doi:10.1038/nature12965

Contents

Data File List

S1.1: Data File Clinical and Demographic Characteristics

S1.1a Data File Table Bladder Pathology Review

S2.2.1 Data File Enrichment of null mutations in all genes.

These are results of the enrichment analysis of null or truncating mutations (frame-shifts indels, nonsense, or splice-site mutations) in total 11739 genes having at least one mutation. This list is ordered by p-values to a maximum of 0.05, which were determined by Binomial test with a background null mutation rate. Columns A to F contain gene names, p-values, false discovery rates (FDR), the number of null mutations, the number of non-null mutations, and the ratio of null mutation to total mutations.

datafile.S2.2.1.BlcaNullAllGenes.xls

S2.2.2 Data File Enrichment of null mutations in epigenetic modifiers.

These are results of the enrichment analysis of null or truncating mutations in 115 chromatin remodeling genes having at least one mutation. Gene lists are ordered by p-values, which were determined by Binomial test with a background null mutation rate. Columns A to F contain gene names, p-values, false discovery rates (FDR), the number of null mutations, the number of non-null mutations, and the ratio of null mutation to non-null mutations.

datafile.S2.2.2.BlcaNullEpigenetic.xls

S2.5.1 Data File Differentially expressed genes in NFE2L2 mutants.

These are results of differentially expressed genes in NFE2L2 mutant samples. Genes are ordered by p-values to a maximum of 0.05, which were determined by Wilcoxon Ran-sum test for log2(RSEM). Columns A to I contain gene names, p-values, false discovery rates (FDR), means of log2(RSEM) in mutants, means of log2(RSEM) in non-mutant samples, the number of altered samples, the number non-altered samples, gene annotations, and fold changes.

datafile.S2.5.1.Blca.NFE2L2.AllMutations.xls

S2.5.2 Data File Differential expressed genes in NFE2L2 hotspot mutants.

These are results of differentially expressed genes in samples having NFE2L2 mutations at DLG or ETGE motifs corresponding to KEAP binding domains. Genes are ordered by p-values to a maximum of 0.05, which were determined by Wilcoxon Ran-sum test. Columns A to H contain gene names, p-values, false discovery rates (FDR), means of log2(RSEM) in mutants, means of log2(RSEM) in non-mutant samples, the number of altered samples, the number non-altered samples, gene annotations, and fold changes.

datafile.S2.5.2.Blca.NFE2L2.HotspotMutations.xls

S2.6.1 Data File Differentially expressed genes in RXRA mutants.

Differentially expressed genes were identified in RXRA mutation status. Genes are ordered by p-values to a maximum of 0.05 from Wilcoxon Ran-sum test. Columns A to I contain gene names, p-values, false discovery rates (FDR), means of log2(RSEM) in mutants, means of log2(RSEM) in non-mutant samples, the number of altered samples, the number non-altered samples, gene annotations, and fold changes.

datafile.S2.6.1.Blca.RXRA.AllMutations.xls

S2.6.2 Data File Differential expressed genes in RXRA hotspot mutants.

Differentially expressed genes were identified according to RXRA mutation at S427, in the ligand binding domain. Genes are ordered by p-values to a maximum of 0.05 from Wilcoxon Rank-sum test. Columns A to I contain gene names, p-values, false discovery rates (FDR), means of log2(RSEM) in mutants, means of log2(RSEM) in non-mutant samples, the number of altered samples, the number non-altered samples, gene annotations, and fold changes.

datafile.S2.6.1.Blca.RXRA.HotspotMutations.xls

S2.6.3 Data File DAVID pathway annotation for differentially expressed genes in RXRA hotspot mutants

(Huang W, Sherman BT, Lempicki RA, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. 2009; 4(1):44-57)**.**

datafile.S2.6.3.Blca.RXRA.DAVID.xls

S2.8.1 Date File Mutual exclusivity correlations for all genetic events in Figure 1.

This is the result of Fisher's exact tests for mutual exclusivity among significantly mutated genes or genes having focal SCNAs. Columns C and D report p-values and false discovery rates (FDR) for a pair of genes in columns A and B, respectively.

datafile.S2.8-1.BlcaMutualExclusivity.xls

S2.8.2 Date File Co-occurrence correlations for all genetic events in Figure 1.

This is the result of Fisher's exact test for co-occurrence among significantly mutated genes or genes having focal SCNAs. Columns C and D report p-values and false discovery rates (FDR) for a pair of genes in columns A and B, respectively. **S2.8-2.BlcaCoOccurrence.xls**

S6.1.1: Data File GISTIC arm-level SCNAs annotations in SNP6.0 Array.

datafile.S6.1.1.Blca.SNP.ArmLevelGisticPeaks.xls

S6.1.2: Data File GISTIC arm-level SCNAs annotations Low Pass Whole Genome.

datafile.S6.1.2.Blca.LowpassWG.ArmLevelGisticPeaks.xls

S6.2.1: Data File GISTIC amplification peak annotations in SNP6.0 Array.

datafile.S6.2.1.Blca.SNP.FocalAmpGisticPeaks.xls

S6.2.2: Data File GISTIC amplification peak annotations in Low Pass Whole Genome.

datafile.S6.2.2.Blca.SNP.FocalAmpGisticPeaks.xls

S6.3.1: Data File GISTIC deletion peak annotations in SNP6.0 Array.

datafile.S6.3.1.Blca.SNP.FocalDelGisticPeaks.xls

S6.3.2: Data File GISTIC deletion peak annotations in Low Pass Whole Genome.

datafile.S6.3.2.Blca.LowpassWG.FocalDelGisticPeaks.xls

S8.2. Data File Cytoscape session of the TieDIE solution for the significantly mutated gene

S8.2.1 . Data File Cytoscape version 3.01 CYS session of the TieDIE

S9.1 Data File Coverage Table

Data File S12.1 APOBEC. BLCA samples data underlying the graphical representation in Figures S12.1 and S12.2. Spreadsheet contains the number of identified mutation clusters and tabulated metrics used in calculating the representation of TCW mutagenesis in each tumor sample. The mRNA expression of APOBEC1, APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3DE, APOBEC3F, APOBEC3G, and APOBEC3H relative to TBP as determined by RNA-seq counts is provided for comparison. A detailed description of the values in each column is provided in an accompanying *Readme for data files S12.1 and S12.2* file

Data File S12.2 APOBEC. Matched normal samples data underlying the graphical representation in Figure S12.1. Spreadsheet contains the mRNA expression data for APOBEC1, APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3DE, APOBEC3F, APOBEC3G, and APOBEC3H relative to TBP within 16 analyzed matched normal samples. A detailed description of the values in each column is provided in an accompanying *Readme for data files S12.1 and S12.2* file.

Leaders: Seth Lerner *slerner@bcm.edu and* ***Hikmat Al-Ahmadie*** *alahmadh@mskcc.org Team Members: Jay Bowen, Bogden Czerniak, Donna Hansel, Tara Lichtenberg, Brina Robinson and Jonathan Rosenberg,*

S1 Biospecimen collection and clinical data:

Text S1 Biospecimen collection and clinical data:

Sample inclusion criteria

Biospecimens were collected from patients diagnosed with muscle-invasive urothelial carcinoma undergoing surgical resection with either transurethral resection or radical cystectomy. No patient had received prior chemotherapy or radiotherapy for their disease. Prior intravesical Bacille Calmette Guerin (BCG) was allowed but no intravesical chemotherapy. Institutional review boards at each tissue source site reviewed protocols and consent documentation and approved submission of cases to TCGA. Cases were staged according to the American Joint Committee on Cancer (AJCC) staging system. Each frozen primary tumor specimen had a companion normal tissue specimen which could be blood/blood components (including DNA extracted at the tissue source site), adjacent normal tissue taken from greater than 2 cm from the tumor, or both. Specimens were shipped overnight from 19 tissue source sites (TSS) using a cryoport that maintained an average temperature of less than -180°C. Each tumor and adjacent normal tissue specimen (if available) were embedded in optimal cutting temperature (OCT) medium and a histologic section was obtained for review. Each H&E stained case was reviewed by a board-certified pathologist to confirm that the tumor specimen was histologically consistent with urothelial carcinoma and the adjacent normal specimen contained no tumor cells. The divergent histologic carcinoma component of the cancer was < 50%. The tumor sections were required to contain an average of 60% tumor cell nuclei with equal to or less than 20% necrosis for inclusion in the study per TCGA protocol requirements.

Table S1.1 Tissue Source Sites

Tissue Source Sites	
BL	Christiana Healthcare
BT	University of Pittsburgh
C4	Indivumed
CF	ILSBio
CU	UNC
DK	Memorial Sloan Kettering
E5	Roswell Park
E7	Asterand
FD	BLN - University Of Chicago
FJ	BLN - Baylor
FT	BLN - University of Miami
G2	MD Anderson
GC	International Genomics Consortium
GD	ABS - IUPUI
GU	BLN - UT Southwestern Medical Center at Dallas
GV	BLN - Cleveland Clinic
H4	Medical College of Georgia
HQ	Ontario Institute for Cancer Research (OICR)
K4	ABS - Lahey Clinic

Sample Processing

RNA and DNA were extracted from tumor and adjacent normal tissue specimens using a modification of the DNA/RNA AllPrep kit (Qiagen). The flow-through from the Qiagen DNA column was processed using a *mir*Vana miRNA Isolation Kit (Ambion). This latter step generated RNA preparations that included RNA <200 nt suitable for miRNA analysis. DNA was extracted from blood using the QiaAmp blood midi kit (Qiagen).

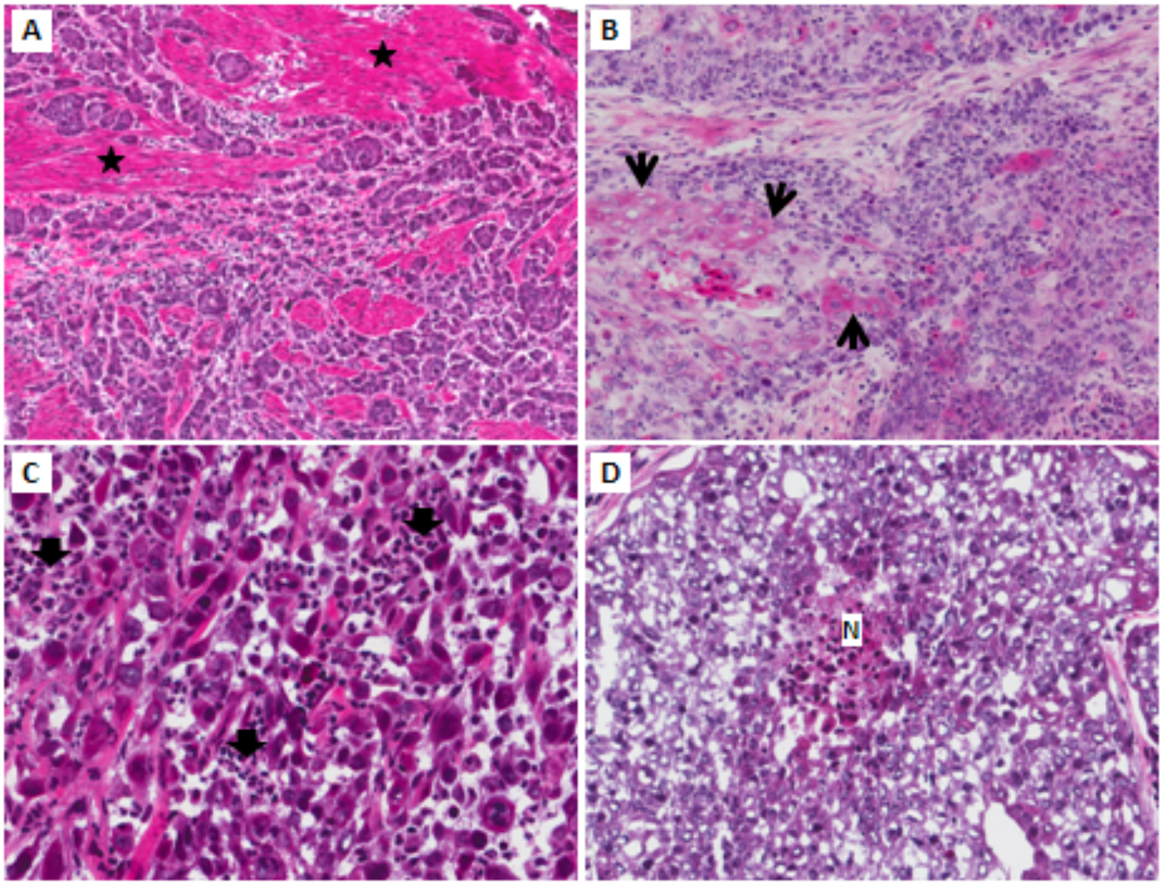
Each specimen was quantified by measuring Abs_{260} with a UV spectrophotometer or by PicoGreen assay. DNA specimens were resolved by 1% agarose gel electrophoresis to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpFISTR Identifiler (Applied Biosystems) was utilized to verify tumor DNA and germline DNA were derived from the same patient. Five hundred nanograms of each tumor and normal DNA were sent to Qiagen for REPLI-g whole genome amplification using a 100 µg reaction scale. Only specimens yielding a minimum of

6.9 µg of tumor DNA, 5.15 µg RNA, and 4.9 µg of germline DNA were included in this study. RNA was analyzed via the RNA6000 nano assay (Agilent) for determination of an RNA Integrity Number (RIN), and only the cases with RIN >7.0 were included in this study. At the time of the data freeze, 345 bladder urothelial carcinoma cases were received by the BCR and 57% passed quality control. A total of 131 cases were included in the data freeze that had complete information from all of the data types.

Methods S1.1a: Pathology review

All cases were subjected to a detailed pathologic review by 4 genitourinary pathologists to confirm the selection criteria and to assess the samples for additional histopathologic features such as the presence of divergent differentiation and its extent, the pattern of invasion and the presence of associated inflammation. In all cases, the amount of divergent histology that was allowed was less than 50% of the tumor. The most common divergent histology in this cohort was squamous differentiation, which was present in 19 of 131 cases (Figure S1.1a). The detailed review resulted in the exclusion of 7 cases due to the presence of divergent histology in >50% of the tumor (3 cases for small cell/neuroendocrine differentiation and 4 cases for squamous differentiation).

Figure S1.1a Pathology – mixed histology



A. Urothelial carcinoma, invasive in the muscularis propria (*)

B. Urothelial carcinoma with squamous differentiation (arrows)

C. Urothelial carcinoma with associated inflammation (block arrows)

D. Urothelial carcinoma with focal tumor necrosis (N)

*Team Leader: **Jaegil Kim** jaegil@broadinstitute.org Team Members: David Kwiatkowski, Jonathan Rosenberg, and Andrew Cherniack*

S2.1: DNA sequencing:

Text S2.1: DNA sequencing

1. 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 [1-4]:

(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 quality control, 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 [1-4].

-

1. Quality control – confirms identity of 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[4] – 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.

-

2. Mutation significance analysis

The statistical significance of mutation frequency in each gene was determined using the algorithm MutSig v1.5 [5] (Lawrence et al., in press). The MutSig algorithm works with an aggregated list of mutations across the entire patient set, and estimates the background mutation rate. The p and q values for a certain gene are determined for the mutation rate observed in that gene in relation to the background model. MutSig uses various factors to accurately estimate the background mutation rate, taking into account the background mutation rates of different mutation categories (i.e. transitions or transversions in different sequence contexts), the non-synonymous to synonymous mutation ratio for each gene, as well as the fact that different samples have different background mutation rates. It then uses convolutions of binomial distributions to calculate the p-value for each gene, which represents the probability that we observe a certain configuration of mutations in a gene by chance, given the background model. Finally, it corrects for multiple hypotheses by calculating a q-value (False Discovery Rate) for each gene.

-

To improve our statistical power of identifying potential "drivers" genes we considered only mutations that had an allele fraction (AF) >= 0.05 to remove potential subclonal mutations. We also performed two independent significance analyses to augment our power to identify genes with potentially significant roles in cancer. 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 (Supplementary Table S2.1.2). 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 (Supplementary Table S2.1.3).

-

3. Differential gene expression depending on mutation status

To identify differentially expressed genes (DEGs) depending on the mutation status of an individual gene we compared log2(RSEM) (RSEM: RNA-Seq by Expectation Maximization) values of 20,502 genes between mutant samples and non-mutant samples. Wilcoxon rank sum test was applied to compute p-values and multiple hypotheses were corrected by calculating q-values (False Discovery Rate) for each gene. To remove the effects of genes expressed at low levels we only considered only genes having RSEM values in at least 90% mutant and non-mutant samples. We also also restricted this comparison of expression effects by gene mutation to specific hotspot or recurrent mutations in NFE2L2 and RXRA. The gene annotation of differentially expressed genes was done from DAVID analysis [6].

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. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, and Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology. 2013 Mar 31(3):213-219.

5. Lawrence MS et al, Mutational heterogeneity in cancer and the search for new cancer genes. Nature. 2013 In press.

6. Huang da W, Sherman BT, Lempicki RA, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. 2009; 4(1):44-57.

4. Exome mutation validation

To validate mutations identified by exome sequencing, we used three approaches in parallel. First we performed targeted PCR followed by deep-coverage sequencing using Fluidigm and Illumina technology, for 753 non-silent mutations in 53 genes: all genes identified as being significantly mutated *(TP53, MLL2, ARID1A, KDM6A, PIK3CA, EP300, CDKN1A, RB1, ERCC2, FGFR3, STAG2, ERBB3, FBXW7, RXRA, ELF3, NFE2L2, TSC1, KLF5, TXNIP, CDKN2A, FOXQ1, RHOB, FOXA1, PAIP1, BTG2, HRAS, ZFP36L1, RHOA, CCND3)*; multiple additional known cancer genes *(NF1, PTEN, PIK3R1, ERBB2, ATM, CTNNB1, APC);* and multiple additional chromatin modifying genes *(MLL, MLL3, MLL4, CREBBP, NCOR1, NCOR2, SRCAP, SETD2, CHD7, NSD1, DOT1L, SMARCC1, SMARCC2, SMARCA4, SMARCA2, ARID2, ARID1B)*. Second and third, we compared mutations identified by Mutect analysis of whole exome sequencing with results obtained from RNA-Seq data for 123 samples, and whole genome (WGS) on 18 samples.

For all three of these approaches to validation, we removed from consideration those sites where the sequence read depth in the validation data set was not sufficient to enable verification of the variant called in the exome data. Powered mutation sites in the validation data were defined as sites where there was > 90% probability to observe at least two variant reads if the mutation call is true, given the allelic fraction of that variant observed in the exome data and the number of reads at that site in the validation data set. We considered two reads with the variant allele in the validation data as sufficient evidence to confirm the call, but also considered a higher threshold of read number for validation which had relatively minor effects on validation rate (see further below). Note that the identification of powered sites was performed independently for each validation approach, targeted resequencing, RNA-Seq, and WGS. Note as well that even if all the original mutation calls are true, in the case in which the sequence read depth combined with the variant allele frequency gives a power of 90%, one would expect that the validation rate would be only 90%. In practice in most instances the power is much greater than 90%, so that the expected validation rate would be much higher.

A. Targeted re-sequencing using Fluidigm

We performed targeted PCR using a microfluidic PCR platform (Fluidigm Access Array, Hollants S. et al, Clinical Chemistry (2012) doi:10.1373/clinchem.2011.173963) followed by next-gen sequencing using Illumina yielding mean 465x coverage, for 753 non-silent mutations in 53 genes on 129 samples (one sample failed QC).

Considering only the powered sites as above, 612 single nucleotide variant (SNV) mutations out of 618 (99%) were validated (Supplemental Figure S2.11.1). The validation rate in the significantly mutated genes was 368 out of 371 (99.2%) SNVs (Supplemental Figure S2.11.2). All 101 (100%) powered indel mutations were validated (Supplemental Figure S2.11.3). Combining these two, the overall validation rate by targeted re-sequencing was 713 of 719, or 99.2%.

We also examined the validation rate as a function of the number of variant reads (N) required for validation (Supplemental Table S.2.11.1-2). This required re-calculation of the number of variants for which there was $\geq$ 90% power of detection for each N. The validation rate remained > 97% for increasing N up to 50, indicating the robustness of the validation. Note that one expects that the validation rate will fall as N increases due to an increasing number of variant sites having power closer to the 90% required.

B. Exome mutation validation using RNA-Seq and WGS data

We again considered only those variants for which there was sufficient power (> 90%) for detection in these alternative data sets, based on allele frequency of the somatic variant allele and read depth at that site. We again considered observation of at least two reads containing the variant nucleotide as sufficient for confirmation.

The validation rate for all SNVs in RNA-Seq data for 123 samples was 9857 of 10629 (92.7%) (Supplemental Figure S2.11.4 bottom) and 275 of 279 (98.6%) for variants in significantly mutated genes (Supplemental Figure S2.11.4 top). Note that we expect that some SNVs, especially nonsense alleles, may be underrepresented in RNA-Seq data due to nonsense-mediated decay.

The validation rate for all non-silent SNVs in WGS data for 18 samples was 3221 of 3259 (99.2%) (Supplemental Figure S2.11.5 bottom), and 51 of 52 (98%) (Supplemental Figure S2.11.5 top).

C. Combined assessment of exome validation

Considering validation by any one of these three approaches as sufficient for validation, the overall validation rate was 741 of 745 (99.5%) (Supplemental Figure S2.11.6). 248 mutations were validated on one platform, 334 mutations were validated on two platforms, and 55 mutations were validated on all three platforms.

5. ERCC2 mutations status and mutation rate

While ERCC2 is one of the significantly mutated genes in this cohort, the correlation between ERCC2 mutation status and overall mutation rate was only marginally significant (Supplemental Figure S2.12(a), P = 0.0278). We also considered the possibility that the effect of ERCC2 mutation on mutation rate might be masked by the strong mutagenic effect of APOBEC cytidine deaminase signature (Supplemental Figure S.2.12(b), P= 0.00275). Hence, we analyzed samples with low APOBEC signature separately (n = 16), to examine the potential effect of ERCC2 mutation (Supplemental Figure S.2.12(c)). In that small set of samples there was a trend toward association between ERCC2 mutation and mutation rate (P=0.087), but the small number of samples with ERCC2 mutations (3 of 16) severely compromised our power.

Table S2.1.1 Categories of mutation types.

category	n	N	rate	rate_per_mb	relative_rate
Tp*C->(T/G)	14993	503254008	2.98E-05	29.79	3.88
Tp*C->A	1109	503254008	2.20E-06	2.20	0.29
(A/C/G)p*C->mut	6057	1431662021	4.23E-06	4.23	0.55
A->mut	3029	1858868904	1.63E-06	1.63	0.21
indel+null	3902	3793784933	1.03E-06	1.03	0.13
double_null	49	3793784933	1.29E-08	0.01	0.00
Total	29139	3793784933	7.68E-06	7.68	1.00

Table S2.1.2 Significantly Mutated Genes (SMGs) identified by MutSig v1.5.

This is the list of significantly mutated genes (SMGs) list from MutSig v1.5, performed across 130 BLCA tumor-normal pairs. Gene lists are ordered by q value to a maximum of 0.1. Columns A to G contain ranks, gene names, the number of non-silent mutations (n), the number of altered samples (npat), the number of unique sites having nonsilent mutations, the number of silent mutations, the number of null + Indel mutations (null + indel). The "null" includes nonsense, frame-shift, and splice site mutations. The remaining columns report p-values, and false discovery rates (FDR) or q-values for each gene tested in MutSig v1.5.

rank	gene	n	npat	nsite	nsil	null+indel	p	q
1	ARID1A	39	33	37	2	27	3.00E-15	1.81E-11
2	CDKN1A	18	18	17	0	13	3.11E-15	1.81E-11
3	TP53	75	64	50	1	18	3.33E-15	1.81E-11
4	RB1	19	17	17	0	12	4.33E-15	1.81E-11
5	KDM6A	32	31	26	2	27	5.00E-15	1.81E-11
6	PIK3CA	26	26	11	1	0	7.66E-15	2.31E-11
7	ELF3	15	11	14	0	8	2.91E-11	7.53E-08
8	ERCC2	16	16	13	2	0	7.36E-10	1.66E-06
9	MLL2	39	35	39	5	17	2.46E-09	4.95E-06
10	FBXW7	16	13	12	0	6	5.79E-09	1.05E-05
11	FOXQ1	7	7	4	1	4	1.54E-08	2.54E-05
12	NFE2L2	12	11	9	0	0	2.49E-08	3.75E-05
13	FGFR3	21	16	11	3	1	7.98E-07	0.00111
14	TXNIP	10	9	10	1	5	1.79E-06	0.00231
15	STAG2	14	14	13	3	12	2.04E-06	0.00246
16	CDKN2A	8	7	8	0	4	2.26E-06	0.00256
17	RHOB	7	7	6	1	0	3.33E-06	0.00337
18	BTG2	6	6	6	0	1	3.35E-06	0.00337
19	FOXA1	7	7	7	1	5	3.77E-06	0.00359
20	RXRA	12	12	6	2	0	4.86E-06	0.0044
21	HORMAD1	8	8	8	1	4	6.56E-06	0.00565
22	EP300	25	20	25	3	8	1.01E-05	0.0083
23	KLF5	11	10	10	2	3	1.85E-05	0.0146
24	GPC5	8	8	8	1	2	2.12E-05	0.016
25	HRAS	6	6	5	0	0	2.80E-05	0.0203
26	ERBB3	14	14	10	2	0	4.20E-05	0.0284

27	ZFP36L1	6	6	6	0	3	4.24E-05	0.0284
28	RHOA	5	5	5	1	0	6.87E-05	0.0444
29	PAIP1	7	7	7	0	1	7.88E-05	0.0492
30	ZFR2	6	6	6	0	0	0.000118	0.0712
31	TSC1	11	11	11	0	8	0.000132	0.0771
32	CCND3	5	5	5	0	1	0.00014	0.0791

Table S2.1.3 Significantly Mutated Genes in COSMIC territory.

This is the list of SMGs in the COSMIC territory from MutSig v1.5. Gene lists are ordered by q value to a maximum of 0.1. Columns A to E contain ranks, gene names, the number of nonsilent mutations, p-values, and FDR qvalues.

rank	gene	n	p	q
1	FGFR3	21	2.57E-13	1.16E-09
2	FBXW7	16	6.25E-13	1.41E-09
3	PIK3CA	26	9.57E-13	1.44E-09
4	TP53	75	1.42E-12	1.51E-09
5	RB1	19	1.67E-12	1.51E-09
6	ERBB2	11	9.25E-10	6.97E-07
7	CDKN2A	8	1.80E-09	1.16E-06
8	HRAS	6	5.30E-09	3.00E-06
9	ATM	19	8.40E-09	4.22E-06
10	ERBB3	14	3.55E-08	1.61E-05
11	CTNNB1	3	1.58E-04	6.52E-02

Table S2.1.4 Enrichments of mutations to CNMF clusters.

