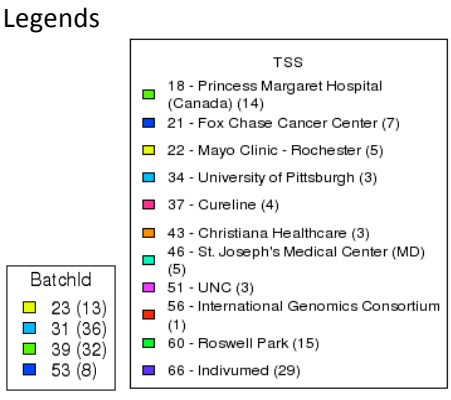


Fig. 25. Hierarchical clustering plot for copy number data (Agilent CGH 1M)

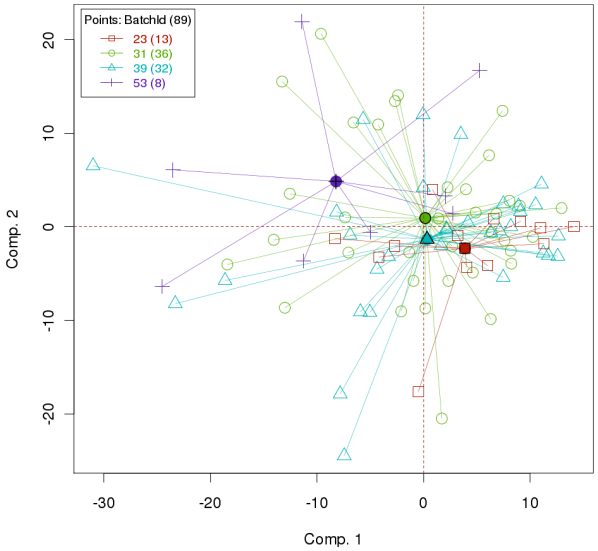


Fig. 26. PCA for copy number data (CGH 1M), with samples connected by centroids according to batch ID.

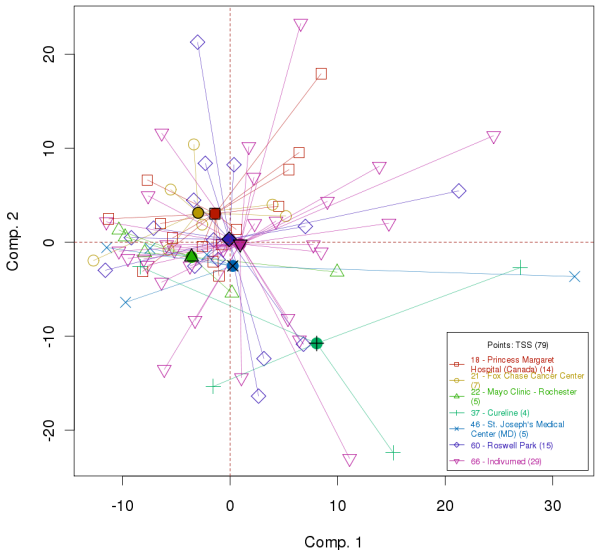


Fig. 27. PCA for copy number data (CGH 1M), with samples connected by centroids according to TSS.

Conclusions

Batch 53 appears distinct from the others in almost all of the data sets, whereas samples from the Tissue Source Site "Cureline" also stand out in most of the data sets. We suspect that the differences between batch 53 and Cureline vs. the others are biological, rather than technical, because (i) batch 53 and Cureline appear to be distinct across several different platforms, and (ii) both DNA and RNA data types show differences. Consequently, we did not try to apply computational batch effects correction to batch 53 or Cureline.

None of the other batches or Tissue Source Sites (TSS) in any data set showed consistent batch effects in both clustering and PCA algorithms. Some batches or TSSs stood out in one algorithm or the other, but not both, reducing our concern about them. Based on the above figures, we believe that technical batch effects in the data sets are reasonably small and unlikely to influence high-level analyses in a major way.