The enrichment of mutations in SMGs to the CNMF clusters in Fig.1 was examined using Chi-square test. Columns A to D contain rgene names, p-values, FDR or q-values, and the number of events. With the criterion of FDR < 0.1 the red cluster in Fig1 had an enrichment of MLL2 mutations, the blue cluster had a significant enrichment of FGFR3, NFE2L2, STAG2, and TSC1 mutations, and the green cluster had an enrichment of TP53, RB1, and KDM6A mutations.

gene	p	q	n.mutation
TP53	7.45E-11	2.16E-09	63
FGFR3	2.25E-05	0.000217344	16
NFE2L2	2.21E-05	0.000217344	11
RB1	0.000220832	0.001601034	17
MLL2	0.000521902	0.003027031	33
STAG2	0.002013929	0.00973399	14
KDM6A	0.004321778	0.017904509	31
TSC1	0.018795631	0.068134161	11

FOXQ1	0.046621768	0.150225696	7
RXRA	0.097777836	0.260778022	12
TXNIP	0.098915802	0.260778022	8
CDKN2A	0.115749745	0.27972855	7
EP300	0.200911287	0.40027548	20
ERCC2	0.245752765	0.40027548	16
ERBB3	0.235257117	0.40027548	14
FOXA1	0.24844685	0.40027548	7
PAIP1	0.195286241	0.40027548	7
ZFP36L1	0.24802928	0.40027548	6
CDKN1A	0.340800656	0.520169422	18
RHOB	0.373170207	0.536843328	7
CCND3	0.388748617	0.536843328	5
PIK3CA	0.413245536	0.544732752	26
ARID1A	0.444189236	0.555685353	33
KLF5	0.459877533	0.555685353	10
RHOA	0.6084785	0.70583506	5
FBXW7	0.651638654	0.726827729	12
BTG2	0.83189671	0.861607306	6
HRAS	0.83189671	0.861607306	6
ELF3	0.972906998	0.972906998	11

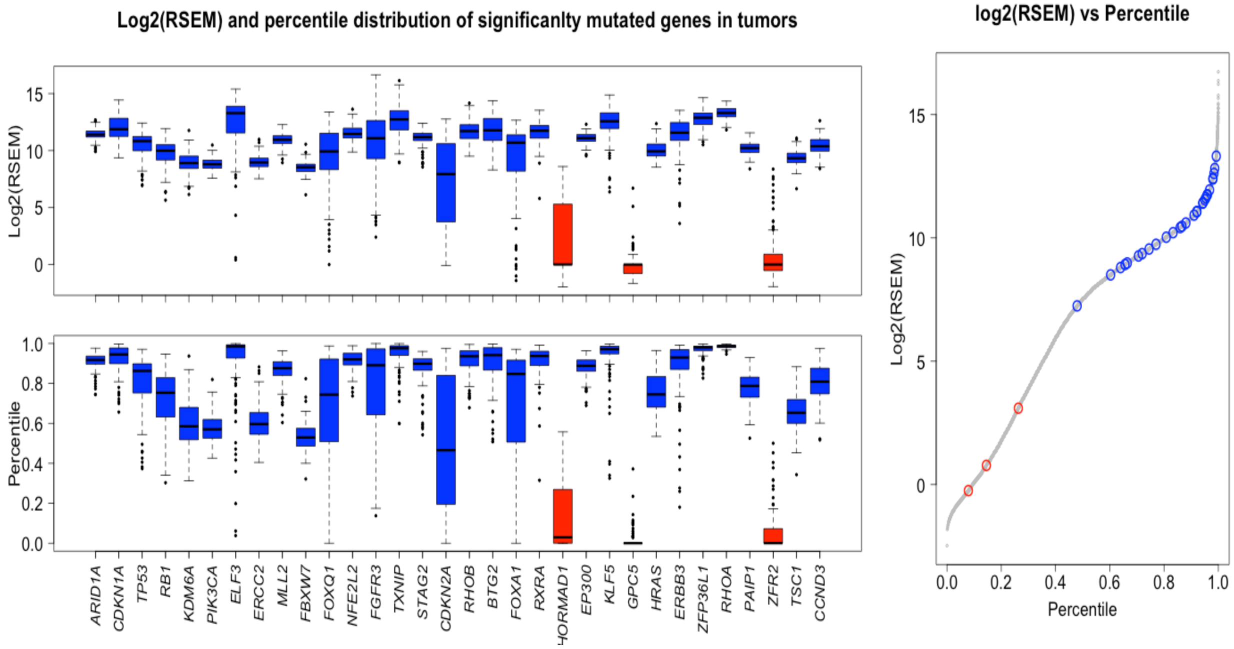
Table S2.1.5 Enrichments of focal SCNAs to CNMF clusters.

The enrichment of focal SNCAs to the CNMF clusters in Fig.1 was examined using Chi-square test. Columns A to D contain gene names, p-values, FDR or q-values, and the number of events. With the criterion of FDR < 0.1 the red cluster in Fig1 had a significant enrichment of focal SCNAs in YWHAZ, PPARG, YAP1, MYC, and PVRL4, the blue cluster had an enrichment of focal SCNAs in CDKN2A and MDM2, the green cluster had an enrichment of focal SCNAs in E2F3 and CCNE1.

gene	p	q	n.copy
CDKN2A	4.29E-15	8.15E-14	60
YWHAZ	2.55E-11	2.42E-10	28
PPARG	6.26E-07	3.97E-06	22
MYC	1.53E-05	7.27E-05	17
PVRL4	4.44E-05	0.000168741	24
E2F3	0.00484999	0.015358303	24
YAP1	0.012500399	0.033929654	5
MDM2	0.035185534	0.083565644	12
CCNE1	0.039640173	0.083684809	15
EGFR	0.122975957	0.233654318	14
ERBB2	0.168532935	0.271608132	9
MYCL1	0.173251523	0.271608132	8
FGFR3	0.185837143	0.271608132	4

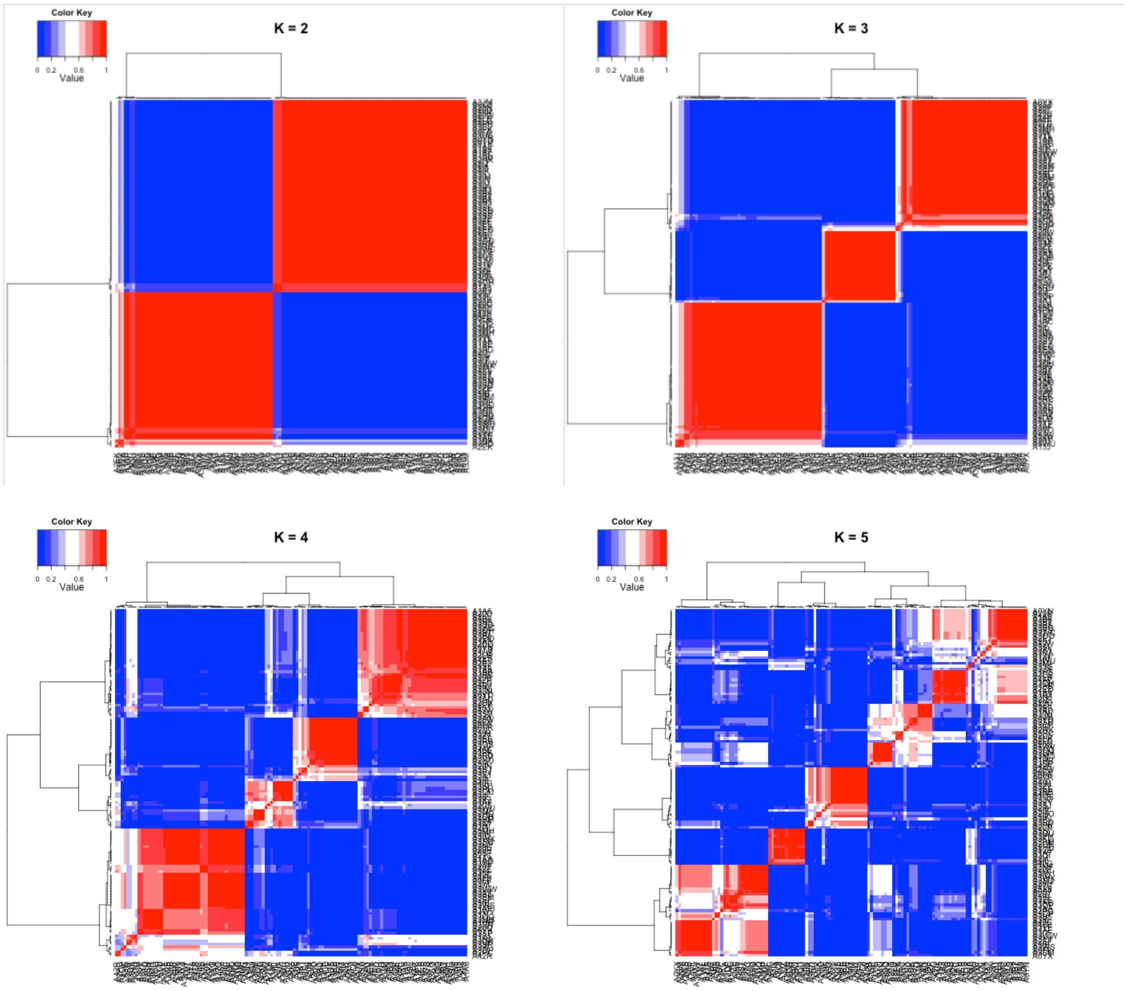
NCOR1	0.234028086	0.317609545	32
PTEN	0.432407628	0.547716328	16
RB1	0.541147534	0.642612696	18
CREBBP	0.755768315	0.844682235	17
CCND1	0.851824745	0.899148342	13
BCL2L1	0.932558695	0.932558695	14

Figure S2.1.1 mRNA expressions levels of significantly mutates genes (SMGs).



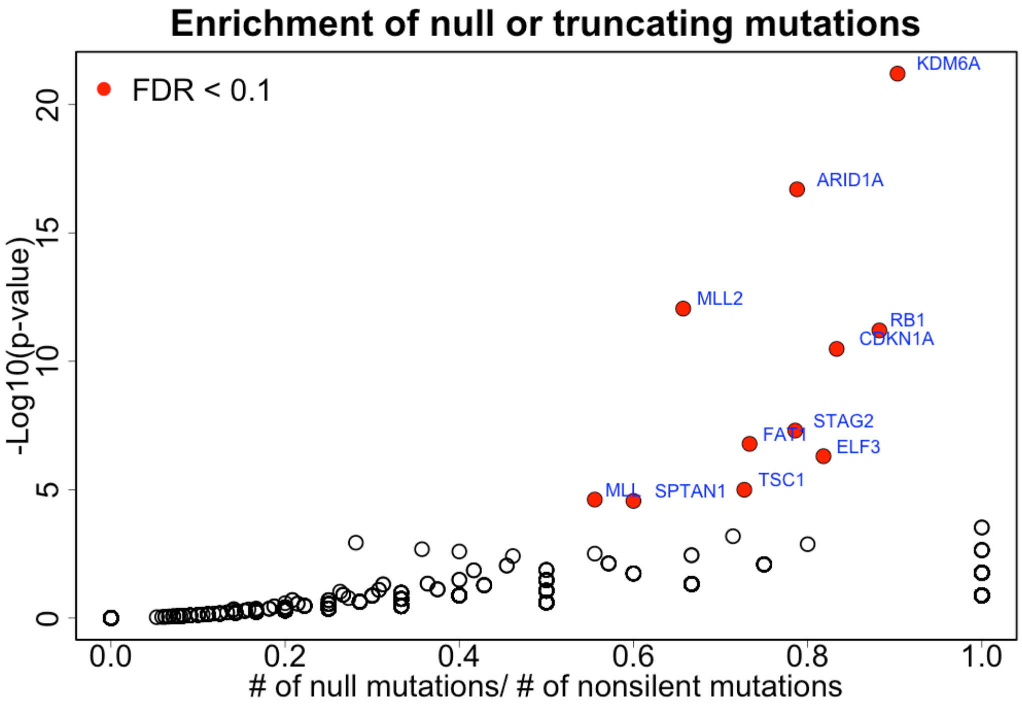
Boxplots of log2 (RSEM) distributions (left top) and rank (percentile) distributions (left bottom) of SMGs from MutSig v1.5. The rank of each gene is defined as a percentile of the corresponding RSEM in descending order in each sample. Three genes, *HORMAD1*, *GPC5*, and *ZFR2* (highlighted by red) were clustered together with a significantly lower mRNA level by PAM clustering for the rank matrix, in which each element represents a rank of RSEM in total 20502 genes in each sample. *HORMAD1* (37 samples have no transcripts) harbors two nonsense, two splice site, and four missense mutations. *GPC5* (73 samples have no transcripts) harbors two nonsense and six missense mutations. *ZFR2* (39 samples have no) harbors six missense mutations. The right figure is a graph of the mean percentile of gene expression (x axis) versus mean log2(RSEM) for every gene on the initial SMG list. Gray dots represent all genes, and blue and red circles correspond to the high expressed and low-expressed SMGs, respectively.

Figure S2.1.2 CNMF clustering of samples based on mutation and copy number events in Figure 1.



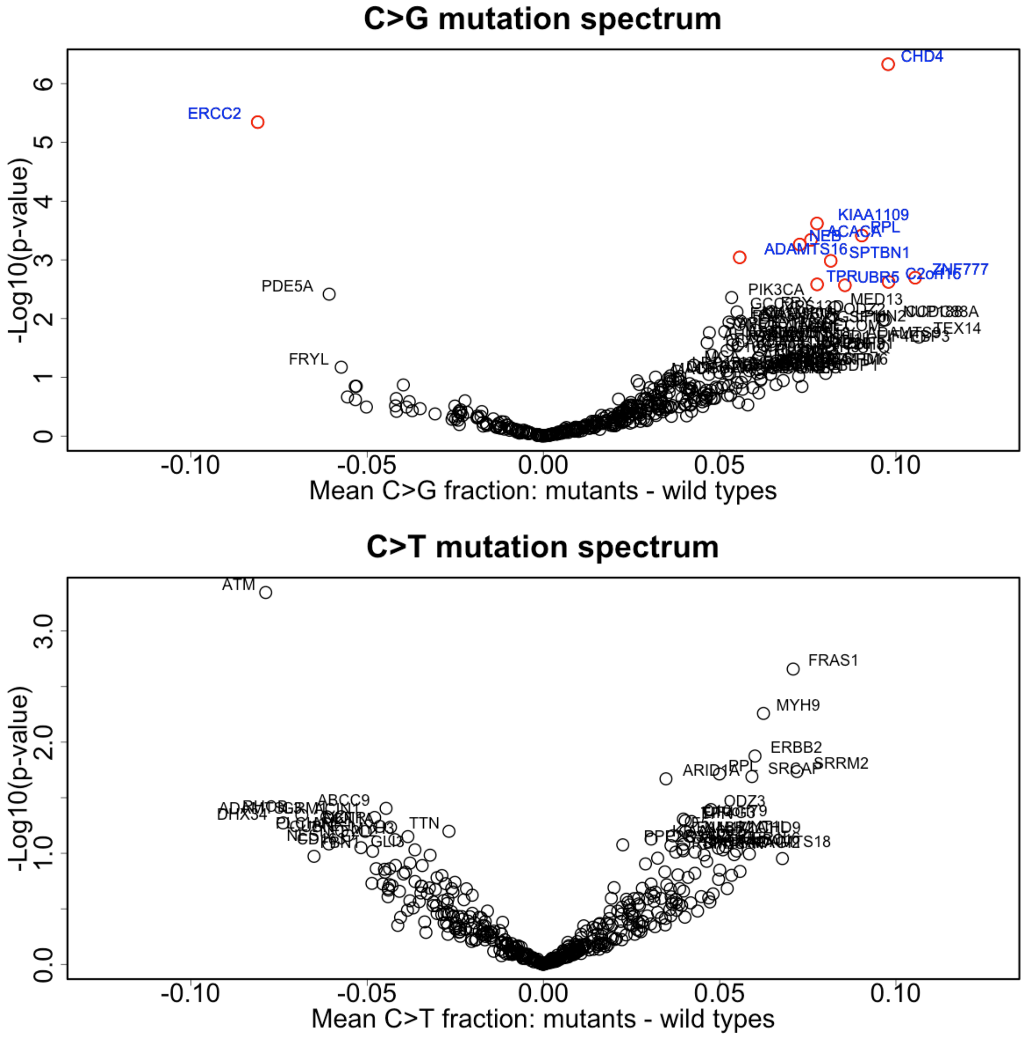
The consensus non-negative matrix factorization (CNMF) method (Jean-Philippe et al, PNAS **101**, 4164 (2004)) was applied to the binary event matrix comprised of mutations in SMGs and focal SCNAs with varying the number of clusters from K = 2 to 5. We only considered only 125 samples by excluding three samples with no copy number data and two samples with no mutations in the SMGs. Based on the visual inspection of a hierarchical clustering of the consensus matrix, defining the average connectivity over 100 clustering runs with different initial conditions, the case of K = 3 was used to arrange samples in Figure 1, giving rise to three clusters highlighted by red, green, and blue colors. The five samples that were not in CNMF clustering were highlighted by gray color.

Figure S2.2 Enrichment of null or truncating mutation.



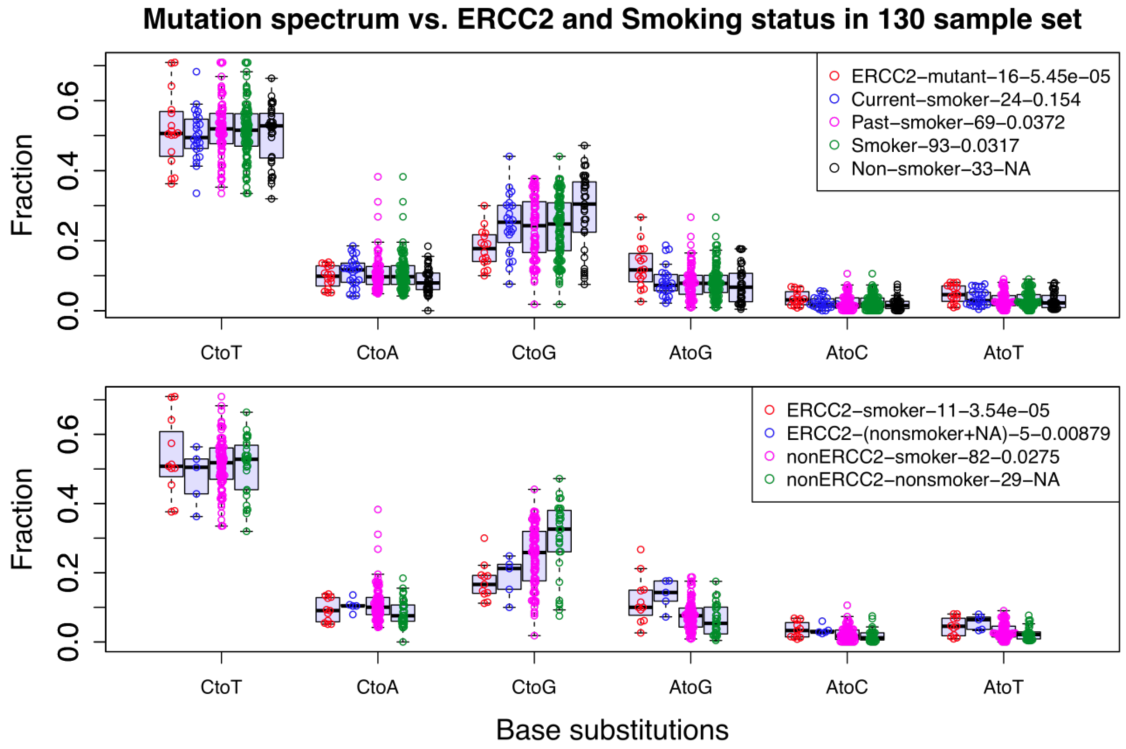
Enrichment analysis of null or truncating mutations in a specific gene or gene set was done by first computing a background rate of null mutations across all genes having at least one nonsilent mutation, yielding the rate, 0.13 = 361/27535 (361 null mutations out of 27535 nonsilent mutations). The Binomial test using the null mutation rate of 0.13 was used to determine p-values. Among 11739 genes, 11 genes (*KDM6A*, *ARID1A*, *MLL2*, *RB1*, *CDKN1A*, *STAG2*, *FAT1*, *ELF3*, *TSC1*, *MLL*, and *SPTAN1* highlighted by red filled circles) met statistical significance with FDR q < 0.1. Eight of these 11 genes had been identified as SMGs using MutSig as above, indicating that null mutations are enriched in SMGs. Note that the most significant three genes (*KDM6A, ARID1A, MLL2*) are all epigenetic modifiers. Null or truncating mutations were highly enriched in epigenetic modifier genes (P = 9.6 x 10^{-30}, Fisher's exact test).

Figure S2.3.1 C>G and C>T mutation spectrums depending on mutations status.



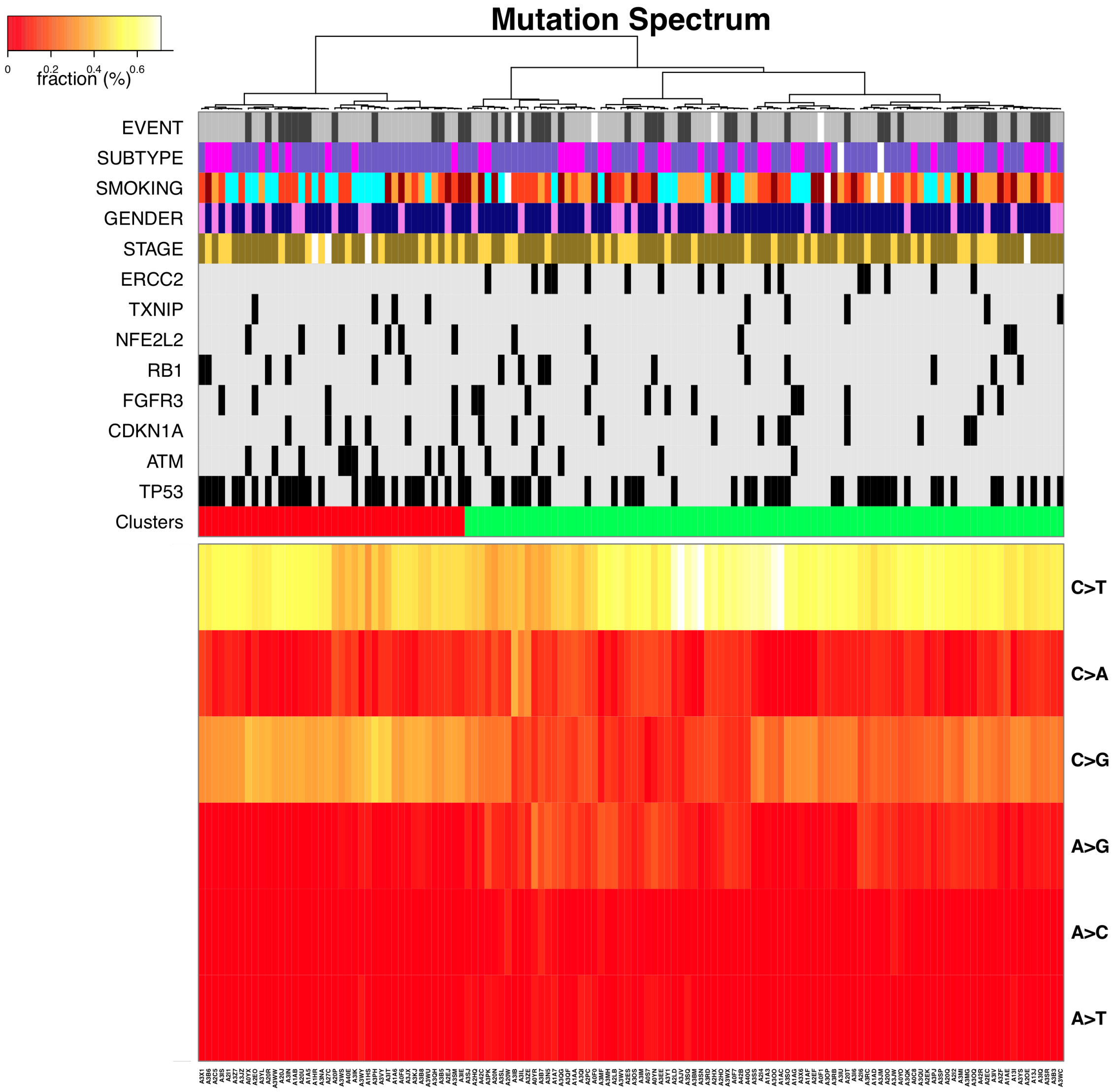
The fraction of the six different base substitutions (C>T, C>A, C>G, A>T, A>C, A>G) was computed for all single nucleotide variants (SNVs) detected in each sample. Samples were then segregated according to the presence or absence of mutation in each of the 456 genes in which mutations were seen in at least 5% samples, and the differences in substitution type fractions were calculated for each group, those with mutations vs. those without. The differences in C>G and C>T fractions according to gene are plotted on the x axis along with the -log10 (p value) according to the Student's t-test on the y axis. Cancers with ERCC2 mutations had a significant reduction in C>G transversion fraction (FDR < 0.1), whereas cancers with CHD4 mutations had a significant increase in C>G fraction (FDR < 0.1). The genes with FDR < 0.1 are highlighted by blue color. In contrast there were no genes with FDR < 0.1 for which mutation appeared to influence the fraction of C>T mutations (graph at bottom, note that scale is different).

Figure S2.3.2 Mutation spectrum according to ERCC2 mutation and smoking status



Boxplots of the fraction of each of six base substitutions (C>T, C>A, C>G, A>T, A>C, A>G) in ERCC2 mutants, current-smokers, past-smokers, all smokers (current or past), and non-smokers. The statistical significances of differences in C>G substitutions among groups were tested by Student t-test. (Top) The legend contains groups, the number of samples in groups, p-value to the non-smoker group. (Bottom) The legend contains group, the number of samples in groups, p-value to the non-ERCC2-nonsmker group.

Figure S2.3.3 Hierarchical clustering of six base substitutions.

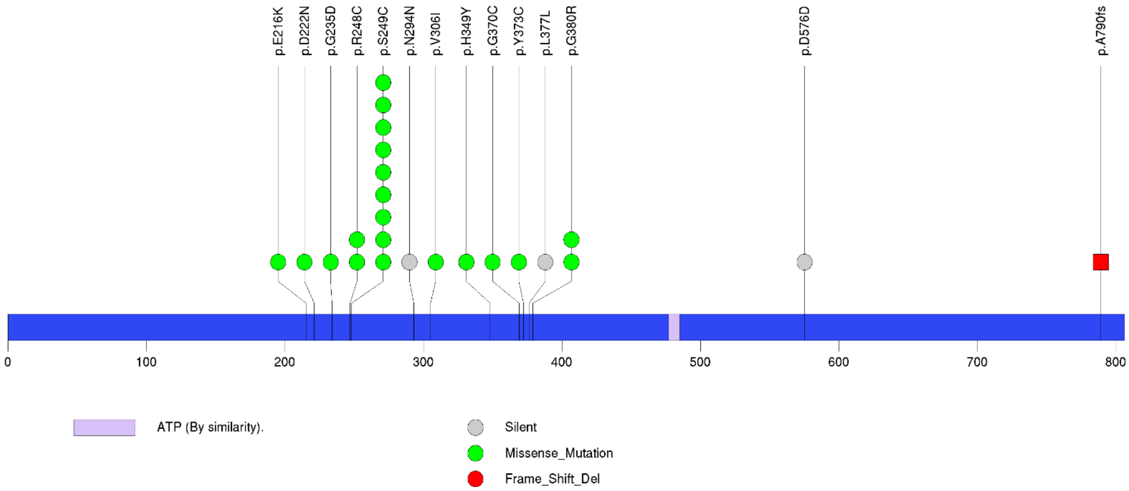


Hierarchical clustering of the fraction of each of six base substitutions (C>T, C>A, C>G, A>T, A>C, A>G) across 130 samples. Cluster I (highlighted by red) is characterized by a high prevalence of both C>T and C>G mutations, while cluster II (highlighted by green) had a predominance of C>T mutations with a modest increase in A>G mutations in comparison to the cluster I. Interestingly, all ERCC2 mutants belonged to cluster II with a significant lower C>G mutation fraction (16 out of 90 and P = 0.0027 by Fisher exact test), while many non-smokers (cyan color in "SMOKING" section) were enriched in the cluster I with higher C>G mutation fraction (16 in 40 samples and P = 0.0087 by Fisher exact test), which is concordant with the observations in Figures S2.3.1 and S2.3.2.

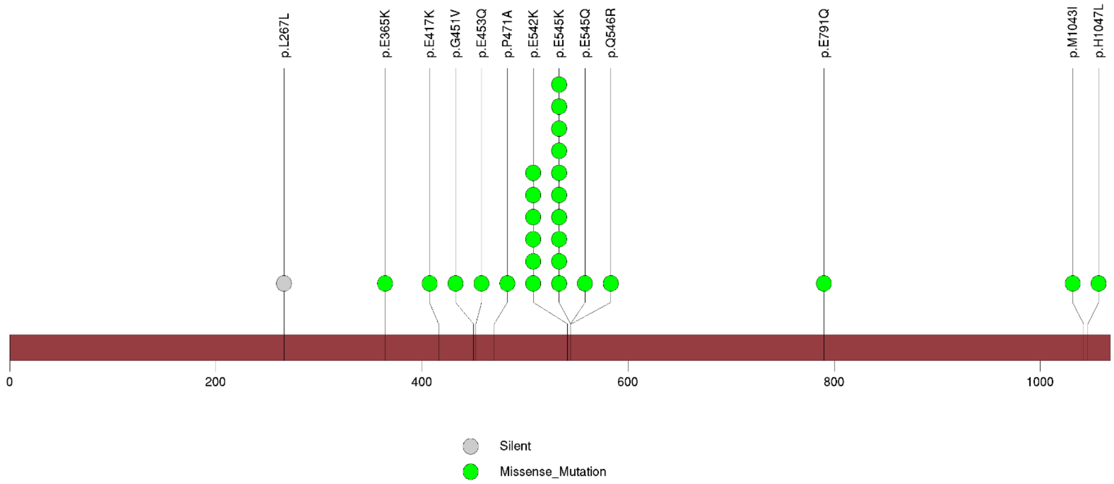
Figure S2.4 Stick figures of mutations in selected SMGs.

Silent mutations are highlighted by gray, frame-shifts mutations were highlighted by red, and missense or nonsense mutations were highlighted by green with corresponding amino acid changes. The amino acid change in nonsense mutations was denoted by *.

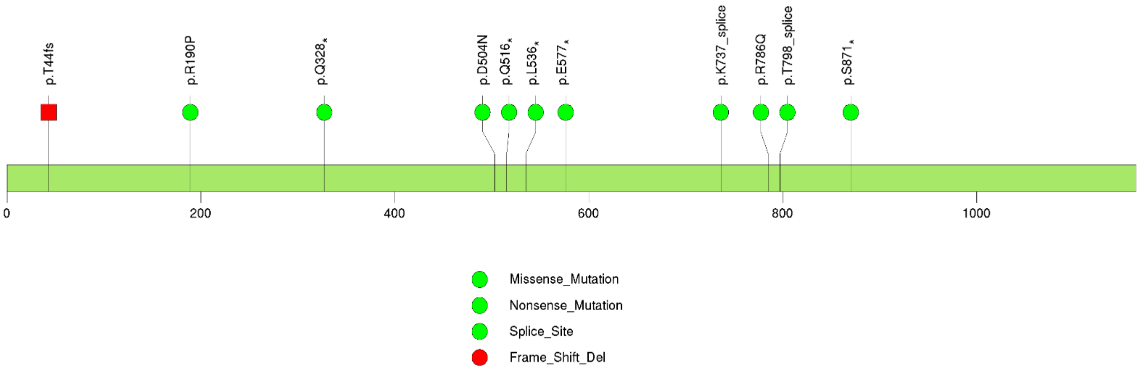
FGFR3



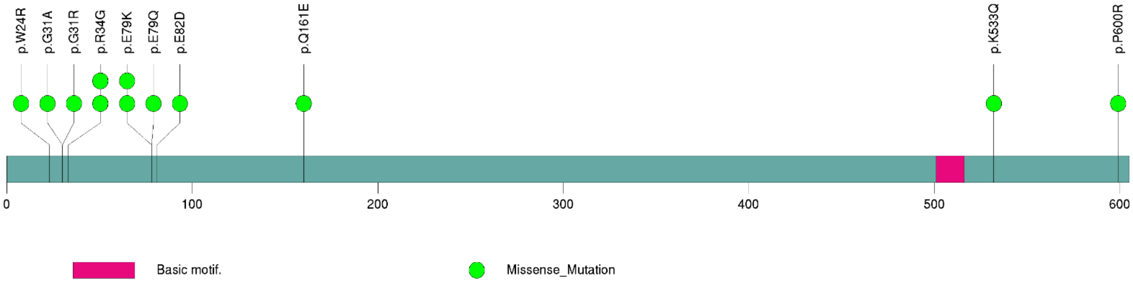
PIK3CA



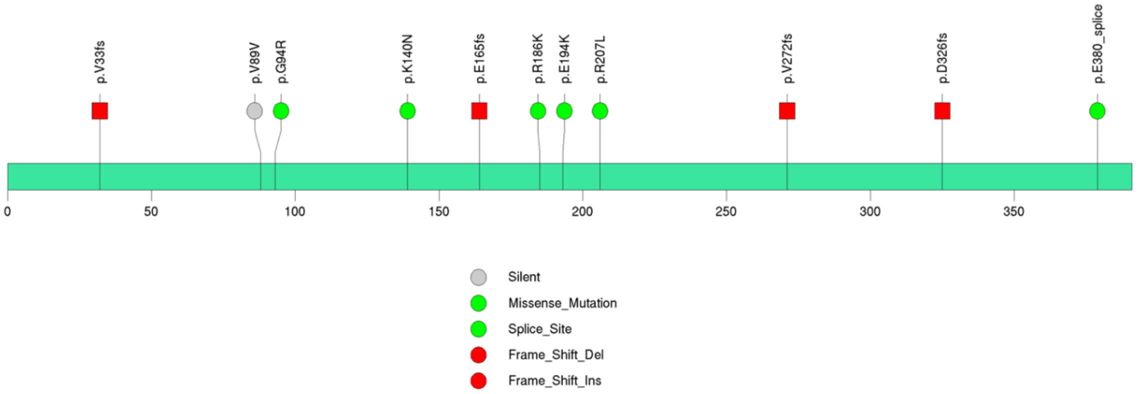
TSC1



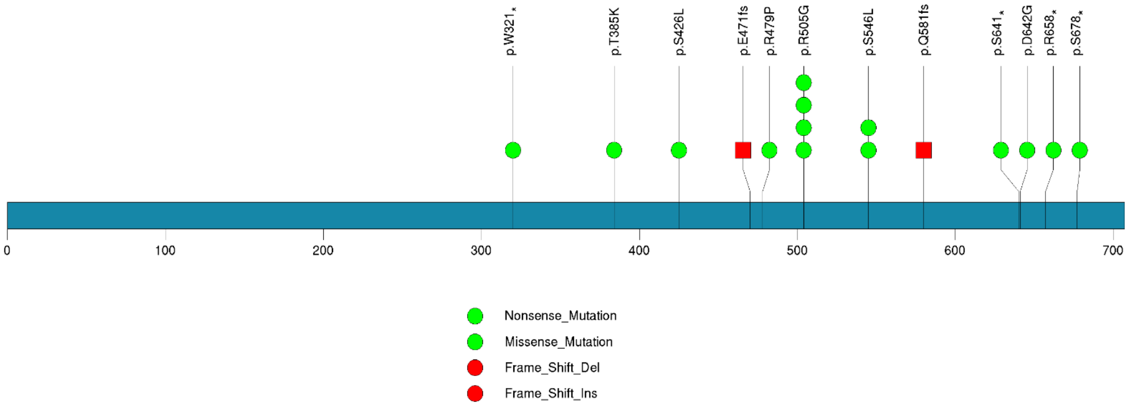
NFE2L2



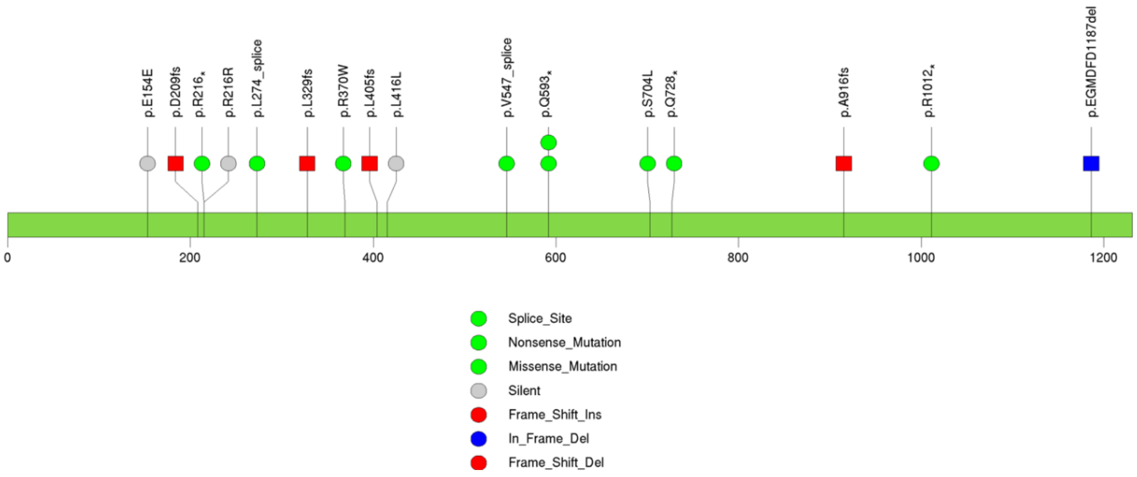
TXNIP



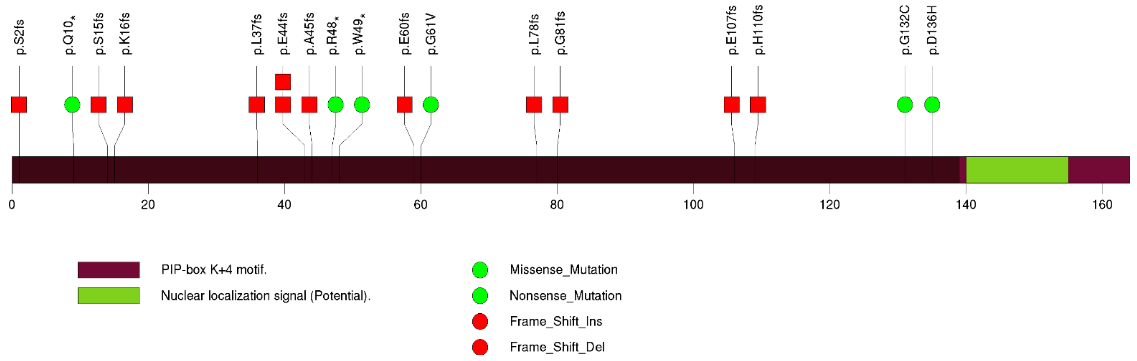
FBXW7



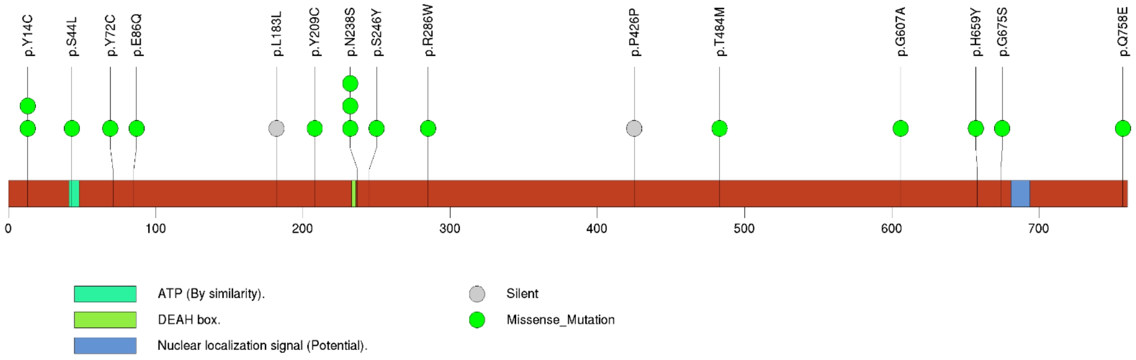
STAG2



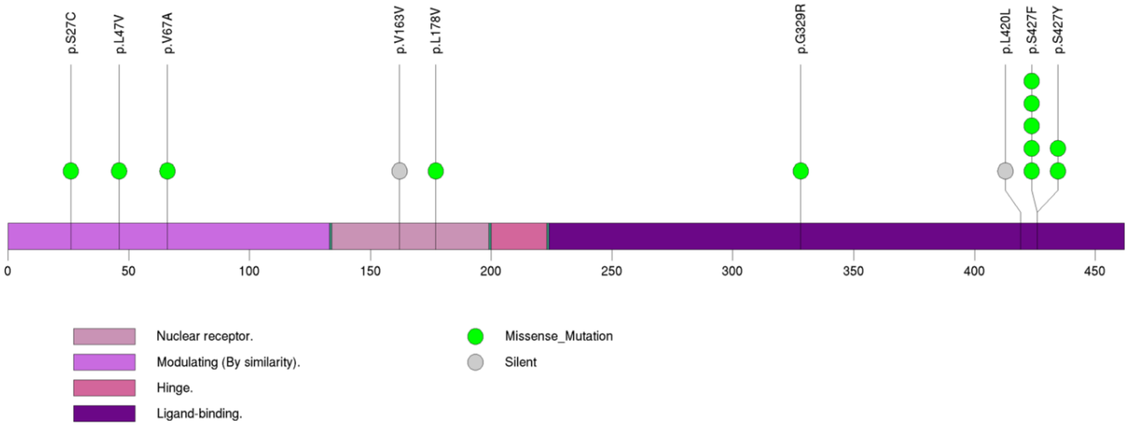
CDKN1A



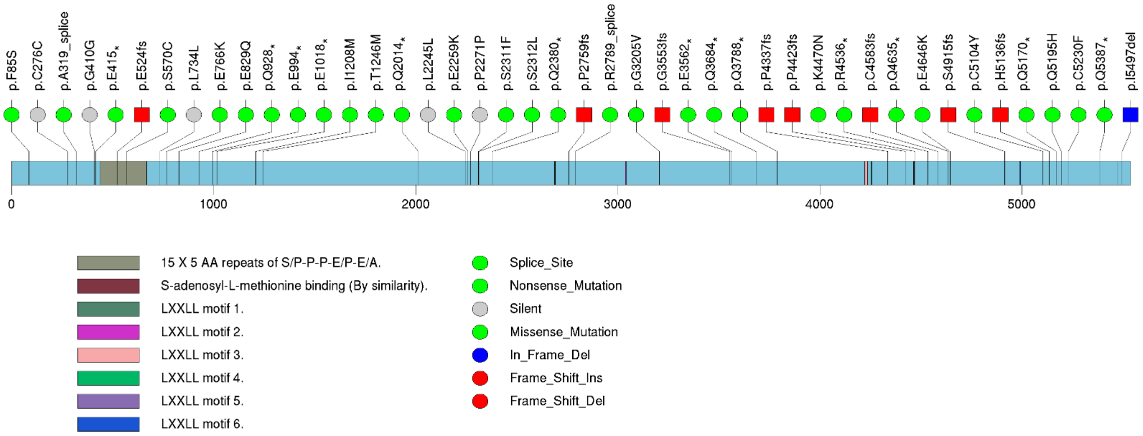
ERCC2


p.Y14C
p.S44L
p.Y72C
p.E86Q
p.L183L
p.Y209C
p.N238S
p.S246Y
p.R286W
p.P426P
p.T484M
p.G607A
p.H659Y
p.G675S
p.Q758E
0
100
200
300
400
500
600
700
ATP (By similarity).
DEAH box.
Nuclear localization signal (Potential).
Silent
Missense_Mutation

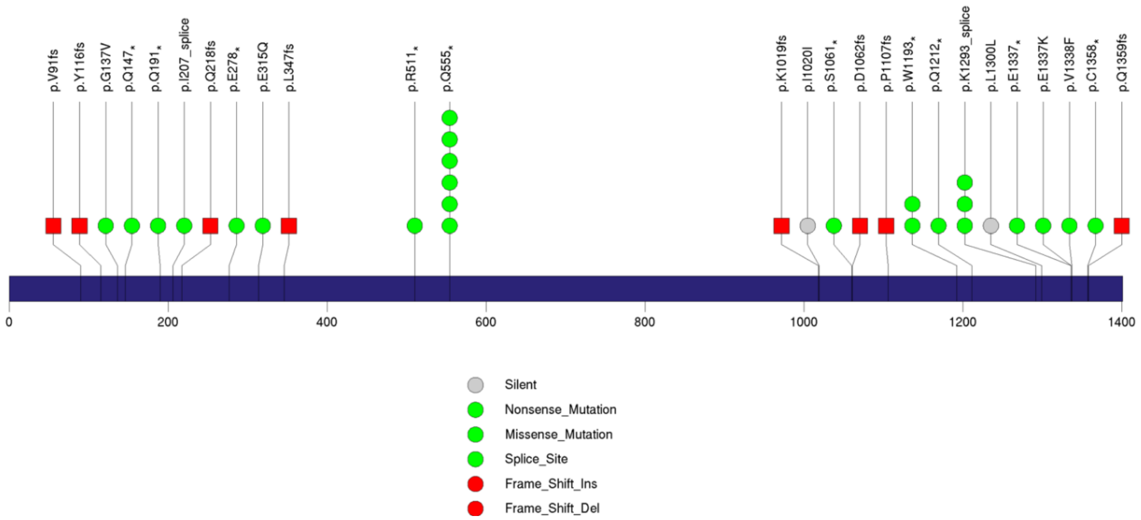
RXRA


p.S27C
p.L47V
p.V67A
p.V163V
p.L178V
p.G329R
p.L420L
p.S427F
p.S427Y
0
50
100
150
200
250
300
350
400
450
Nuclear receptor.
Modulating (By similarity).
Hinge.
Ligand-binding.
Missense_Mutation
Silent

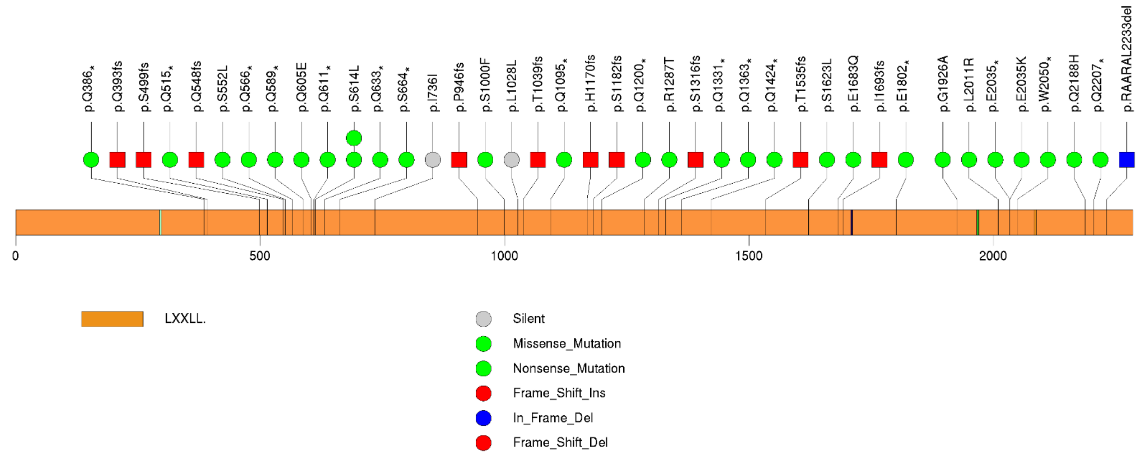
MLL2


p.F85S
p.C276C
p.A319_splice
p.G410G
p.E415*
p.E524fs
p.S570C
p.L734L
p.E766K
p.E829Q
p.Q928*
p.E994*
p.E1018*
p.I1208M
p.T1246M
p.Q2014*
p.L2245L
p.E2259K
p.P2271P
p.S2311F
p.S2312L
p.Q2380*
p.P2759fs
p.R2789_splice
p.G3205V
p.G3553fs
p.E3562*
p.Q3684*
p.Q3788*
p.P4337fs
p.P4423fs
p.K4470N
p.R4536*
p.C4583fs
p.Q4635*
p.E4646K
p.S4915fs
p.C5104Y
p.H5136fs
p.Q5170*
p.Q5195H
p.C5230F
p.Q5387*
p.I5497del
0
1000
2000
3000
4000
5000
15 X 5 AA repeats of S/P-P-P-E/P-E/A.
S-adenosyl-L-methionine binding (By similarity).
LXXLL motif 1.
LXXLL motif 2.
LXXLL motif 3.
LXXLL motif 4.
LXXLL motif 5.
LXXLL motif 6.
Splice_Site
Nonsense_Mutation
Silent
Missense_Mutation
In_Frame_Del
Frame_Shift_Ins
Frame_Shift_Del

KDM6A


p.V91fs
p.Y116fs
p.G137V
p.Q147*
p.Q191*
p.I207_splice
p.Q218fs
p.E278*
p.E315Q
p.L347fs
p.R511*
p.Q555*
p.K1019fs
p.I1020I
p.S1061*
p.D1062fs
p.P1107fs
p.W1193*
p.Q1212*
p.K1293_splice
p.L1300L
p.E1337*
p.E1337K
p.V1338F
p.C1358*
p.Q1359fs
0
200
400
600
800
1000
1200
1400
Silent
Nonsense_Mutation
Missense_Mutation
Splice_Site
Frame_Shift_Ins
Frame_Shift_Del

ARID1A



EP300

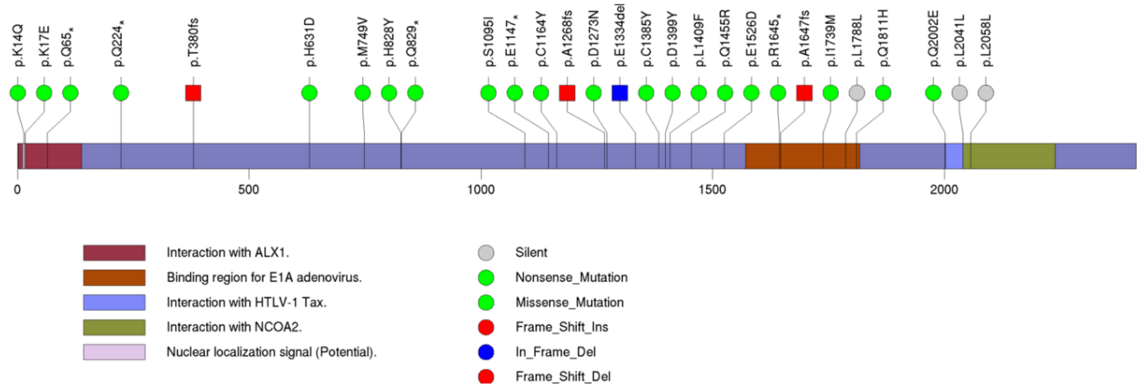
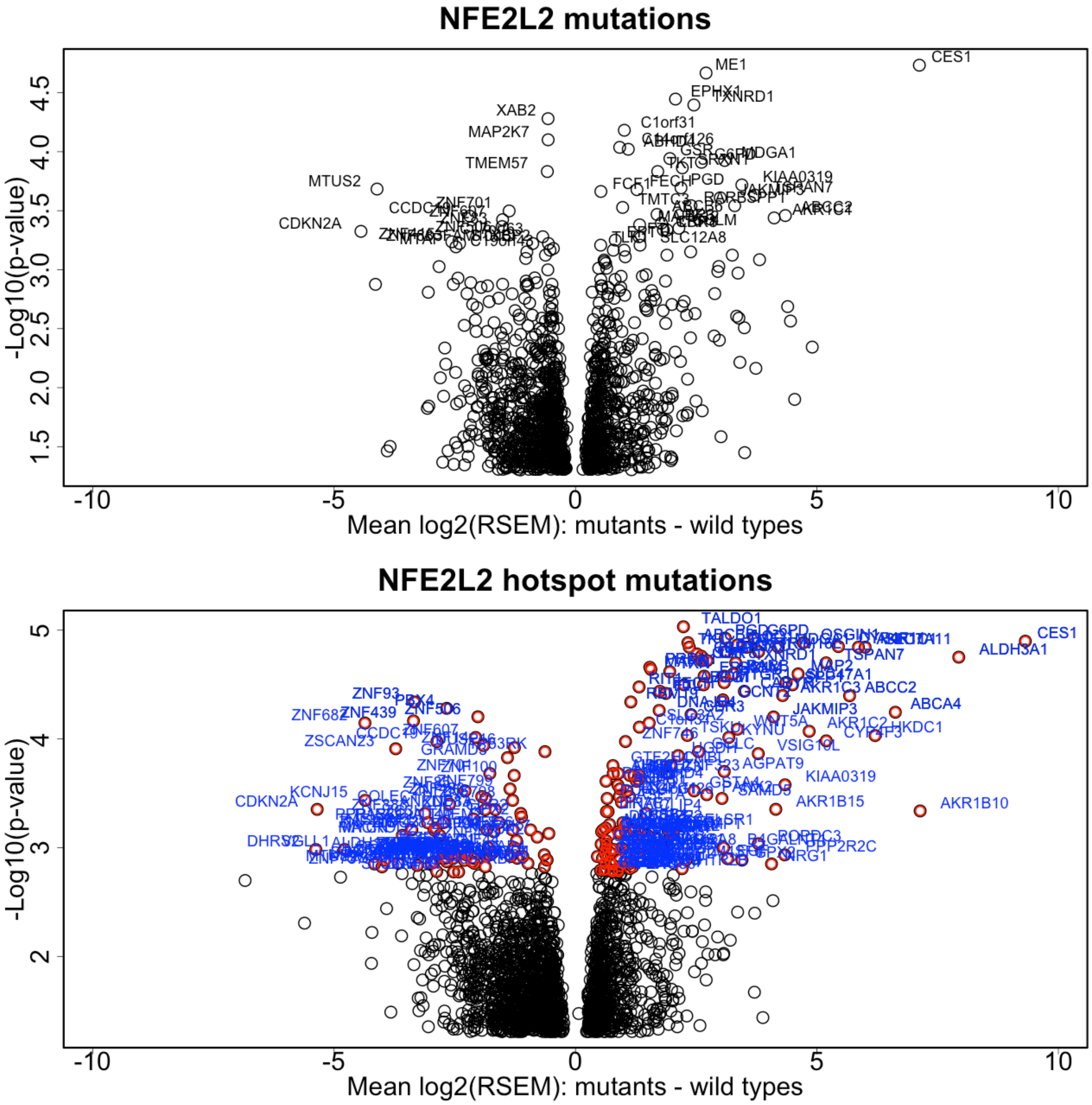
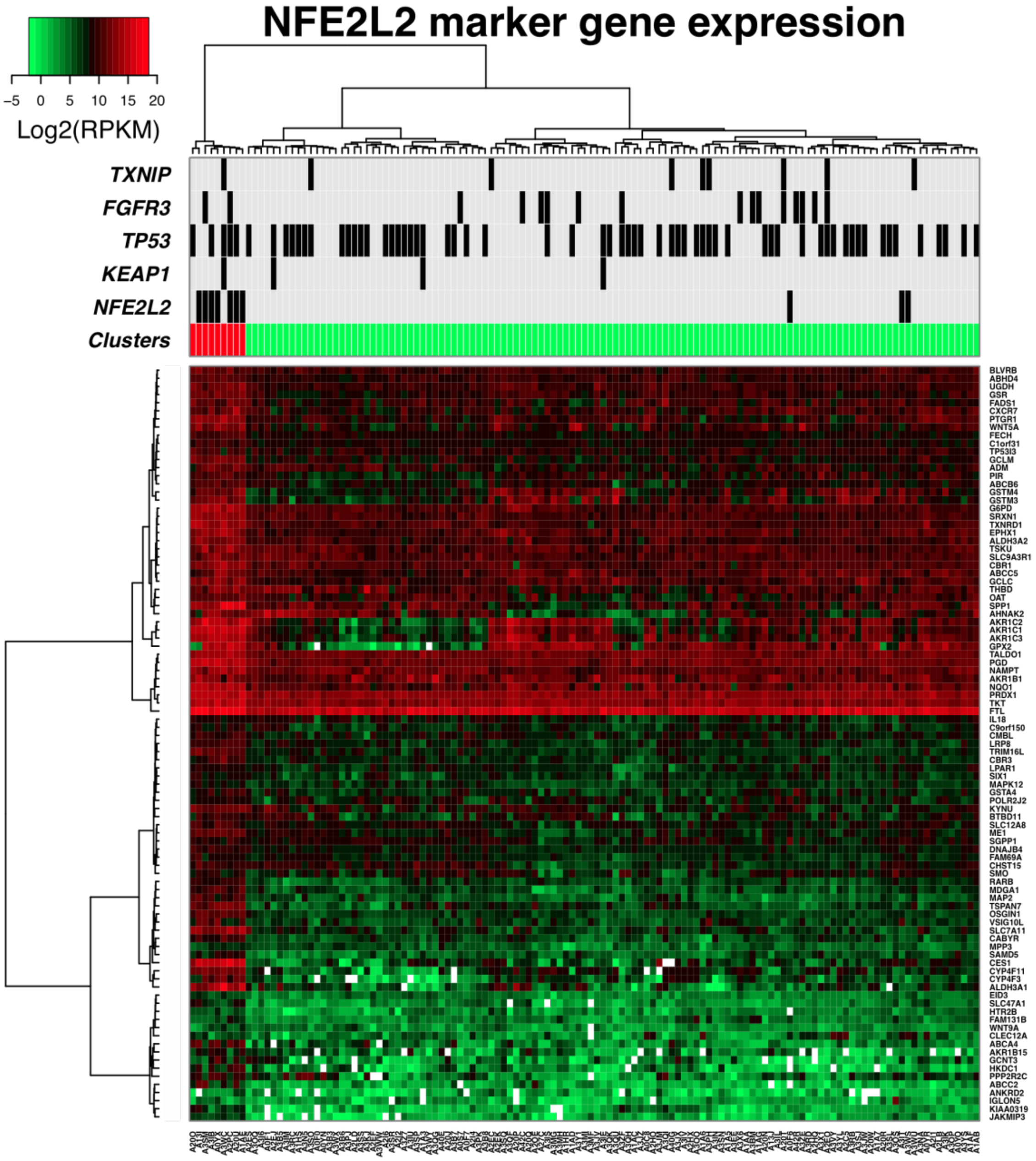


Figure S2.5.1 Differentially expressed genes in NFE2L2 mutants and hotspot mutants



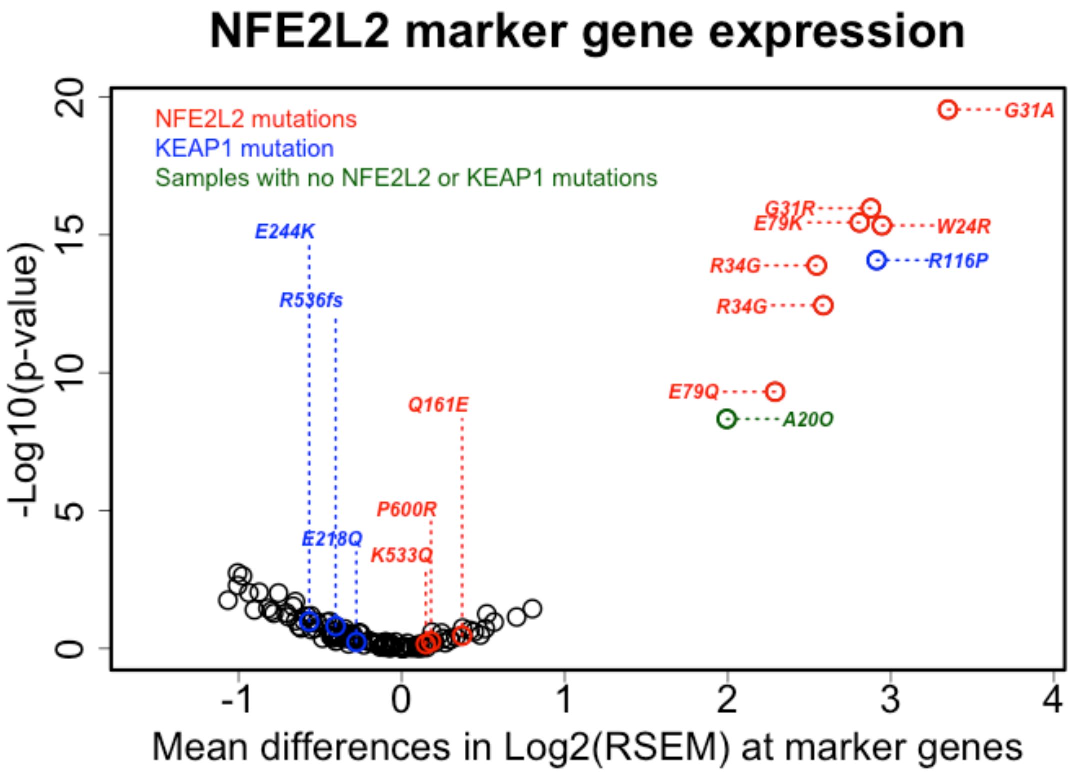
(Top) Differentially expressed genes (DEGs) in NFE2L2 mutant samples were identified by comparison of the log2(RSEM) values of NFE2L2 mutant and non-mutant sample sets. Each gene for which the p-value according to Wilcoxon Rank-sum test was < 0.05 is indicated in the graph by a circle, positioned according to the difference in the mean log2(RSEM) (x axis) and the –log10(p-value) (y axis). (Bottom) DEGs are shown as in a, but considering only NFE2L2 mutations affecting the DLG or ETGE motifs. Genes with FDR q < 0.1 are highlighted by red circles and blue names. The number of significantly DEGs was much higher in samples with hotspot mutations and degree of change in expression was dramatic for many genes.

Figure S2.5.2 Hierarchical clustering of bladder cancer samples using NFE2L2 marker genes.



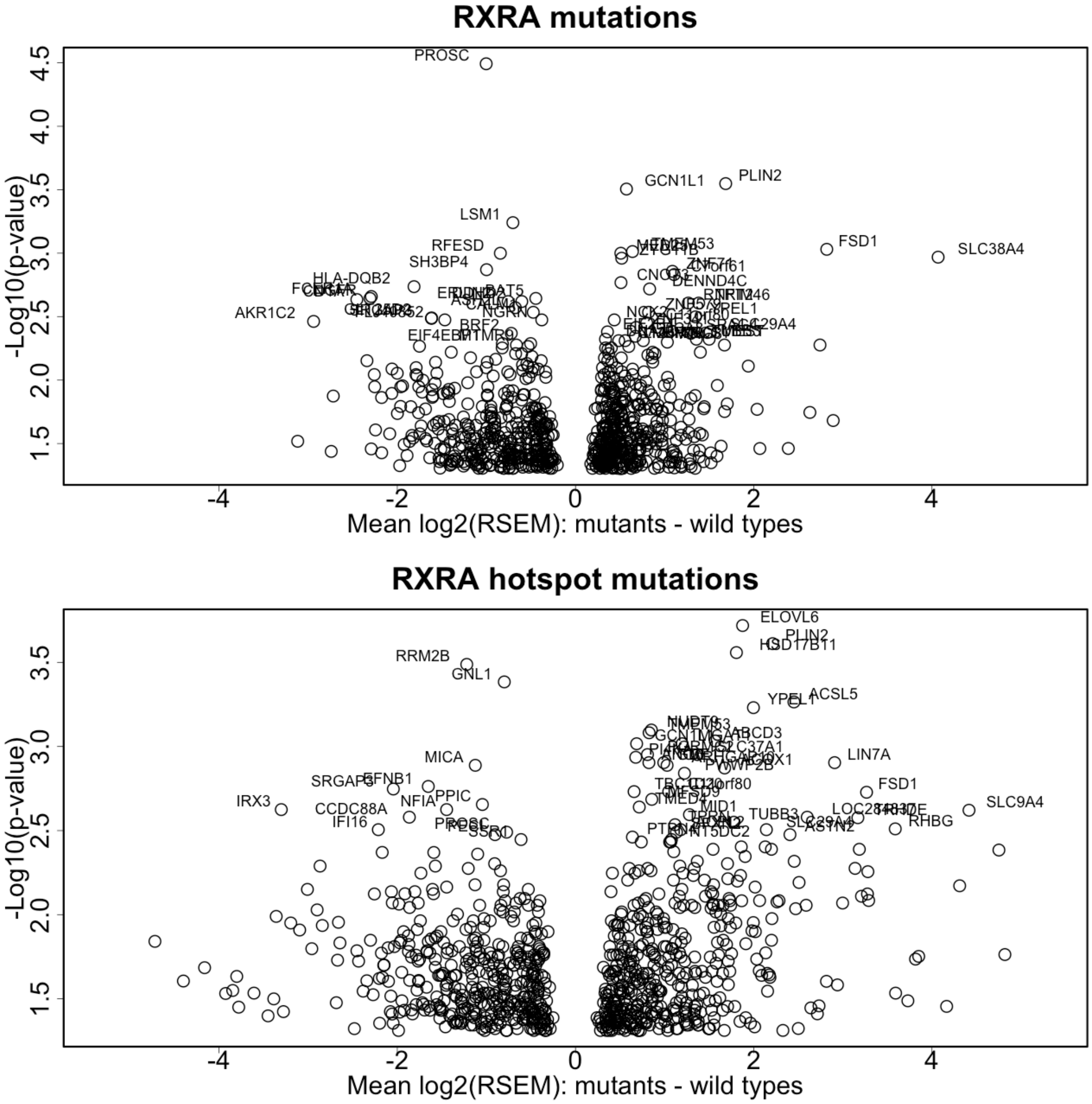
By selecting NFE2L2 marker genes with p-value < 0.01 and fold change > 2 we performed a hierarchical clustering of samples according to NFE2L2 marker gene expressions. Seven NFE2L2 hotspot mutants and one KEAP1 mutant were co-clustered together (red cluster at left), showing a significant elevation of mRNA expressions in these NFE2L2 marker genes. One additional sample with no mutations in NFE2L2 or KEAP1 was also in this cluster.

Figure S2.5.3 Sample-specific NFE2L2 marker gene expressions vs Amino acid substitutions in NFE2L2 and KEAP1.



NFE2L2 marker gene expression values were compared to the average for each sample and plotted to indicate NFE2L2 and KEAP1 mutation status for each sample. The x-axis is the mean difference of log2(RSEM) and the y axis is –log10 (P-value) by Wilcoxon rank sum test. Note that all NFE2L2 hotspot mutants at KEAP1 binding domain (DLG or ETGE motif) and a KEAP1 mutant with an amino acid change of R116P showed a dramatic differential expression in NFE2L2 marker genes. In addition, one sample (A20O) with no mutations in NFE2L2 or KEAP1 also showed a high level mRNA expression of these marker genes.

Figure S2.6 Differentially expressed genes in RXRA mutants.

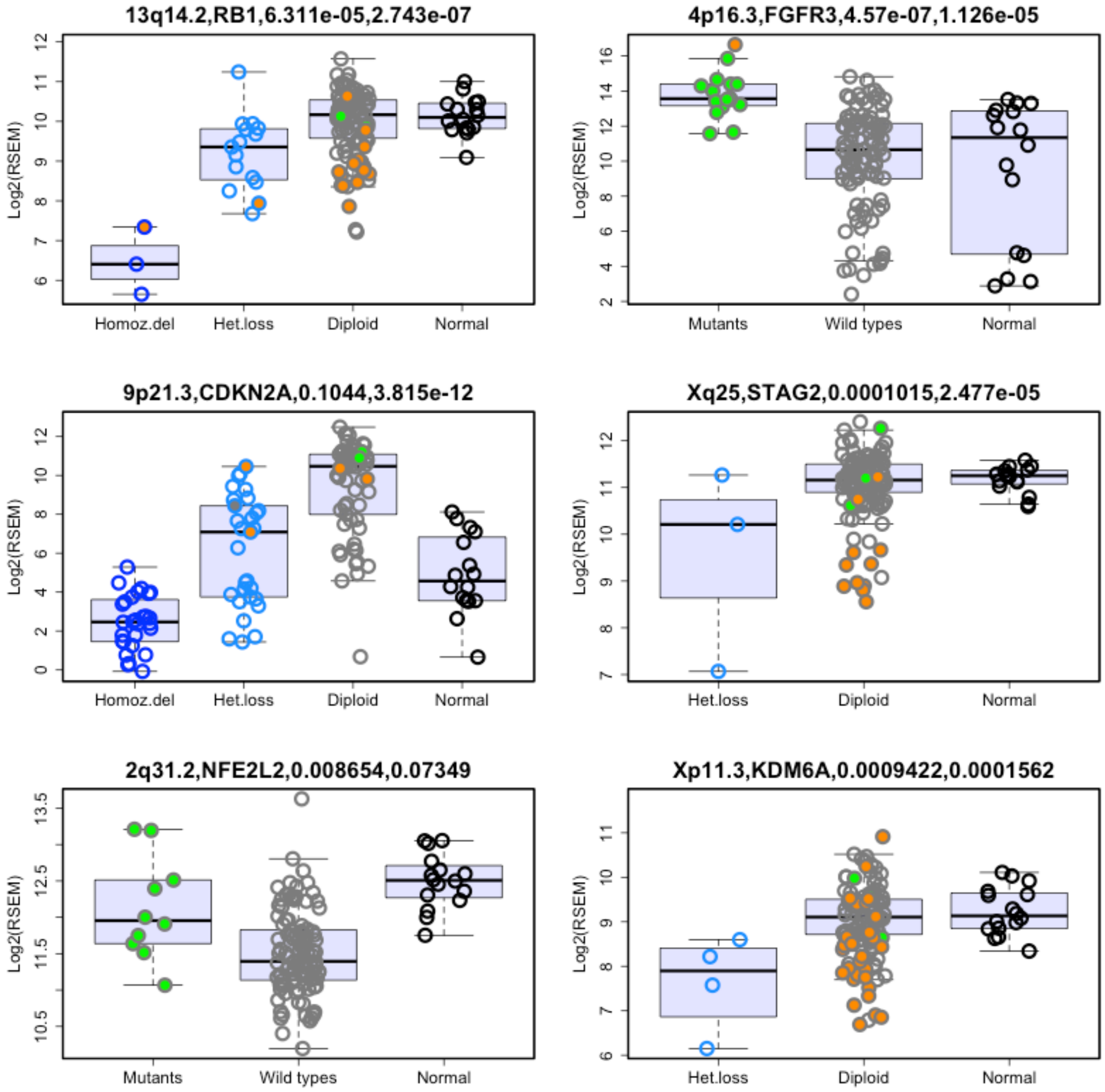


RXRA mutations

RXRA hotspot mutations

(Top) Differentially expressed genes (DEGs) in RXRA mutant samples were identified by comparing log2(RSEM) values between RXRA mutants and non-mutant samples. The p-values were determined by Wilcoxon Rank-sum test. Genes ordered by p-values to a maximum 0.05 were named. (Bottom) Differentially expressed genes in samples RXRA recurrent mutations at S427 were identified by the same method above. Gene names by p-value to a maximum 0.05 were shown. The number of DEGs were much higher in samples harboring recurrent mutations at S427. Up-regulated genes in hotspot mutants were enriched for those involved in the PPARG pathway (P = 0.0016) and lipid metabolic processes or adipocyte differentiation process (P = 0.006).

Figure S2.7 Correlations between mRNA levels and mutation for selected significantly mutated genes.



The title above each figure indicates the cytoband, gene name, p-value for comparison of log2(RSEM) between mutant and non-mutant samples, and p-value for comparison between altered (mutations or SCNAs) and non-altered samples. Dark blue circles denote “Homozygous deletion” (< 1 copy) , light-blue circles represent “Heterozygous loss” (between 1 copy and 1.5 copy), gray circles are CN normal (“Diploid”), red circles represent copy number gain (> 3 copy). Samples with missense mutations are indicated by green fill; those with null or truncating mutations by orange-fill. Normal (black circles) are normal samples. All p-values were computed by Wilcoxon rank-sum test.

Table S2.8.1 Mutual exclusivity correlations.

These are the results of mutual exclusivity analyses to a maximum p-value of 0.05 by Fisher's exact test for significantly mutated genes (no suffix) or genes having focal SCNAs (suffix: .copy). Columns C and D report p-values and false discovery rates (FDR) for a pair of genes in columns A and B.

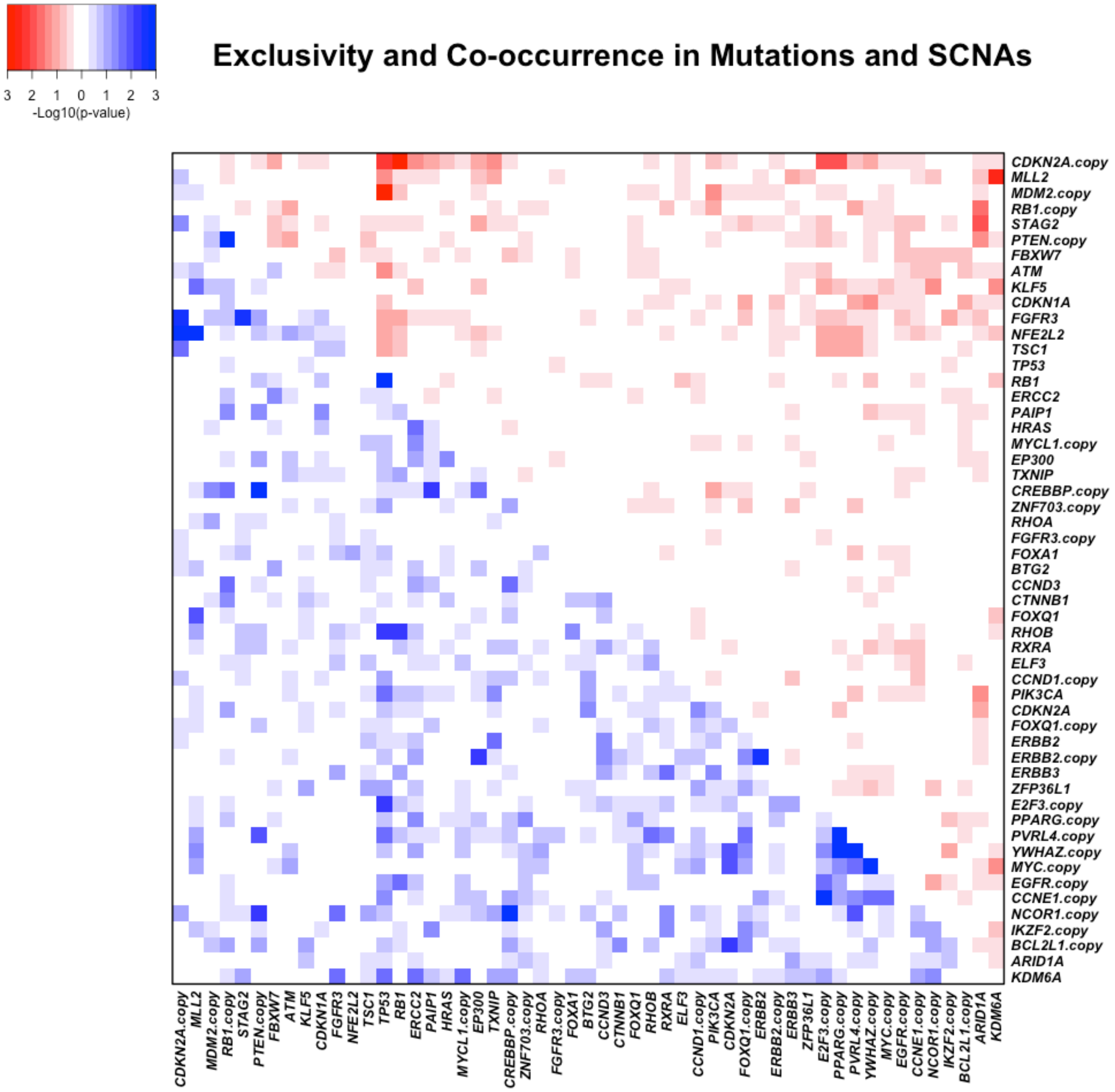
gene1	gene2	pval	qval
RB1	CDKN2A.copy	6.32E-06	0.00904392
TP53	MDM2.copy	0.000153	0.1094715
MLL2	KDM6A	0.00244	1
TP53	CDKN2A.copy	0.00477	1
CDKN2A.copy	PPARG.copy	0.00965	1
ARID1A	STAG2	0.0113	1
CDKN2A.copy	E2F3.copy	0.0129	1
ARID1A	RB1.copy	0.0241	1
ARID1A	PTEN.copy	0.0432	1
TXNIP	CDKN2A.copy	0.0438	1
KDM6A	MYC.copy	0.0443	1
ARID1A	PIK3CA	0.0456	1
KLF5	NCOR1.copy	0.0483	1
ERCC2	CDKN2A.copy	0.0488	1

Table S2.8.2 Co-occurrence correlations.

These are results of co-occurrence tests to a maximum q-value of 0.1 by Fisher's exact tests for significantly mutated genes or genes having focal SCNAs. Columns C and D report p-values and false discovery rates (FDR) for a pair of genes in columns A and B, respectively.

gene1	gene2	pval	qval
YWHAZ.copy	MYC.copy	2.76E-07	0.000394956
PPARG.copy	YWHAZ.copy	3.01E-06	0.002153655
YAP1.copy	MYC.copy	2.43E-05	0.0115911
CREBBP.copy	PTEN.copy	0.000132	0.0440748
NFE2L2	CDKN2A.copy	0.000154	0.0440748

Figure S2.8 Mutual exclusivity and co-occurrence correlations



The upper right triangle (red) represents mutual exclusivity relationships and the lower left triangle (blue) represents co-occurrence relationships between SMGs and genes having focal SCNAs. All statistical tests were done by Fisher’s exact test and the values in heatmap refer to –log10(P-value).

Table S2.9.1 Clinical associations of mutations in significantly mutated genes

P-values of Fisher's exact tests for the association of mutations in SMGs to the clinical data, including death events (alive vs deceased), gender, subtype (papillary vs non-papillary), smoking status (smoker vs non-smoker), and stage (stage I or II vs III or IV).

gene	event	gender	subtype	smoking	stage	n.mutation
ARID1A	0.831	0.815	0.182	0.489	0.272	33
CDKN1A	0.576	0.238	0.271	0.78	0.775	18
TP53	0.345	0.84	0.000513	0.106	0.846	64
RB1	0.0485	0.562	0.265	1	1	17
KDM6A	1	0.633	0.272	1	0.495	31
PIK3CA	0.492	0.801	0.354	0.612	0.475	26
ELF3	1	0.463	0.739	0.721	0.0317	11
ERCC2	1	0.76	0.261	1	0.775	16
MLL2	0.00247	0.502	0.666	0.357	0.665	35
FBXW7	0.00636	0.306	0.0622	1	0.341	13
FOXQ1	0.0943	1	0.677	0.675	0.667	7
NFE2L2	1	0.137	0.501	1	0.171	11
FGFR3	0.0867	0.355	0.0398	0.761	0.382	16
TXNIP	0.48	1	0.0562	0.696	0.722	9
STAG2	0.383	1	0.365	0.759	1	14
CDKN2A	0.423	1	0.0976	1	0.0258	7
RHOB	1	0.362	1	1	0.101	7
BTG2	1	0.335	0.665	0.652	1	6
FOXA1	0.0943	0.0617	1	1	0.177	7
RXRA	0.532	1	0.752	0.73	1	12
EP300	0.449	0.78	0.607	0.589	0.791	20
KLF5	1	0.0657	1	1	1	10
HRAS	0.663	0.159	0.376	1	1	6
ERBB3	0.773	0.515	0.0112	1	0.753	14
ZFP36L1	1	0.159	1	1	0.667	6
RHOA	0.664	0.0954	1	0.112	0.637	5
PAIP1	1	1	1	1	0.667	7
TSC1	0.0982	1	0.00393	0.154	0.722	11
CCND3	0.664	0.333	1	0.325	1	5

Table S2.9.2 Clinical associations of genes harboring focal SCNAs.

P-values of Fisher's exact tests for the association of copy number events in genes harboring focal SCNAs to the clinical data, including death events (alive vs deceased), gender, subtype (papillary vs non-papillary), smoking status (smoker vs non-smoker), and stage (stage I or II vs III or IV).

gene	event	gender	subtype	smoking	stage	n.copy
MYCL1	1	0.403	0.71	0.204	0.673	8
PVRL4	1	0.795	1	1	0.328	24
PPARG	1	0.277	0.623	0.794	0.21	22
FGFR3	1	1	0.0961	1	0.0809	4
E2F3	1	1	0.23	0.794	0.619	24
EGFR	0.142	0.515	0.767	0.354	0.353	14
YWHAZ	0.25	0.62	0.491	0.146	1	28
MYC	0.58	1	0.578	0.76	0.583	17
YAP1	1	1	0.176	0.572	0.0283	5
CCND1	0.521	1	0.222	0.511	0.208	13
MDM2	0.101	0.164	1	0.296	0.732	12
ERBB2	0.468	0.112	1	0.444	1	9
CCNE1	0.545	1	0.141	1	1	15
BCL2L1	0.545	0.327	0.0599	0.112	0.0599	14
CDKN2A	0.848	0.54	0.18	1	0.325	60
RB1	0.789	0.237	1	0.152	0.245	18
CREBBP	0.263	1	0.578	0.233	1	17
NCOR1	0.383	1	0.19	0.644	1	32
PTEN	0.264	0.537	0.266	0.559	0.775	16
WWOX	0.663	0.335	1	0.338	0.321	6

Table S2.10 Enrichments of non-silent mutations in chromatin remodeling genes in BLCA

The enrichment of non-silent mutations in 146 chromatin remodeling genes in BLCA was compared to that in published TCGA tumor types, including breast (BRCA), colorectal (COAD), brain (GBM), lung (LUSC), blood (LAML), ovarian (OV), and endometrial (UCEC) cancers. The p-values were determined by Binomial test with a background non-silent mutation rate in 146 chromatin remodeling genes across tumors. Columns A to E contain tumor types, number of non-silent mutations (n_all), number of non-silent mutations in 146 chromatin remodeling genes (n_chromatin), p-values, and false discovery rate (FDR) q values. Except for LAML, BLCA had the most significant enrichment of non-silent mutations in chromatin remodeling genes with FDR < 0.1.

tumor	n_all	n_chromatin	pval	qval
BLCA	29140	664	1.44E-23	5.77E-23
BRCA	24584	351	0.872123958	0.999999316
COAD	68359	954	0.995074106	0.999999316
GBM	453	9	0.251546713	0.670791234
LAML	1963	162	5.31E-66	4.25E-65
LUSC	48746	612	0.999999316	0.999999316
OV	15001	194	0.989136657	0.999999316
UCEC	140677	2034	0.982873001	0.999999316

Table S2.11.1 Validation rate of exome-called SNVs by targeted re-sequencing while varying the number of required variant alleles for validation.

Variant allele	# validated	# powered sites (≥90%)	validation rate (%)
2	612	618	99.03%
5	599	616	97.24%
10	579	593	97.64%
20	541	553	97.83%
40	488	503	97.02%
50	455	468	97.22%
100	325	346	93.93%
150	207	227	91.19%

Table S2.11.2 Validation rate of exome-called SNVs in significantly mutated genes by targeted re-sequencing while varying the number of required variant alleles for validation.

Variant allele	# validated	# powered sites (≥ 90%)	validation rate(%)
2	368	371	99.19%
5	361	369	97.83%
10	352	358	98.32%
20	335	338	99.11%
40	312	321	97.20%
50	293	301	97.34%
100	210	222	94.59%
150	143	159	89.94%

Figure S2.11.1 Validation of SNV mutations in 53 genes by targeted re-sequencing.

Validation of single nucleotide mutations by targeted re-sequencing across 53 genes including significantly mutated genes, and selected cancer genes and chromatin modifying genes. Validation status is shown by gene (top) and sample (bottom). The numbers on the top of each bargraph (top) are the number of powered but not validated mutations over the number of validated mutations at powered sites.

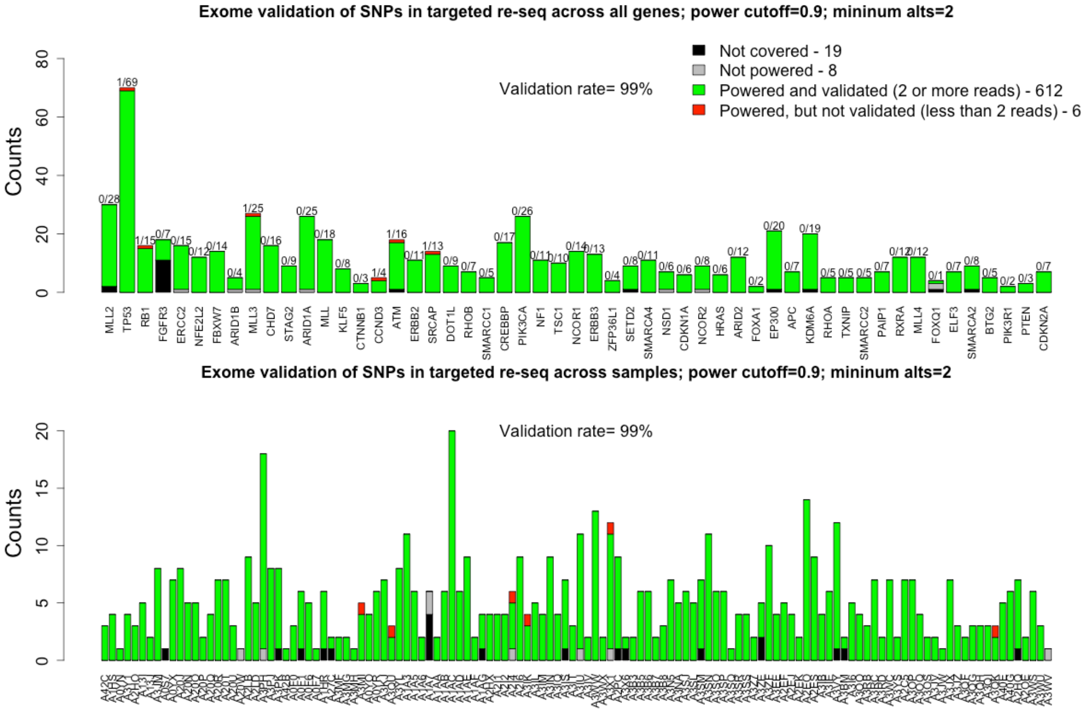


Figure S2.11.2 Validation of mutations in significantly mutated genes by targeted re-sequencing.

Validation of single nucleotide mutations in significantly mutated genes by targeted re-sequencing across genes (top) and across samples (bottom). The numbers on the top of each bargraph represent the validated mutations at powered sites, and invalidated mutations at powered sites.

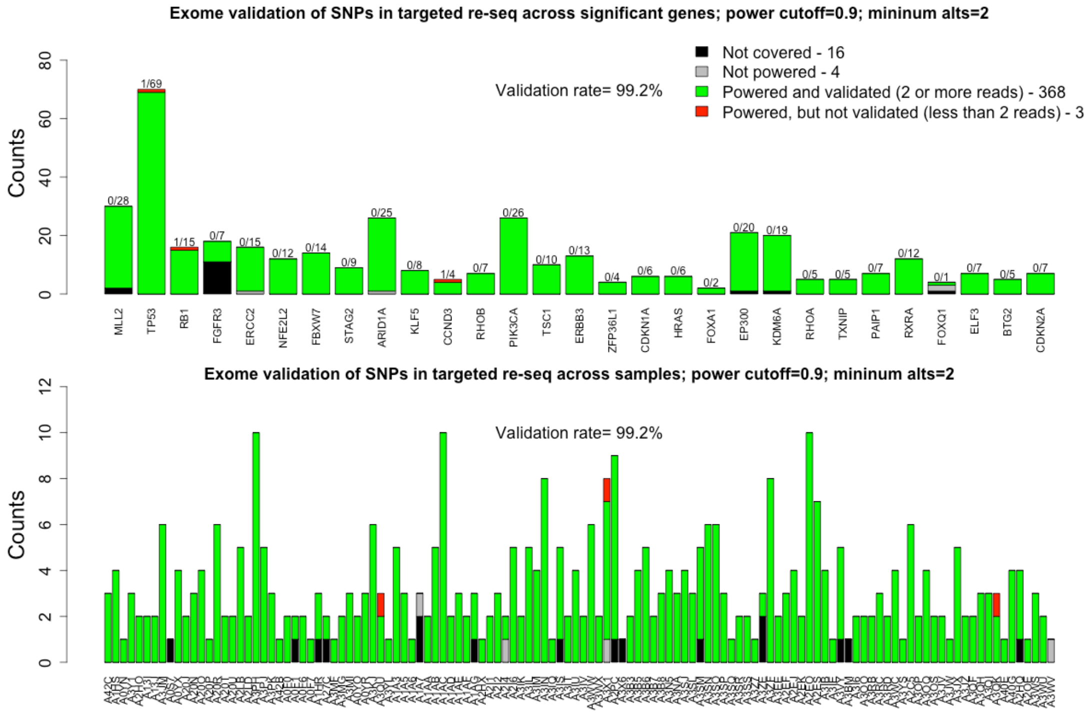


Figure S2.11.3 Validation of indel mutations in 53 genes by targeted re-sequencing.

Validation of indel mutations by targeted re-sequencing across 53 genes including significantly mutated genes, and selected cancer genes and chromatin modifying genes. Validation status is shown by gene (top) and sample (bottom). The numbers on the top of each bargraph (top) are the number of powered but not validated mutations over the number of validated mutations at powered sites.

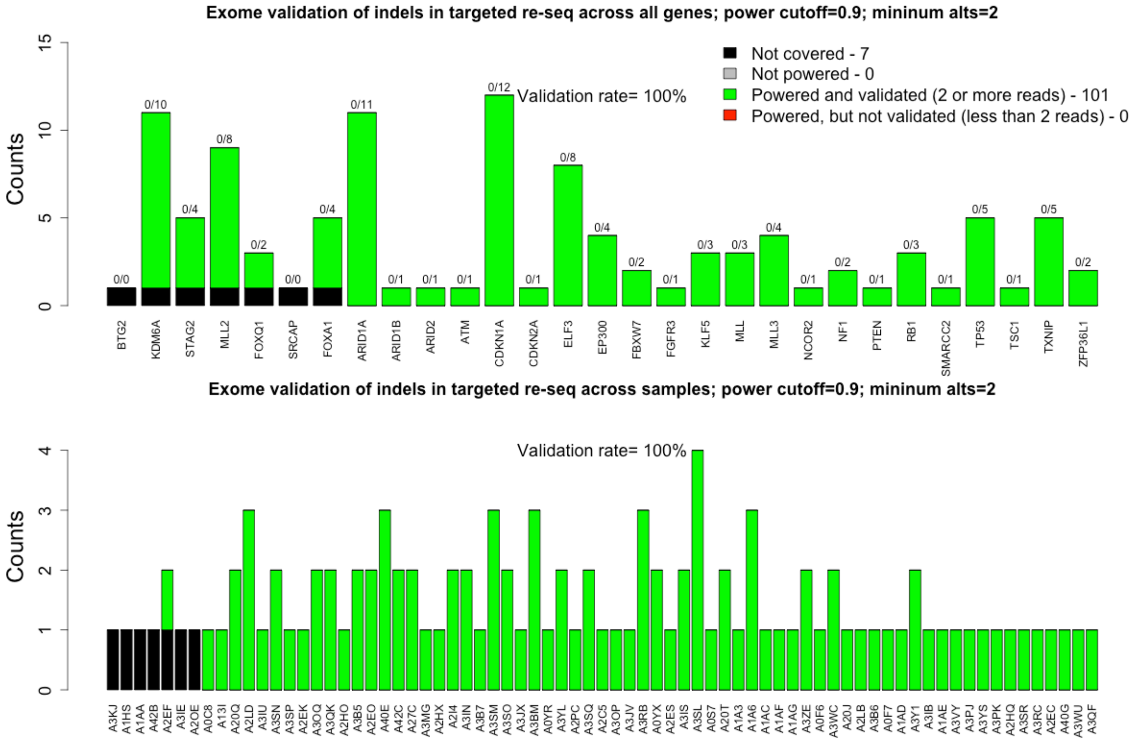


Figure S2.11.4 Validation of SNV mutations identified by Mutect analysis of whole exome sequencing using RNA-seq data for 123 samples.

Validation status for SNV mutations in 123 RNA-Seq samples in 29 significantly mutated genes (top). Validation status for all SNV mutations in 123 RNA-Seq samples shown by sample (bottom). The numbers on the top of each bargraph (top) are the number of powered but not validated mutations over the number of validated mutations at powered sites.

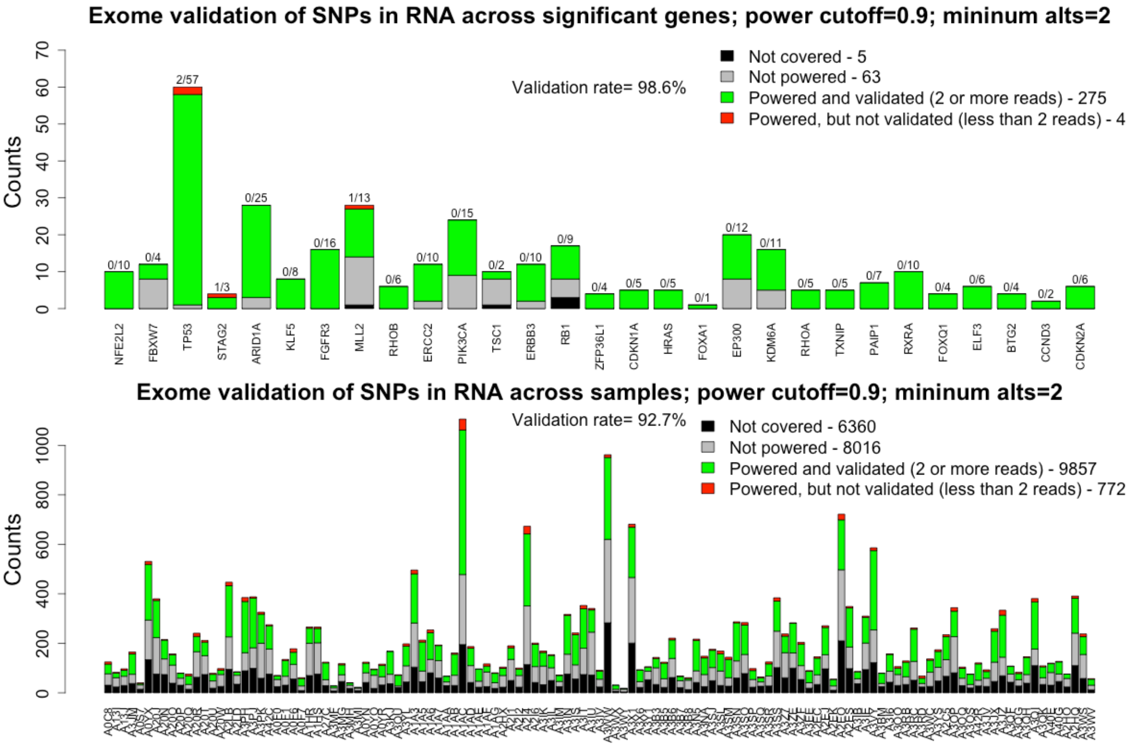


Figure S2.11.5 Validation of SNV mutations identified by Mutect analysis of whole exome sequencing using WGS data for 18 samples.

Validation status for SNV mutations in 18 WGS samples in 29 significantly mutated genes (top). Validation status for all SNV mutations in 18 WGS samples shown by sample (bottom). The numbers on the top of each bargraph (top) are the number of powered but not validated mutations over the number of validated mutations at powered sites.

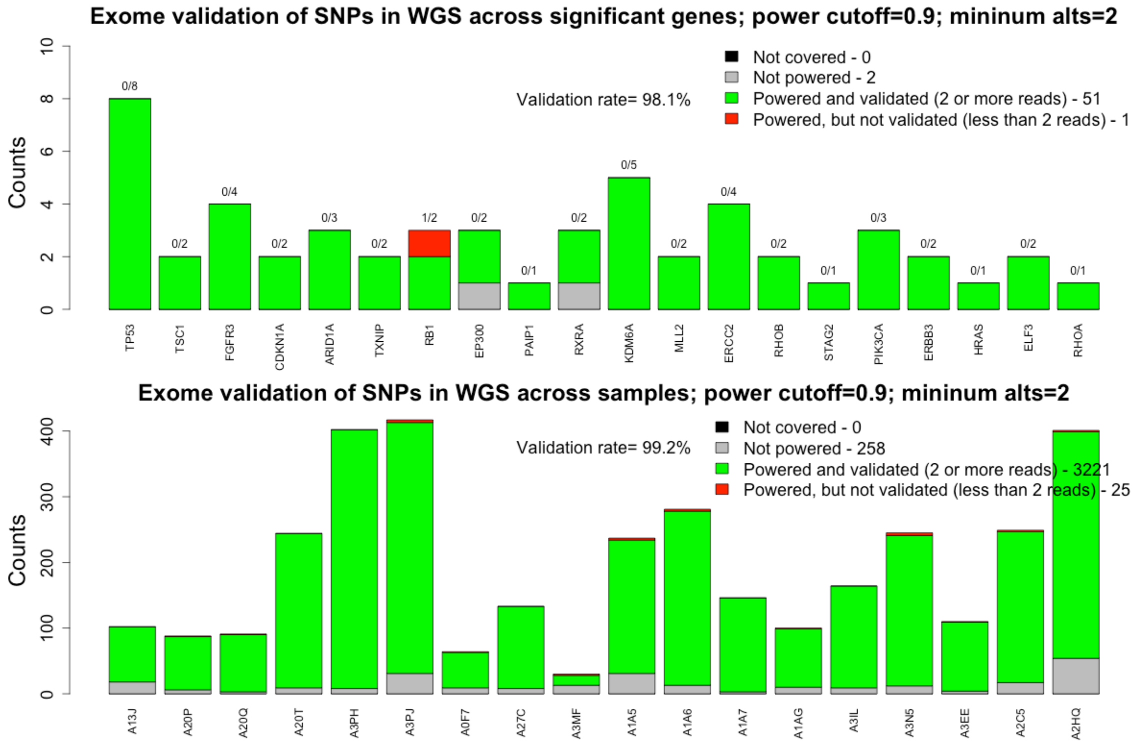


Figure S2.11.6 Combined validation results for 753 non-silent mutations in 53 genes assessed by targeted re-sequencing, RNA-Seq, and WGS data.

Validation of 753 exonic non-silent mutations in 53 genes by any of targeted re-sequencing, RNA-Seq data for 123 samples, or WGS data for 18 samples. Validation status is shown by gene (top) and sample (bottom). The numbers on the top of each bargraph (top) are the number of powered but not validated mutations over the number of validated mutations at powered sites.

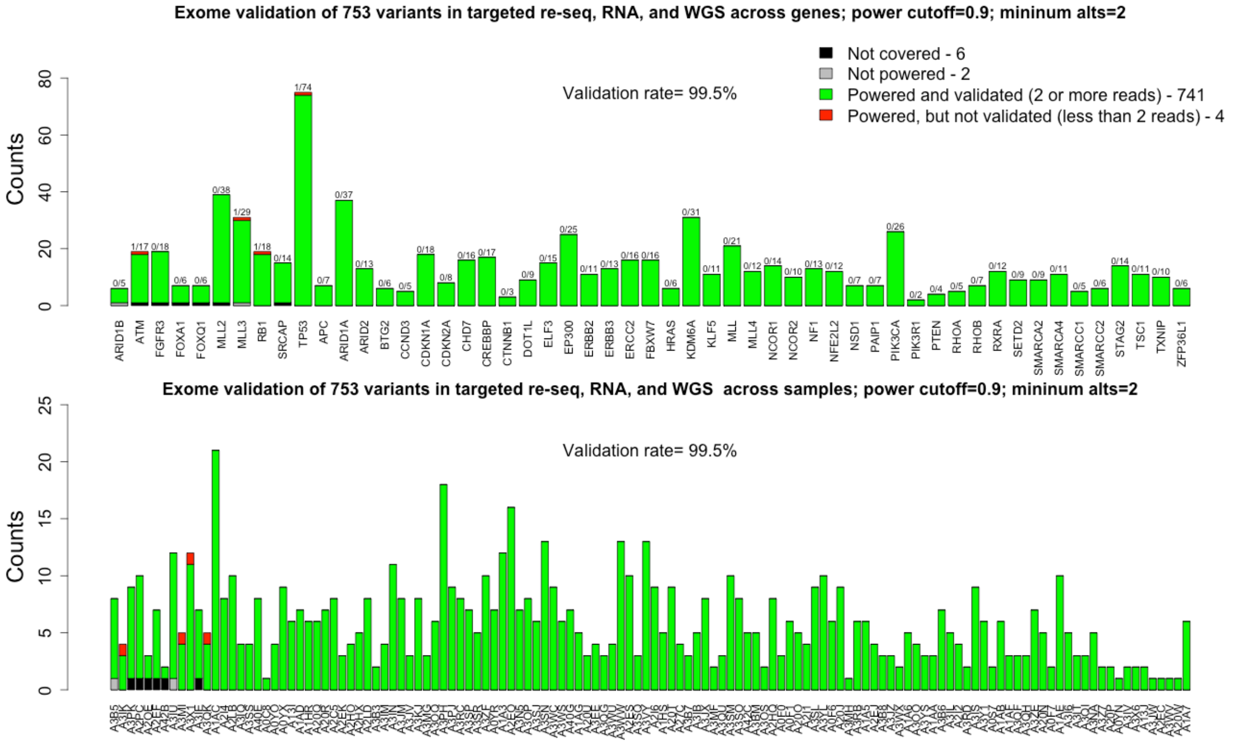
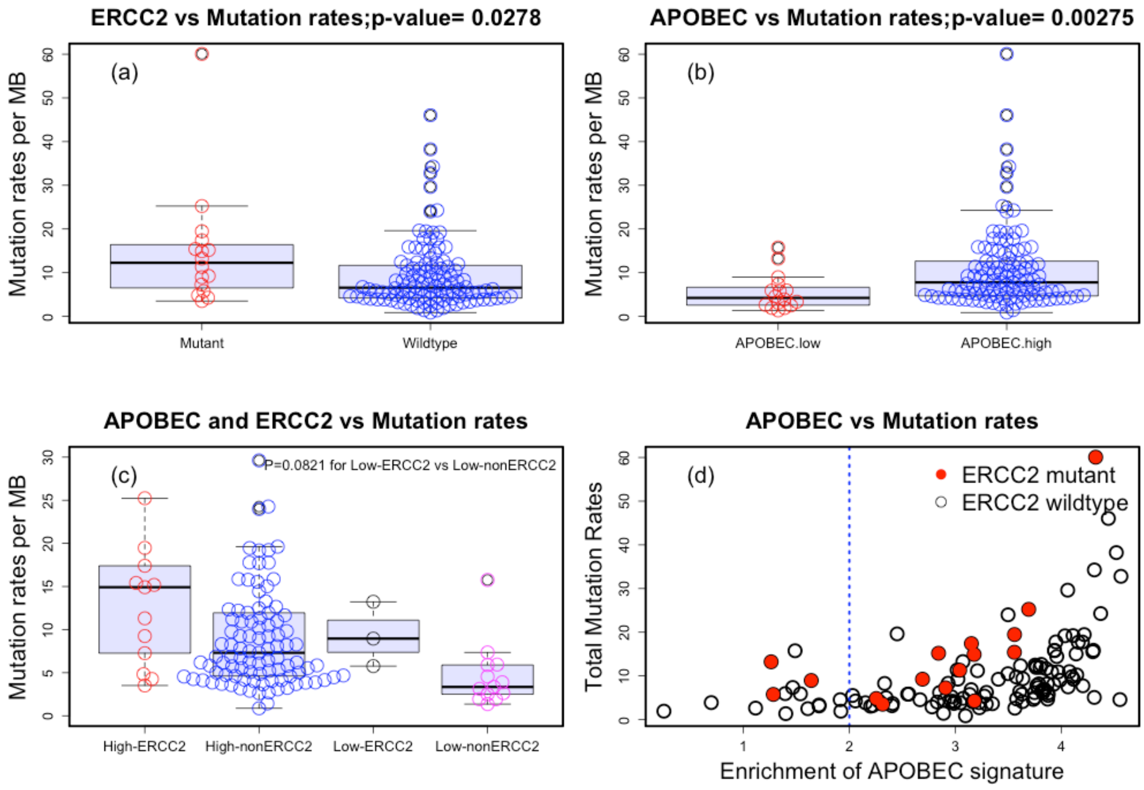


Figure S2.12 ERCC2 mutation status and APOBEC mutagenesis levels vs mutation rate.

Boxplots for mutation rate per MB for 130 samples are shown. (a) The two groups are stratified by ERCC2 mutation status (P= 0.0278 by Mann-Whitney); (b) the two groups are stratified by enrichment with APOBEC mutation signature (low (≤2) vs. high (>2); taken from Data File S12.1 and Figure S12.2b), (P=0.00275 by Mann-Whitney); (c) the four groups are stratified by both ERCC2 mutation status and APOBEC mutation signature (P=0.0821 by Mann-Whitney for Low-ERCC2 vs. Low-nonERCC2). (d) Scatterplot of mutation rate vs. APOBEC mutation signature enrichment in ERCC2 mutant samples (filled red circles) and wild-type samples (black circles). In (c), “High-ERCC2” refers to ERCC2 mutant samples that show an APOBEC mutation signature enrichment > 2; “High-nonERCC2” refers to ERCC2-wildtype samples with APOBEC mutation signature enrichment > 2; “Low-ERCC2” refers to ERCC2 mutant samples with APOBEC mutation signature enrichment ≤ 2; and “Low-nonERCC2” refers to ERCC2-wild-type samples with APOBEC mutation signature enrichment ≤ 2.



*Team Leader **Raju Kucherlapati** Rkucherlapati@partners.org, Team Members Netty Santoso, and Semin Lee*

S3: Low Pass Whole Genome Sequencing: Chromosomal rearrangement

Text S3.1 Supplementary Methods for Genome Sequencing:

WGS (low-pass) Based Analysis of Structural Variations. From 700 to 500 ng of each sample gDNA were sheared using Covaris E220 to about 250 bp fragments, than converted to a pair-end Illumina library using KAPA Bio kits with Caliper (PerkinElmer) robotic NGS Suite according to manufacturer's protocols. All libraries were sequenced by HiSeq 2000 using one sample – one lane, pair-end 2x51 bp setup. Tumor and its matching normal were usually loaded to the same flowcell. Average sequence coverage was found to be 6.07, read quality 38.6, 94% reads mapped. Raw data were converted to FASTQ format then were fed to BWA alignment software to generate .bam files. **Identification of copy number variants.** To characterize somatic copy number alterations in the tumor genome, we applied a new algorithm called BIC-seq to low-coverage whole genome sequencing data. First, we counted the uniquely-aligned reads in fixed-size, non-overlapping windows along the genome. Given these bins with read counts for tumor and matched normal genomes, BIC-seq attempts to iteratively combine neighboring bins with similar copy numbers. Whether the two neighboring bins should be merged is based on Bayesian Information Criteria (BIC), a statistical criterion measuring both fitness and complexity of a statistical model. Segmentation stops when no merging of windows improves BIC, and the boundaries of the windows are reported as a final set of copy number breakpoints. Segments with copy ratio difference smaller than 0.1 (log2 scale) between tumor and normal genomes were merged in the post-processing step to avoid excessive refinement of altered regions with high read counts.

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

We detected 2529 candidate structural variant (inter, intra, del, inv) events (average=22.18 /tumor). Among 1973 translocation events that involved at least one gene, 820 had one of the breakpoints in an intergenic region, whereas the remaining 1153 juxtaposed coding regions of two genes in putative fusion events. Some recurrent SVs and genes involved are listed (Table S3.1)

Validations of translocations hits. To understand the translocations at the structural level, we PCR amplified the junction fragments using primers from regions of the two chromosomes close to the region of putative breakpoints and the DNA from this product was subjected to sequencing using the Sanger method on a capillary electrophoresis. By using this approach, we successfully validated 62 out of 109 candidate translocations (57% validation rate). We also attempted to validate the translocations by detection of reads that span the translocation junction (split reads) through MEERKAT. We found 334 out of 1153 gene-gene SV events that have the split reads (detected by both MEERKAT and BreakDancer). Finally, we also confirmed our SVs events with RNA seq. Based on this approach, we found 33 gene-gene SV events that are detected by all MEERKAT, BreakDancer and RNAseq, while another 44 gene-gene SVs were detected by both BreakDancer and RNAseq.

Table S3.1 Genes involved in recurrent translocations (gene-gene) in bladder cancer from 114 T/N pairs

Genes	Type	T/N pairs
CDKAL1	CTX,INV,ITX	8
FLJ22536	INV,CTX	7
SHANK2	CTX,DEL,ITX,INV	7
TTC28	CTX	7
LOC285045	CTX	5
PHACTR1	CTX,INV	5
TACC3	ITX,INV	2
TACC3-FGFR3	ITX	3
FHIT	ITX,INV,CTX,DEL	4
IKZF3	ITX,INV,CTX	4
NOS1AP	ITX,CTX,INV	4
PTPRD	INV,DEL,CTX,ITX	4
COPA	INV, CTX,DEL	3
CPM	ITX,INV,CTX,DEL	3
CPSF6	INV	3
DLG2	CTX	1
CALN-DLG2	CTX	2
ERBB2	ITX,CTX	3
FRS2	INV,CTX	3
MTAP-CDKN2BAS	DEL	3
PPFIA1	CTX,ITX	3
SLC26A3	CTX	3
WWOX	ITX,CTX	3
ZMAT4	CTX,ITX	3

ITX: Intrachromosomal translocation
CTX: Interchromosomal translocation
DEL: Deletion
INV: Inversion

Team Lead: ***Katherine Hoadley*** *hoadley@med.unc.edu Team Members: Wei Zhang, Yuexin Liu, Bradley Broom, and Rehan Akbani*

S4: RNA sequencing:

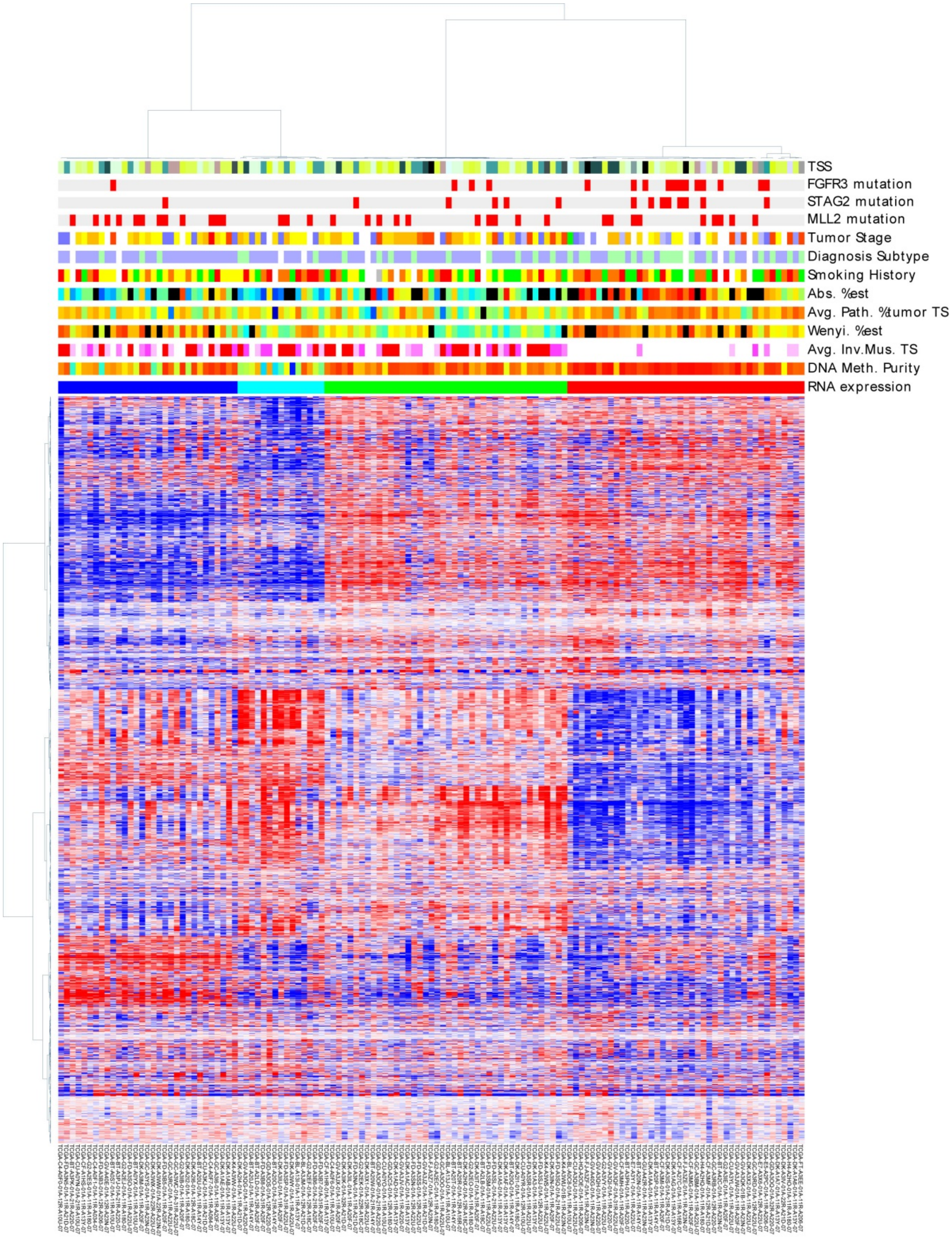
Text S4.1 Expression quantification

RNA was extracted, prepared into mRNA libraries, and sequenced by Illumina HiSeq resulting in paired 50nt reads, and subjected to quality control as previously described[1]. RNA reads were aligned to the hg19 genome assembly using Mapsplice.[2] RNA fusion events were automatically detected by MapSplice as previously described[1] Gene expression was quantified for the transcript models corresponding to the TCGA GAF2.1[3], using RSEM[4] and normalized within-sample to a fixed upper quartile. For further details on this processing, refer to Description file at the DCC data portal under the V2_MapSpliceRSEM workflow.[5] Data for genes were median centered across samples for down stream analysis.

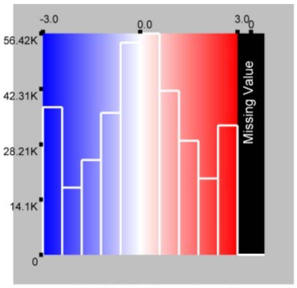
Text S4.2 Unsupervised clustering

We performed unsupervised clustering on the gene expression data using a bootstrapped ensemble clustering algorithm that merges the output of hierarchical and k-means clustering. The method clustered the samples by taking 2000 bootstrap resamples of the genes and counted how frequently two samples occurred in the same cluster using both hierarchical clustering and k-means clustering. The sample co-occurrence matrix was then clustered using hierarchical clustering. In both hierarchical clustering steps, the distance metric was Pearson correlation squared and Ward was used as the linkage algorithm. The resampling analysis identified four robust sample clusters. The four clusters and their protein expression patterns are shown in (Supplemental Figure S4.1)

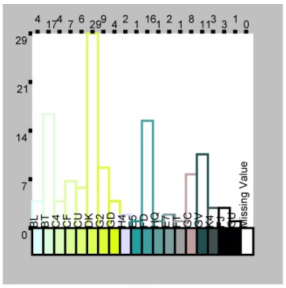
Figure S4.1: RNA expression



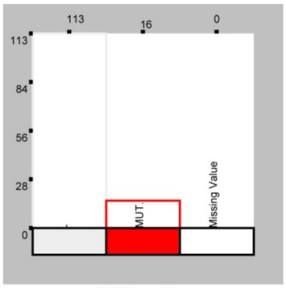
RNA expression - Illumina HiSeq for 2708 variable genes in 129 TCGA BLCA samples.



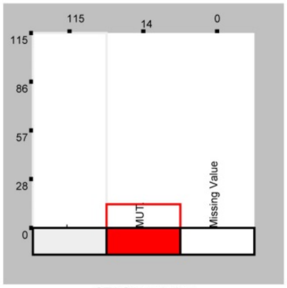
Row Centered map for cm1



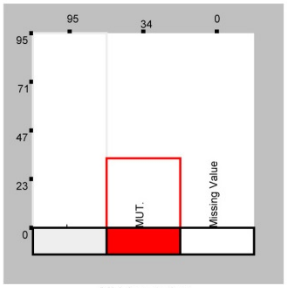
TSS



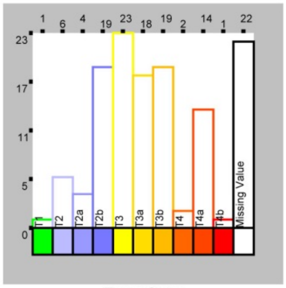
FGFR3 mutation



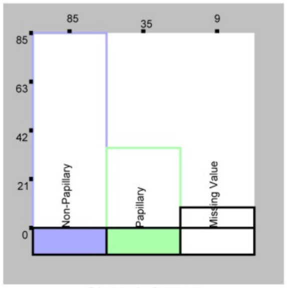
STAG2 mutation



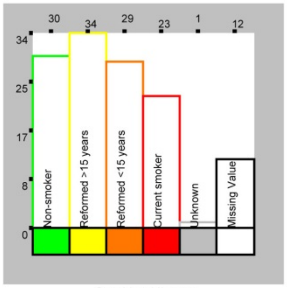
MLL2 mutation



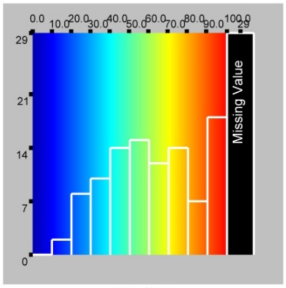
Tumor Stage



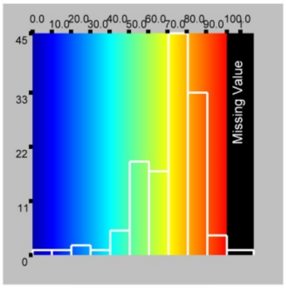
Diagnosis Subtype



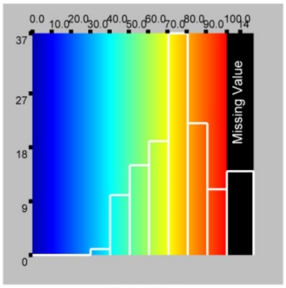
Smoking History



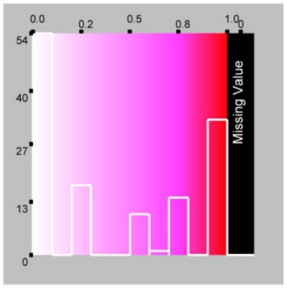
Abs. %est



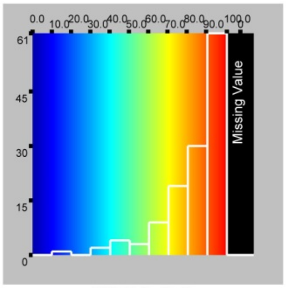
Avg. Path. %tumor TS



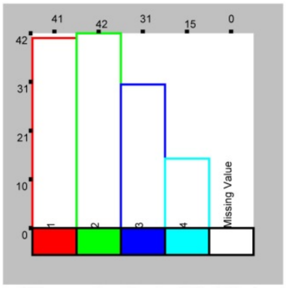
Wenyi. %est



Avg. Inv.Mus. TS



DNA Meth. Purity

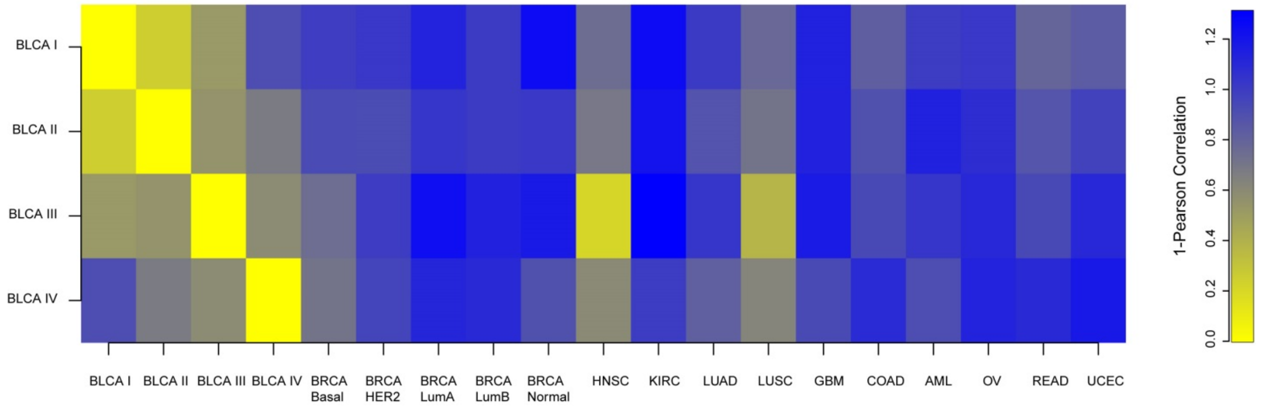


RNA expression - Illumina (HiSeq) Cluster

Text S4.3 Expression Correlation of Bladder tumors to other Tumor types

Using all genes, the 1-Pearson correlation was calculated for all pairwise correlations. Bladder subtype III is highly correlated with Head and Neck Squamous cancer, Lung Squamous cancer and Basal-like breast cancer (Supplemental Figure S4.2).

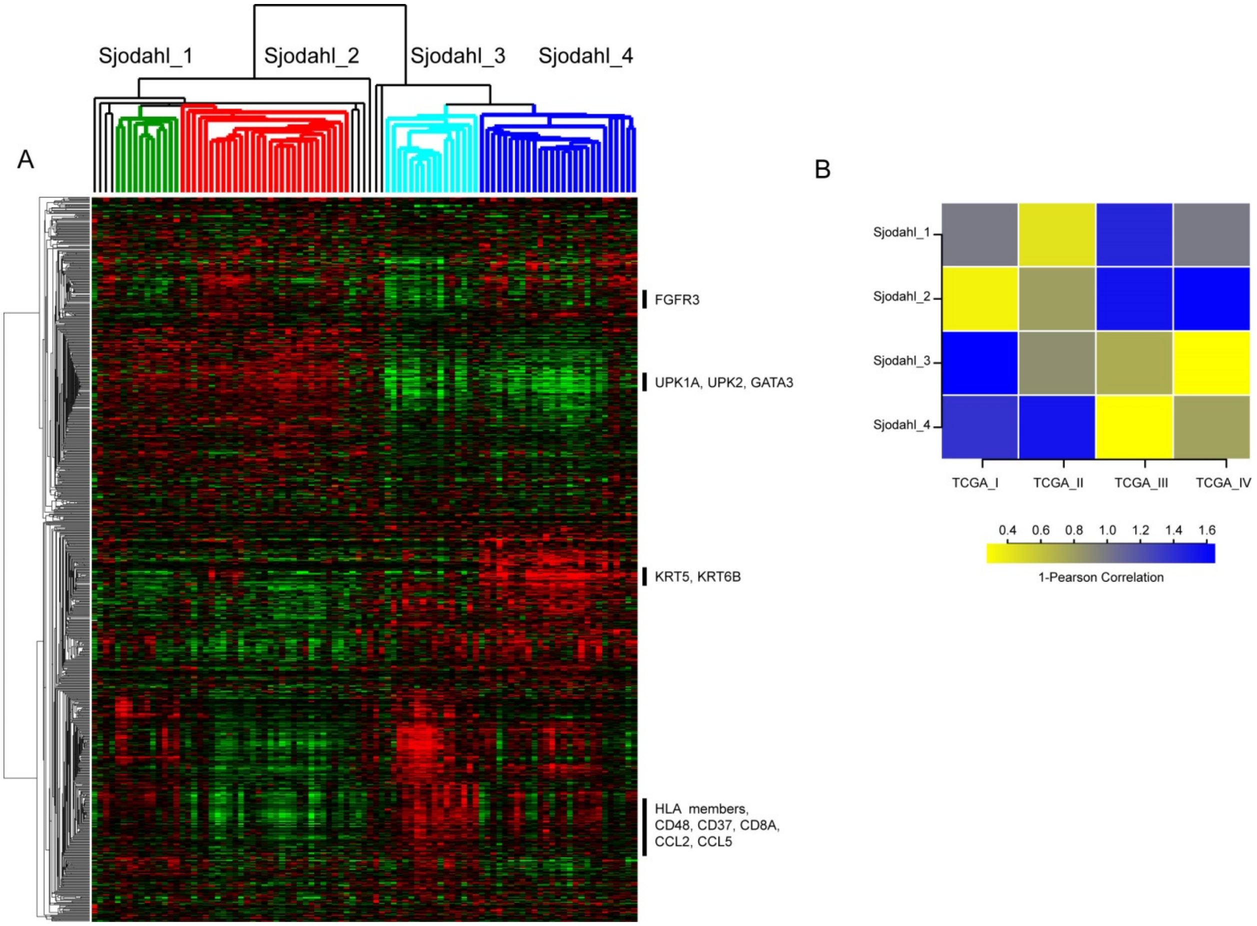
Figure S4.2: RNA Bladder subtypes and correlation to other TCGA tumor types.



Text S4.4 Validation of Unsupervised Clusters

93 muscle invasive tumors from Sjodahl et al (6) were hierarchically clustered using the gene list derived for Supplemental Figure S4.1. Clustering revealed 4 distinct subgroups. Pairwise correlations were made between the 4 Sjodahl and 4 TCGA subtypes using the median gene expression of the genes used for clustering the TCGA dataset (Supplemental Figure S4.3).

Figure S4.3: RNA expression clusters in a validation data set



(A) Muscle invasive tumors (n=93) from Sjodahl et al (6) hierarchically clustered using the gene list derived for Supplemental Figure S4.1. The expression of genes that defined the subgroups seen in the TCGA data (Supplemental Figure S4.1) also seemed to have subgroup specific expression patterns including examples such as FGFR3, UPK1A, UPK2, GATA3, KRT5, KRT6B, and immune related genes (ie. HLA members, CD48, CD8A, CCL2, CCL5) are highlighted. Subgroups were selected from the nodes (highlighted in green, red, cyan, and blue for Sjodahl subgroup 1, 2, 3, and 4 respectively). (B) Pairwise correlations were made between the 4 Sjodahl and 4 TCGA subtypes using the median gene expression of the genes used for clustering the TCGA dataset (Supplemental Figure S4.1). This correlation was then visualized by plotting the 1-pearson correlation for all pairwise comparisons (yellow = correlation, blue=anti-correlation).

Text S4.5 Correlation of expression subtypes with clinical outcome.

Cox proportional hazards analysis was performed in R (7) to estimate the hazard ratio associated with cluster expression and known clinical variables (Supplemental Table S4.1).

Table S4.1. Cox Proportional Hazards Analysis

	Univariate Analysis					Multivariate Analysis			
Variable	HR	Lower 95% CI	Upper 95% CI	p-value		HR	Lower 95% CI	Upper 95% CI	p-value
Diagnosis Subtype Non-Pap vs Pap	1.91	0.851	4.262	0.12		-	-	-	-
Regional nodes N+ vs N0	1.45	0.796	2.637	0.22		-	-	-	-
Stage pT3/4 vs pT2	1.58	0.753	3.313	0.23		-	-	-	-
Age at Diagnosis Continuous	1.04	1.007	1.07	**0.015**		1.03	1.001	1.065	**0.040**
Expression subtypes									
II vs I	1.34	0.598	3.023	0.47		1.29	0.573	2.897	0.541
III vs I	2.34	1.009	5.411	**0.048**		1.90	0.803	4.516	0.144
IV vs 1	1.68	0.644	4.382	0.29		1.68	0.644	4.388	0.288

References

1. The Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012 Sep 27;489(7417):519-25
2. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J. (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Research, 2010 Oct;38(18):e178
3. http://tcga-data.nci.nih.gov/docs/GAF/GAF.hg19.June2011.bundle/outputs/TCGA.hg19.June2011.gaf
4. Li B, Dewey CN. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Aug 4;12:323.
5. https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/blca/cgcc/unc.edu/illuminahiseq_rnaseqv2/rnaseqv2/unc.edu_BLCA.IlluminaHiSeq_RNASeqV2.mage-tab.1.9.0/DESCRIPTION.txt
6. Sjodahl, G. *et al.* A molecular taxonomy for urothelial carcinoma. *Clin Cancer Res* **18**, 3377-3386 (2012).
7. R Development Core Team, 2008. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria (2008)

Team Leaders: ***Andrew Mungall*** *amungall@bcgsc.ca and* ***Gordon Robertson*** *grobertson@bcgsc.ca*

S5.1 MicroRNA Sequencing

Text S5.1 miRNA library construction and sequencing

MicroRNA library construction and sequencing Library construction, sequencing, and analysis of sequence data were as described in (Cancer Genome Atlas Network 2012).

Unsupervised consensus clustering For miRNA-seq data, read count data for 131 tumor samples were extracted from Level 3 data archives on the TCGA Data Portal website (tcga.cancer.gov/dataportal). The set of isoform.quantification.txt files, which give read counts at base pair resolution, was processed to report total read counts for 5p and 3p strands (corresponding to miRBase v16 MIMAT identifiers, Fig. S5.1a), and read counts for each sample were normalized to RPM, i.e. to reads per million reads aligned to miRBase strands. Strands corresponding to miRNAs that had been removed from v18 miRBase (miRNA.dead) were eliminated from the data matrix. Mature and star strands were ranked by RPM variance across the samples, and the most variant 25% (214 MIMATs) were input into NMF v0.5.02 or v0.5.06 (Gaujoux and Seoighe 2010) in R v2.12.0 for unsupervised consensus clustering. The default Brunet algorithm was applied, using 50 iterations for the rank survey and 200 iterations for the clustering runs. A preferred cluster result was selected by considering profiles of cophenetic score and average silhouette width (Rousseeuw 1987) of the consensus membership matrix, for clustering solutions having between 3 and 15 clusters (Fig. S5.1b). Silhouette results were generated from the NMF consensus membership matrix using the R ‘cluster’ package v1.14.1. Silhouette width profiles were generated by reordering samples to match the sample order in the NMF heatmap, and typical vs. atypical members were identified for each unsupervised group using a silhouette width threshold set to a fraction (e.g. 0.90) of the maximum width in that group (Fig. S5.1c).

To generate abundance heatmaps for an NMF result, tumor samples in the RPM abundance matrix were ordered to correspond to the NMF output order, and abundance data for the 15 matching normal samples was added. Records were retained for the subset of 32 unique miRNA 5p or 3p strands to which NMF had assigned the top 5% of scores in each metagene in its W matrix. Using Cluster 3 (http://bonsai.hgc.jp/~mdehoon/software/cluster/), the 32 miRNA abundance profiles across tumor and normal samples were log-transformed and median-centred, then were hierarchically clustered using an absolute centred correlation and average linkage. The result was visualized with then Java Treeview (http://jtreeview.sourceforge.net/).

Purity and ploidy results were reported by Absolute (Carter et al. 2012). Association p-values for covariate contingency tables were calculated using R v3.0.1's Fisher exact test. Spearman miR-to-gene correlations were calculated with R 3.0.1's corr.test.

Differentially abundant miRNAs (Fig. S5.2) miRNA 5p and 3p strands that were differentially abundant for samples in each unsupervised tumor group, relative to samples in all other groups, were identified with SAMseq v2.0 (Li and Tibshirani 2011) in R v3.0.0, using an RPM abundance matrix as input and an FDR threshold of 0.05.

Relationships between miRNA-seq and mRNA-seq unsupervised groups (Fig. S5.4) were visualized using Bezier curves (Mathematica v9, Wolfram Research, Champaign IL).

References

Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418):61-70.

Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012; 30(5):413-21.

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

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 J, Tibshirani R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2011 Nov 28. [Epub ahead of print]

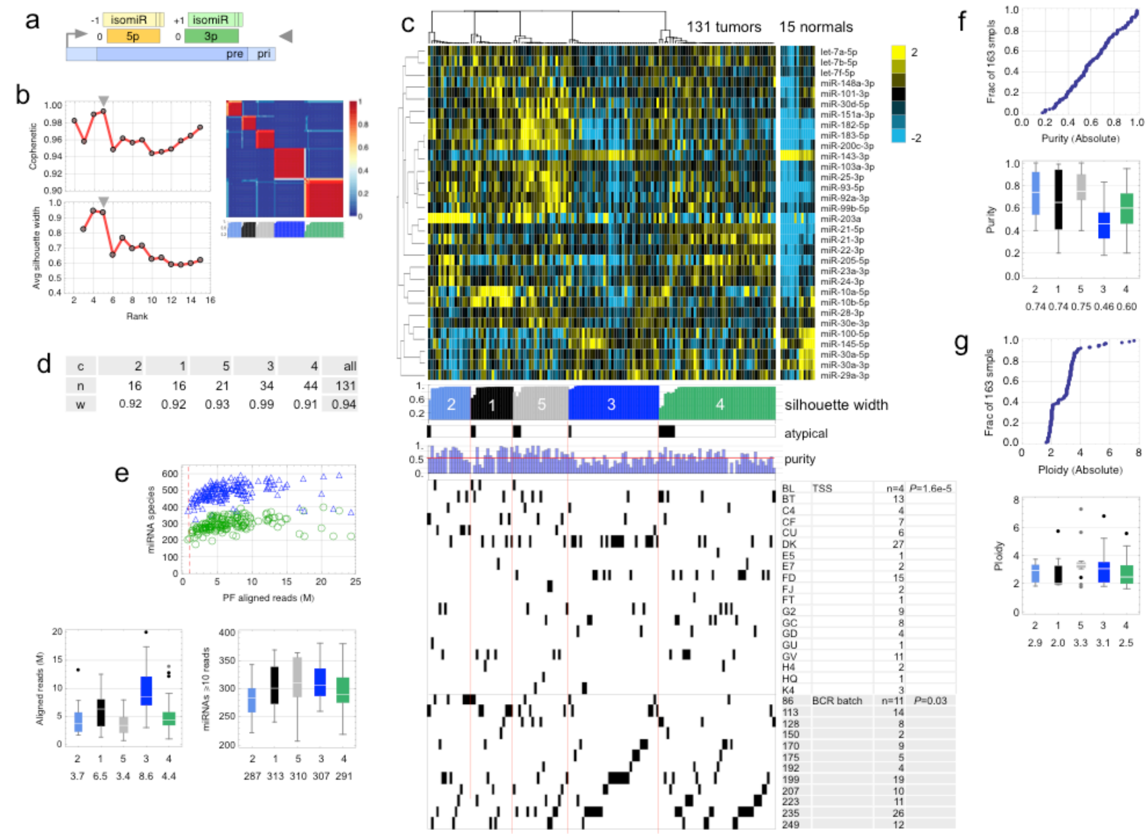
Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754-60.

Rousseeuw PJ. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics. 1987;20:53–65.

Table S5.1 Annotation priorities that are used to resolve multiple database matches for a single alignment location and for multiple alignment locations for small RNA sequencing reads.

Priority	Annotation type	Database
1 2 3 4 5	mature strand star strand precursor miRNA stemloop, from 1 to 6 bases outside the mature strand, between the mature and star strands "unannotated", any region other than the mature strand in miRNAs where no star strand is annotated	miRBase v16
6 7 8 9 10 11 12	snoRNA tRNA rRNA snRNA scRNA srpRNA Other RNA repeats	UCSC small RNAs, RepeatMasker
13 14 15 16,17	coding exons with zero annotated CDS region length 3' UTR 5' UTR coding exon, intron	UCSC genes
18 19 20 21 22 23 24 25 26	LINE SINE LTR Satellite RepeatMasker DNA RepeatMasker Low complexity RepeatMasker Simple Repeat RepeatMasker Other RepeatMasker Unknown	UCSC RepeatMasker

Figure S5.1 Unsupervised clustering of miRNA-seq data.

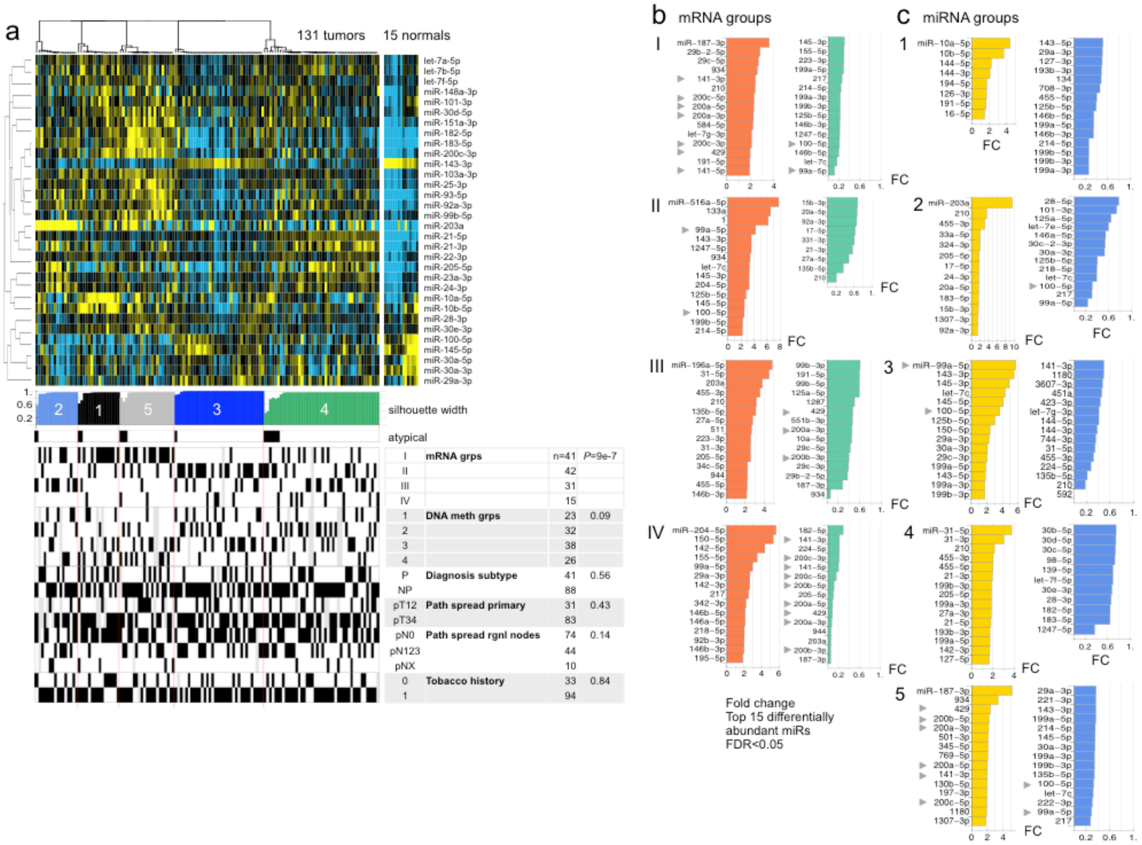


BL	TSS	n=4	P=1.6e-5
BT		13	
C4		4	
CF		7	
CU		6	
DK		27	
E5		1	
E7		2	
FD		15	
FJ		2	
FT		1	
G2		9	
GC		8	
GD		4	
GU		1	
GV		11	
H4		2	
HQ		1	
K4		3	
86	BCR batch	n=11	P=0.03
113		14	
128		8	
150		2	
170		9	
175		5	
192		4	
199		19	
207		10	
223		11	
235		26	
249		12	

c	2	1	5	3	4	all
n	16	16	21	34	44	131
w	0.92	0.92	0.93	0.99	0.91	0.94

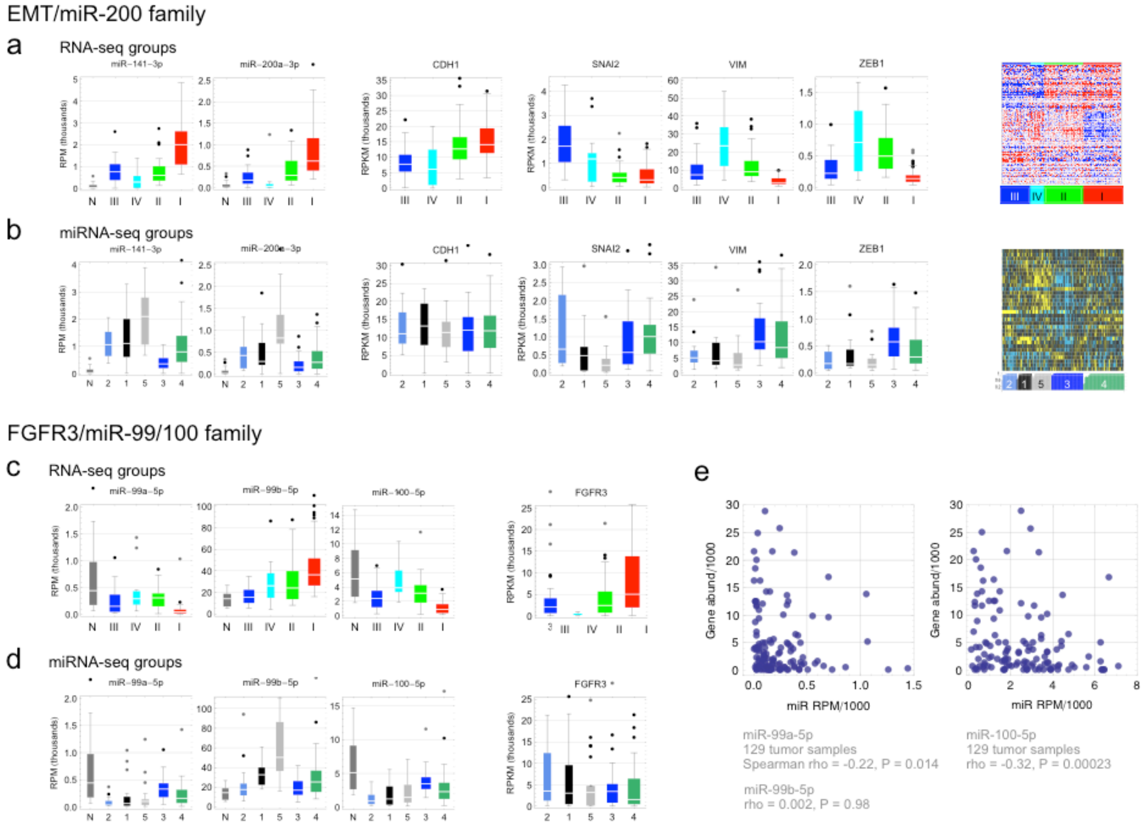
Unsupervised clustering of miRNA-seq data. a) Schematic of an miRNA primary transcript (pri), the trimmed pre-miRNA (pre), reference miRBase 5p and 3p strands, and 5' and 3' isomiR variation. The gray triangle indicates the 5p/3p-strand data representation used. **b)** From the NMF rank survey (Gaujoux and Seoighe 2010), both cophenetic correlation coefficient and average silhouette width suggests a five-group solution. The consensus membership heatmap indicates that most samples were unambiguously clustered. **c)** NMF consensus clustering. Top to bottom: normalized abundance heatmap for 32 discriminatory miRNAs; silhouette width profile; 'atypical' group members, which are samples with a width below 0.9 of the maximum in a group; a profile of sample purity (see **f**); and covariate tracks showing tissue source site and BCR batch number, with Fisher exact association P-values. **d)** Summary table of group number (c), number of samples (n) and average silhouette width (w). **e)** Top: The number of sequencing reads aligned to miRBase annotations and the number of miRBase annotations with at least 1 (blue) or 10 (green) reads aligned. Lower left: Distributions of the number of sequencing reads aligned to miRBase annotations in each unsupervised group. Lower right: Distributions of the number of miRBase annotations with at least 10 reads aligned in each group. Tables give median values. **f)** Sample purity and **g)** ploidy. Upper: distribution function. Lower: distributions in each group.

Figure S5.2 Molecular and clinical covariates, and differentially abundant miRNAs



Molecular and clinical covariates, and differentially abundant miRNAs. **a)** NMF normalized abundance heatmap with covariate tracks showing (top to bottom) mRNA groups, DNA methylation groups, diagnosis subtype (P=papillary vs. NP=non-papillary), pathlogic spread primary tumor (pT1 or 2 vs. pT3 or 4), pathologic spread regional nodes (pN0 vs. pN1, 2 or 3), and tobacco smoking history (0=never smoked vs. 1=all other categories). P-values are from Fisher exact tests. **b)** Fold changes for miRNA 5p or 3p strands that are differentially abundant (FDR<0.05) for each mRNA sample group (i.e. cluster) relative to all other tumor samples. Up to 15 of the largest (red) and smallest (green) fold changes (FC) are shown. **c)** Fold changes for miRNA 5p or 3p strands that are differentially abundant (FDR<0.05) for each miRNA sample group relative to all other tumor samples. Up to 15 of the largest (gold) and smallest (blue) fold changes (FC) are shown.

Figure S5.3 Distributions of normalized abundances of selected miRNAs and genes in the four mRNA groups and five miRNA groups



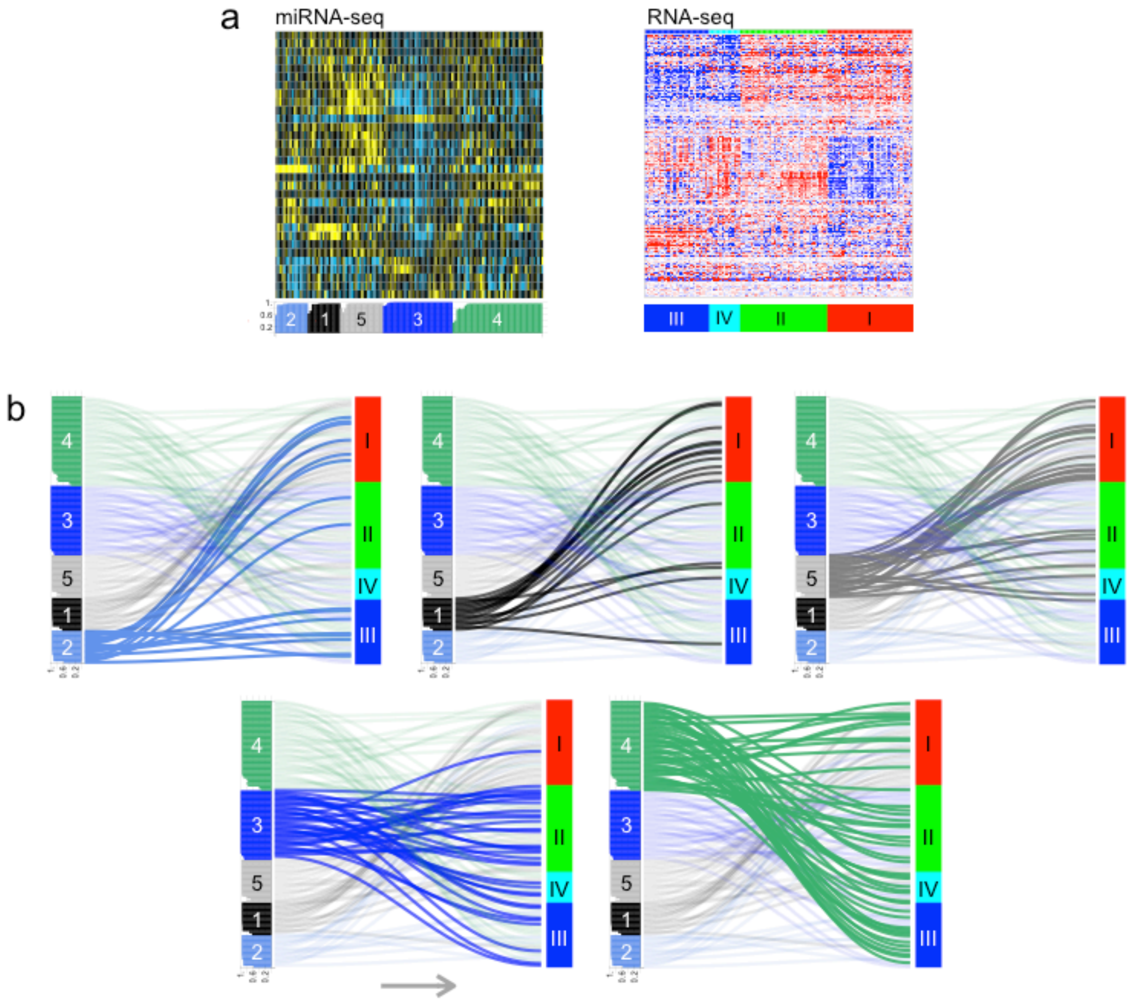
miR-99a-5p
129 tumor samples
Spearman rho = -0.22, P = 0.014

miR-99b-5p
rho = 0.002, P = 0.98

miR-100-5p
129 tumor samples
rho = -0.32, P = 0.00023

Distributions of normalized abundances of selected miRNAs and genes in the four mRNA groups and five miRNA groups. **a,b)** EMT-related miRNAs and genes. **a)** For the four mRNA-based groups, miR-141 and -200a from the miR-200 family, and CDH1, SNAI2, VIM and ZEB1. **b)** as **(a)**, but for the five miRNA-based groups. **c,d)** miR-99/100-family miRNAs and FGFR3. **c)** for the four mRNA-based groups, **d)** for the five miRNA-based groups. **e)** Scatterplots of miR vs gene abundance, with Spearman correlation results for miR-99a and -100. Spearman results are indicated for miR-99b.

Figure S5.4 Relationship between miRNA-seq and mRNA-seq unsupervised groups.



Relationship between miRNA-seq and mRNA-seq unsupervised groups. Above: normalized abundance heatmaps for five miRNA-seq clusters and four mRNA-seq clusters. Below: Bezier curves show how the samples in each miRNA-seq cluster (left) are distributed across the heatmap-ordered mRNA-seq samples (right). The curves should be read left-to-right, and are drawn with the silhouette width profile for miRNA clusters on the left. A separate Bezier diagram is shown for each miRNA cluster. In each diagram, curves for the samples in an miRNA cluster are assigned the color that cluster has in the silhouette width profile.

*Team Leader: **Jaegil Kim** jaegil@broadinstitute.org Team Member: Andrew Cherniack, David Kwiatkowski, and Jonathan Rosenberg*

S6.1: Copy Number

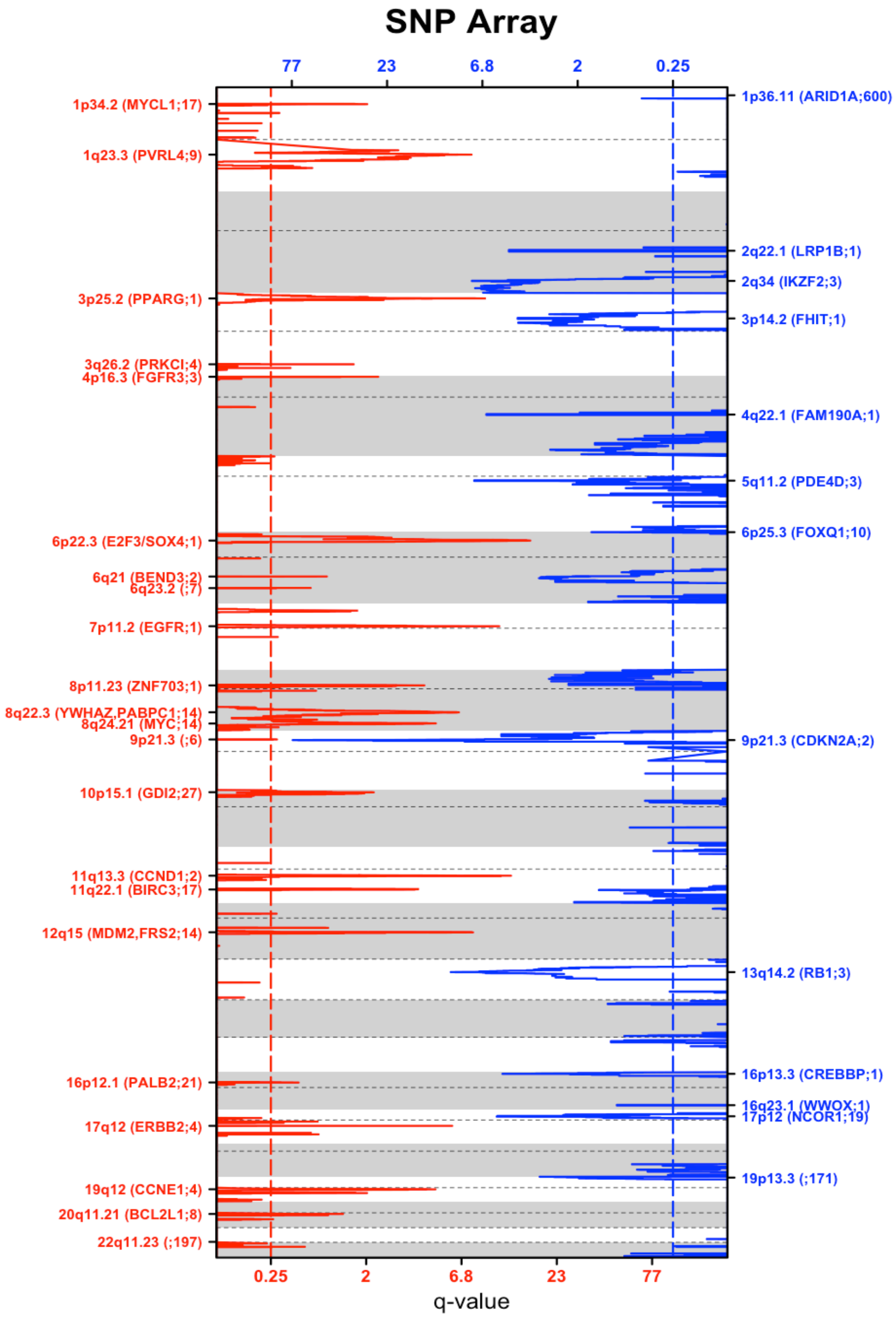
1. SNP6.0 array processing

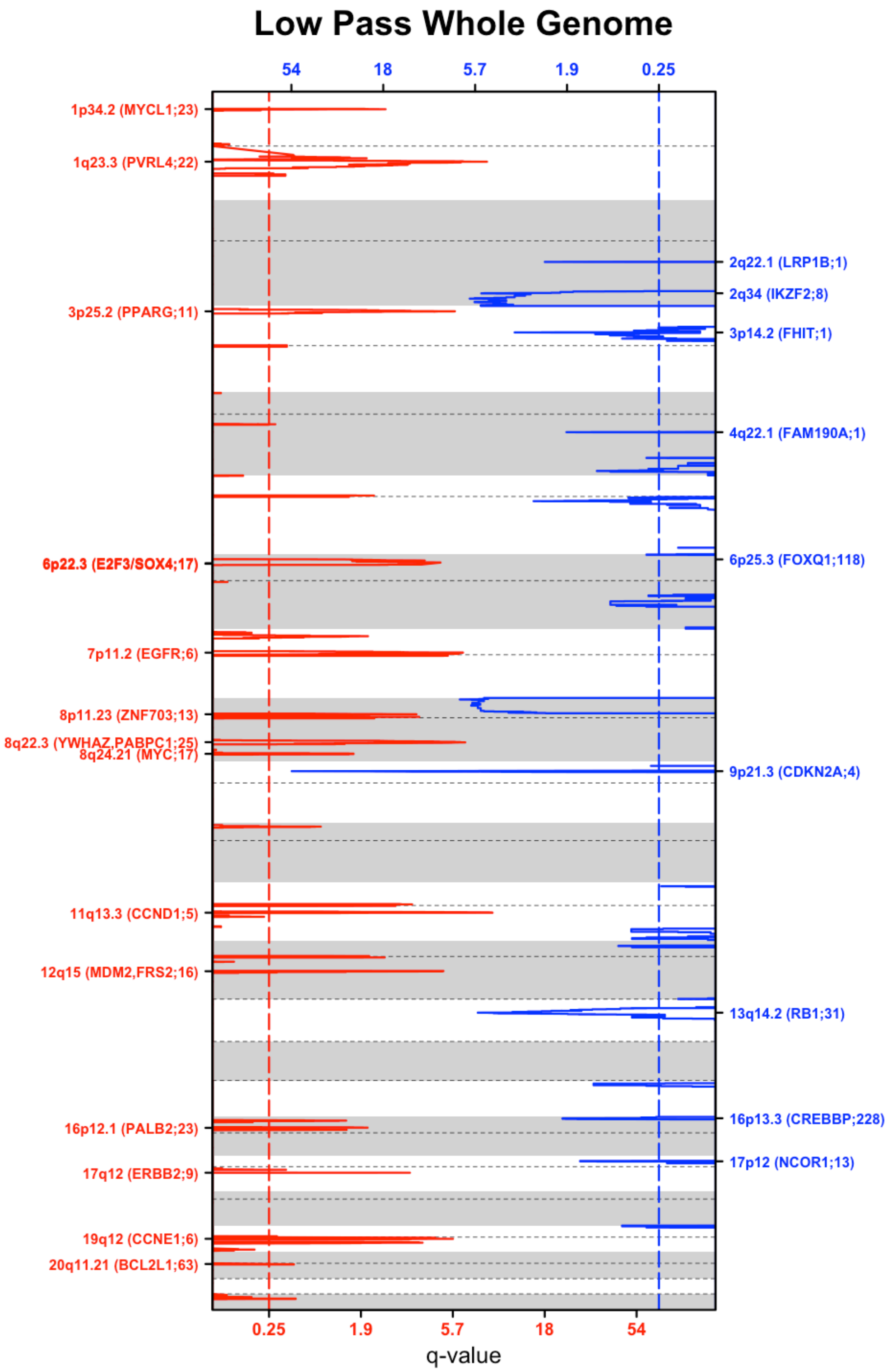
DNA from each tumor or germline-derived sample was hybridized to the Affymetrix SNP 6.0 arrays using protocols at the Genome Analysis Platform of the Broad Institute [1]. From raw .CEL files, Birdseed was used to infer a preliminary copy-number at each probe locus [2]. For each tumor, genome-wide copy number estimates were refined using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumor [3]. This linear combination of normal samples tends to match the noise profile of the tumor better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy-number profile. Individual copy-number estimates then undergo segmentation using Circular Binary Segmentation [4]. As part of this process of copy-number assessment and segmentation, regions corresponding to germline copy-number alterations were removed by applying filters generated from either the TCGA germline samples from the ovarian cancer analysis or from samples from this collection.

2. GISTIC analysis

Segmented copy number profiles for tumor and matched control DNAs were analyzed using Ziggurat Deconstruction, an algorithm that parsimoniously assigns a length and amplitude to the set of inferred copy number changes underlying each segmented copy number profile [4]. Analysis of broad copy number alterations was then conducted as previously described [2] (Data File S6.1). Significant focal copy number alterations were identified from segmented data using GISTIC 2.0 [5] (Figure S6.1; Data Files S6.1 and S6.2).

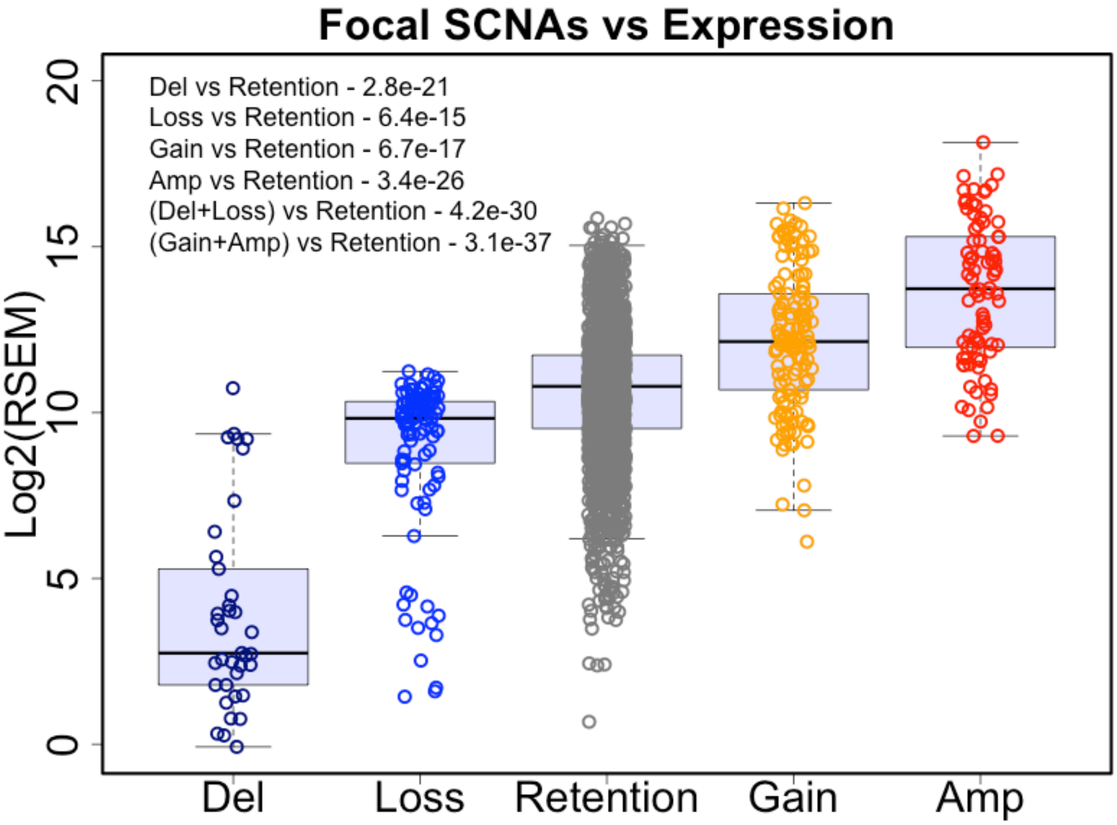
Figure S6.1 Merged GISTIC peaks for focal SCNAs in SNP6.0 Array and Low Pass WholeGenome





Graphical representation of significant amplification and deletion events in 128 bladder cancers (SNP6.0 array) and 114 bladder caners (Low Pass Whole Genome). GISTIC2.0 was used to identify statistically significant focally amplified (red) and deleted (blue) regions, which are plotted on a genome scale with chromosome 1 at top and 22 at bottom. The scale in red and blue is the false discovery rate (FDR) q-value in log format. Known or putative genes, which are the targets of amplification or deletion, are shown next to each peak. The number of candidate genes within the peak is shown following each gene name, e.g. (MYCL1; 17) means that there are 17 candidate genes within the focal peak.

Figure S6.2. Correlations between focal SCNAs vs mRNA expression



Log2(RSEM) values of twenty genes harboring focal SCNAs in Figure 1 were grouped together according to the copy number status stratified with Amp > 5 copy, 3 copy < Gain < 5 copy, 1 copy < Loss < 1.5 copy, and Del < 1 copy. Wilcoxon rank sum tests were performed to compare mRNA expression level of each group (Del, Loss, Gain, Amp, Del + Loss, and Gain + Amp) to that of the copy retention group. Each circle in the figure represents a log2 (RSEM) of focally amplified or deleted genes in a specific sample and the number in each legend denotes a p-value for comparison.

Section References

1. McCarroll, S.A. *et al*. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**:1166-1174 (2008).
2. Korn, J.M. *et al*. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**:1253-1260 (2008).
3. The Cancer Genome Atlas Research Network, Integrated genomic analyses of ovarian carcinoma. *Nature* **474**:609-615 (2011).
4. Olshen, A.B. *et al*. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**:557-572 (2004).
5. Mermel, C.H. *et al*. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**:R41 (2011).

Team Leader: ***Peter Laird*** *plaird@usc.edu Team Member: Toshinori Hinoue*

S7.1: DNA Methylation:

Text S7.1: Methylation

Array-based DNA methylation assay

We used the Illumina Infinium HumanMethylation450 (HM450) platform (Illumina, San Diego, CA) to obtain DNA methylation profiles of 131 TCGA invasive urothelial carcinoma samples and 18 adjacent histologically normal-appearing bladder tissue samples. Twelve control cell line technical replicates were also included in the assay to monitor technical variations. The Infinium HM450 assay analyzes the DNA methylation status of up to 482,421 CpG sites and 3,091 non-CpG sites throughout the genome. It covers 99% of RefSeq genes with multiple probes per gene, 96% of CpG islands from the UCSC database and their flanking regions. The DNA methylation score for each locus is presented as a beta (β) value (β = (M/(M+U)) in which M and U indicate the mean methylated and unmethylated signal intensities for each locus, respectively. β-values range from zero to one, with scores of zero indicating no DNA methylation and scores of one indicating complete DNA methylation. A detection *P* value also accompanies each data point and compares the signal intensity difference between the analytical probes and a set of negative control probes on the array. Any data point with a corresponding *P* value greater than 0.01 is deemed not to be statistically significantly different from background and is thus masked as "NA" in TCGA level 3 data packages, as detailed below. Further details on the Illumina Infinium HM450 DNA methylation assay technology has been described previously[1]. The assay probe sequences and information on each interrogated CpG/CpH site on the Infinium HM450 BeadChip are available from Illumina (www.illumina.com).

Sample and data processing

We performed bisulfite conversion on 1 µg of genomic DNA from each sample using the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, CA) according to the manufacturer's instructions. We assessed the amount of bisulfite converted DNA and completeness of bisulfite conversion using a panel of MethyLight-based quality control (QC) reactions as previously described[2]. All the TCGA samples passed our QC tests and entered the Infinium DNA methylation assay pipeline. Bisulfite-converted DNAs were whole genome amplified (WGA) and enzymatically-fragmented prior to hybridization to BeadChip arrays. BeadArrays were scanned using the Illumina iScan technology to produce IDAT files. TCGA DNA methylation data packages were generated using the *EGC.tools* R package (version 1.3.0) after processing raw IDAT files for each sample with the *methylumi* R package (version 2.3.22).

TCGA Data Packages

The data levels and the files contained in each data level package are described below and are present on the TCGA Data Portal website (http://tcga-data.nci.nih.gov/tcga/). Please note that as continuing updates of genomic databases and data archive revisions frequently become available, the data packages on TCGA Data Portal are updated accordingly.

Level 1: Level 1 data contain raw IDAT files (two per sample) as produced by the iScan system. *Level 2*: Level 2 data contain background-corrected methylated (M) and unmethylated (U) summary intensities as extracted by the *methylumi* R package. Non-detection probabilities (*P* values) were computed as the minimum of the two values (one per allele) for the empirical cumulative density function of the negative control probes in the appropriate color channel. Background correction is performed via normal-exponential deconvolution (currently not stratified by probe sequence). Multiple-batch archives have the intensities in each of the two channels multiplicatively scaled to match a reference sample (sample with R/G ratio closest to 1.0). *Level 3*: Level 3 data contain β-value calculations with HGNC gene symbol, chromosome (UCSC hg19, Feb 2009), and genomic coordinate (UCSC hg19, Feb 2009) for each targeted CpG/CpH site on the array. Probes having a common SNP (MAF > 0.01, per dbSNP build 135 via the UCSC snp135common track) within 10bp of the interrogated CpG site or having 15bp from the interrogated CpG site overlap with a repetitive element (as defined by RepeatMasker and Tandem Repeat Finder Masks based on UCSC hg19, Feb 2009) are masked as "NA" across all samples, and probes with a non-detection probability (*P* value) greater than 0.01 in a given sample are masked as "NA" on that chip. Probes that are mapped to multiple sites on hg19 are annotated as "NA" for chromosome and 0 for CpG/CpH coordinate.

The following data archives were used for the analyses described in this manuscript.

jhu-usc.edu_BLCA.HumanMethylation450.Level_3.1.8.0
jhu-usc.edu_BLCA.HumanMethylation450.Level_3.2.8.0

jhu-usc.edu_BLCA.HumanMethylation450.Level_3.3.8.0
jhu-usc.edu_BLCA.HumanMethylation450.Level_3.4.8.0
jhu-usc.edu_BLCA.HumanMethylation450.Level_3.5.8.0
jhu-usc.edu_BLCA.HumanMethylation450.Level_3.6.8.0
jhu-usc.edu_BLCA.HumanMethylation450.Level_3.7.8.0
jhu-usc.edu_BLCA.HumanMethylation450.Level_3.8.8.0
jhu-usc.edu_BLCA.HumanMethylation450.Level_3.9.8.0
jhu-usc.edu_BLCA.HumanMethylation450.Level_3.10.8.0
jhu-usc.edu_BLCA.HumanMethylation450.Level_3.11.8.0
jhu-usc.edu_BLCA.HumanMethylation450.Level_3.12.8.0
jhu-usc.edu_BLCA.HumanMethylation450.Level_3.13.8.0
jhu-usc.edu_BLCA.HumanMethylation450.mage-tab.1.8.0

Unsupervised clustering analysis of DNA methylation data

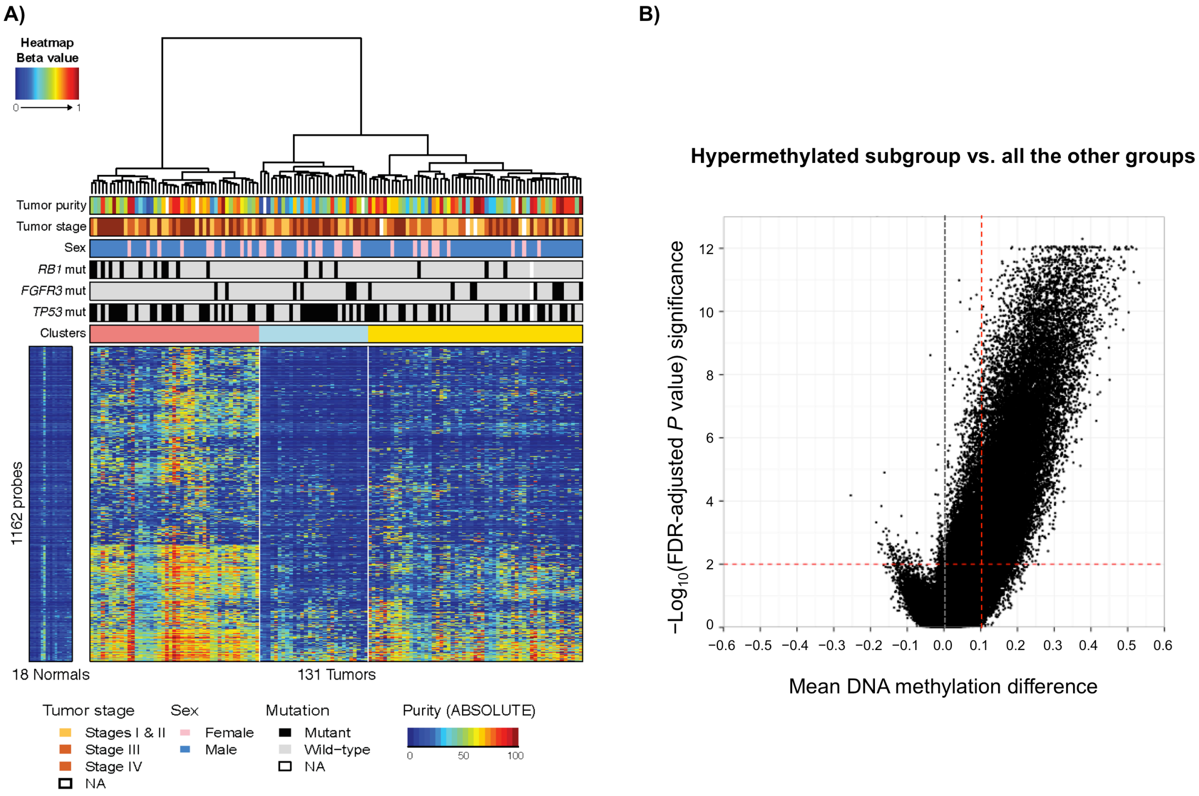
We used the Level 3 DNA methylation data contained in the packages listed above for analyses. We first removed probes which had any "NA"-masked data points and probes that were designed for sequences on X and Y chromosomes. We started with CpG sites that were located in the promoter regions (defined as the 3kb region spanning from 1,500 bp upstream to 1,500 bp downstream of the transcription start sites) and CpGs associated with CpG islands extracted from the UCSC Genome Browser (http://genome.ucsc.edu). To capture cancer-specific DNA hypermethylation events, we further eliminated sites that were methylated (mean β-value ≥0.2) in the adjacent histologically normal-appearing bladder tissues. However, a clustering analysis can be strongly confounded by the purity of tumor samples. To alleviate the potential influence of variable levels of tumor purity in our sample set on our clustering result, we dichotomized the data using a β-value of >0.3 as a threshold for positive DNA methylation. We then performed unsupervised hierarchical clustering on 11,622 CpG sites with this threshold that are methylated in at least 10% of the tumors using a binary distance metric for clustering and Ward's method for linkage. The cluster assignments were generated by cutting the resulting dendrogram.

A heatmap was generated based on the original β-values to visualize a subset (10%) of randomly selected 1,162 CpG sites used in the hierarchical clustering. The probes are arranged based on the order of unsupervised hierarchal clustering of the dichotomous data using a binary distance metric and Ward's linkage method. We performed Fisher's exact tests to quantify associations between mutations and clustering assignments results. To identify probes that show significant DNA methylation differences between the hypermethylated subgroup (Cluster 1; n= 45) and all the other groups (Clusters 2 and 3; n = 86), we performed Wilcoxon rank-sum test on β-values across all loci after "NA"-masked and sex-linked probes are eliminated (n = 380,836). The resulting *P* values were corrected using the Benjamini-Hochberg procedure. Approximately 7.7% (n = 29,371) of all loci examined exhibited statistically significant (adjusted *P* value <0.01) differences in mean DNA methylation greater than 10%.

Others

Statistical analysis and data visualization were carried out using the R/Biocoductor software packages (http://www.bioconductor.org).

Figure S7.1 Methylation Clustering



A) Unsupervised clustering of promoter CpG island methylation data revealed three major subgroups. Shown is heatmap representation of DNA methylation β-values of 1,162 randomly selected CpG sites that are located in promoter CpG island, and those that showed cancer-specific DNA hypermethylation. Data for 131 urothelial tumors and 18 adjacent histologically normal-appearing bladder tissues are plotted. DNA methylation levels are indicated by a color spectrum from dark blue (low DNA methylation) to red (high DNA methylation). Three major cluster assignments are annotated as a vertical bar above the heatmap: *lightcoral*, cluster 1 (n=45); *lightsky blue*, cluster 2 (n=29) and *yellow*, cluster 3 (n=57). Selected molecular and clinical features of each tumor sample are also shown as color bars above the heatmap, as indicated in the legends to the right of the heatmap. **B)** Volcano plot comparing DNA methylation profiles in the hypermethylated subgroup (cluster 1) and all the other groups (clusters 2 and 3 combined). Mean DNA methylation β-value differences between hypermethylated subgroup and all the other groups are plotted on the x-axis, and −1 × $\log_{10}$-transformed FDR-adjusted *P* values are plotted on the y-axis for each probe (n = 380,836). We identified 29,371 sites that are significantly more frequently hypermethylated in the hypermethylated group using FDR-adjusted *P* = 0.01 and |Δβ| = 0.1 as a cutoff for differential methylation.

Section References

1. Bibikova, M. et al. High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288-295 (2011).
2. Campan, M., Weisenberger, D. J., Trinh, B. & Laird, P. W. MethyLight. *Methods Mol Biol* **507**, 325-337 (2009).

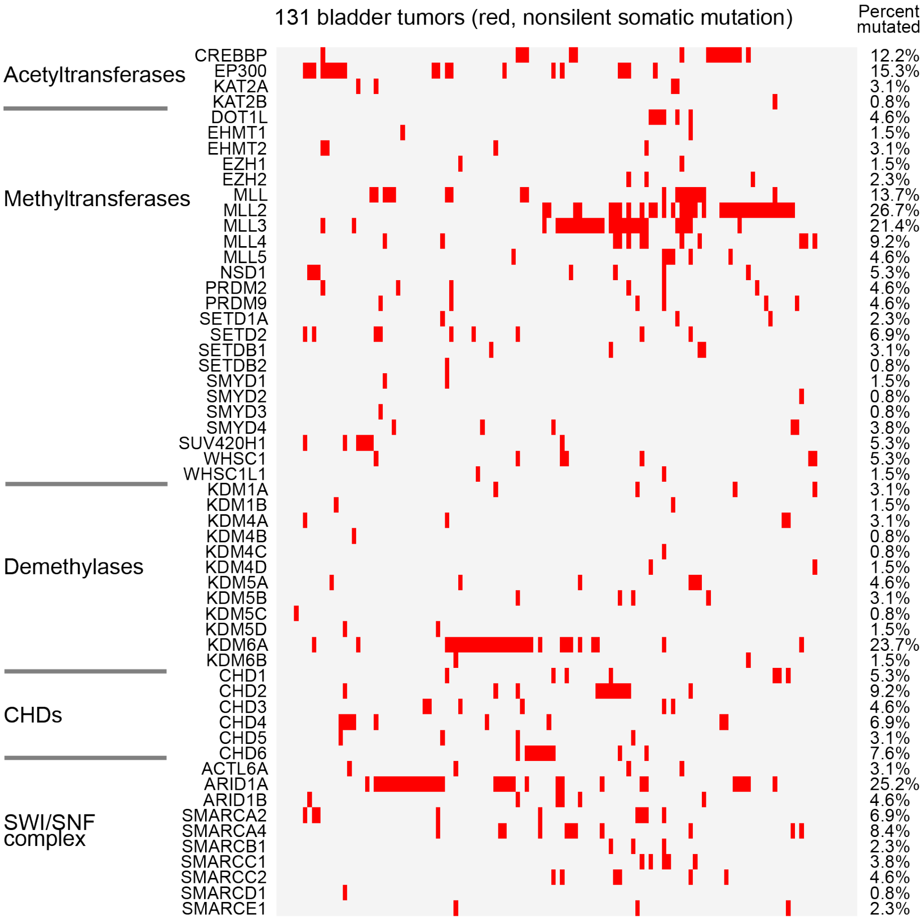
Team Leader: ***Chad Creighton*** *creighto@bcm.edu Team Members: Niki Schultz, Josh Stuart, Wei Zhang, Ilya Shmulevich, David Kwiatkowski, and Sheila Reynolds*

S8.1: Pathways and integrated analyses:

Text S8.1: Methods and Materials integrated analysis

We surveyed the mutation and copy-number data in the context of well-studied pathways, including p53/Rb, PI(3)K, and chromatin remodeling (Figure 4a). Percentages of samples were tabulated, denoting activation or inactivation involving at least one allele. For genes with known oncogenic roles, genetic alterations inferred to be activating were tabulated; for genes with tumor suppressive roles, alterations inferred to be inactivating were tabulated. For inactivating alterations, we considered either nonsilent mutation or loss of a single copy (log2[CN ratio] < -0.42, using "all_data_by_gene" Firehose output, as was applied in Figure 1). For activating alterations, we considered either gene copies of three for more (log2[CN ratio] > 0.58) or canonical activating mutation, as inferred using prior knowledge and the literature. Mutations deemed canonically activating were as follows: *PIK3CA*: p.E542K, p.E545K, p.E545Q, p.H1047L, p.Q546R, p.E453Q, p.M1043I; *ERBB2*: p.S310F, p.L313V, p.T733I, p.L755S, p.D769N, p.T862A; *FGFR3*: p.R248C, p.G380R, p.S249C, p.Y373C, p.G370C; *HRAS*: p.G13R, p.G12D, p.Q61K. Alteration percentages were calculated, using the 131 sample freeze list. For genes involved in histone modification, a summary figure provided by Lawlor and Thiele (2012) was used as the basis of the pathway diagram, with a more complete list of chromatin modifier genes being considered using a heat map representation (Figure S8.1).

Figure S8.1. Mutations involving chromatin modifier genes



(red, nonsilent somatic mutation). Percentages computed using the 131 sample freeze set

Reference

Lawlor ER, Thiele CJ. Epigenetic changes in pediatric solid tumors: promising new targets. Clin Cancer Res. 18(10):2768-79, 2012.

Text S8.2 PARADIGM Pathway Analysis

PARADIGM Pathway Analysis: Inferring gene activity from pathway analysis of copy number and expression data. Integration of copy number, mRNA expression and pathway interaction data was performed on 126 out of the 131 BLCA samples that had both copy number and RNA-seq gene expression data using the PARADIGM software (Vaske et al 2010). Briefly, this procedure infers integrated pathway levels (IPLs) for genes, complexes, and processes using pathway interactions and genomic and functional genomic data from a single patient sample. The mRNA data was converted to relative mRNA expression levels by subtracting each gene's median computed over 14 tumor-adjacent normal controls from its level observed in each patient

sample. Level 3 copy number data (segmented and normalized to reflect the difference in copy number between a gene's level detected in tumor versus normal blood) was mapped to the genome using the UCSC hg19 Knowngenes track. Gene-level copy number estimates were then derived by taking the median of all segments falling within the length of the gene. Both expression and gene-level copy number data were then rank transformed before use by the PARADIGM analysis.
Pathways were obtained in BioPax Level 3 format, and included the NCIPID and BioCarta databases from http://pid.nci.nih.gov, the Reactome database from http://reactome.org, and the set of signaling and metabolic pathways in the last public release of the KEGG database. Gene identifiers were unified by UniProt ID then converted to Human Genome Nomenclature Committee's HUGO symbol using mappings provided by HGNC (http://www.genenames.org/). Interactions from all of these sources were then combined into a merged Superimposed Pathway (SuperPathway).

Genes, complexes, and abstract processes (e.g. "cell cycle" and "apoptosis") were retained and referred to collectively as pathway features. Before merging gene features, all gene identifiers were translated into HUGO standard identifiers. All interactions, even those introducing cycles and conflicting paths were retained as PARADIGM's inference procedure has been shown to be robust to both circular and contradictory regulatory logic that may reside within pathway databases or as a result from the merging of databases. A breadth-first traversal starting from the feature with the highest number of interactions was performed to build one single component. The resulting pathway structure contained a total of 14,171 concepts, representing 6122 proteins, 6370 complexes, 1130 families, 43 RNAs, 15 miRNAs and 491 processes.

TieDIE identification of connections between genomic perturbations and transcriptional changes.

TieDIE Method. We asked whether the genomic perturbations were significantly associated with the transcriptional hubs identified by the PARADIGM (Vaske et al., 2010) analysis. To this end, we developed an integrative approach called Tied Diffusion Through Interacting Events (TieDIE) to search for significant interconnections between genomic perturbations and downstream transcriptional changes. TieDIE uses a heat diffusion process to identify relevant pathways. TieDIE can be distinguished from HotNet in that it takes as input two distinct sets and searches for interlinking pathways connecting the genes in the two sets to one another. It uses a set of sources, in this case the mutated histone-modifying genes and a set of targets, in this case transcriptional hubs, whose state in the tumor cells is assumed influenced by one or more of the upstream sources. TieDIE then diffuses heat separately from the sources and targets to determine a linker set of genes as those that gain more heat from both of the diffusion processes than would be gained from diffusion from any one set alone. The method then identifies a sub-network connecting the source, target, and linker sets by selecting edges where both adjacent nodes are in one of these sets. For each solution network, the TieDIE algorithm computes an influence score measuring the degree to which the proportion of diffused heat ends up on a common intersecting set of genes between the two input sets (manuscript in preparation). The method can also generalize to three input sets, but diffusing additional heat from a third 'signaling' set of genes and then finding linker nodes that have significantly more heat from two of the three input sets than gained by any one set alone.

TieDIE Significance Analysis. We determined if the TieDIE solutions were significant by performing a constrained permutation analysis to evaluate the significance of the resulting influence scores. One random simulation was generated by permuting the set of sources while maintaining the given set of targets. A source set that contains a lot of hubs, genes with a large number of connections, could produce a significantly interlinked network due trivially to the fact that many targets are more likely to be reached from paths emanating from hubs. The random simulation therefore needs to control for the degree distribution represented among the sources. We therefore performed a constrained permutation of the sources such that random genes selected to be the *i*th source had approximately the same number of neighbors. To do this, we sorted all of the genes by their degree. We then created non- overlapping bins by collecting *K* consecutive genes from the sorted list and putting them into the same bin together. Note that it is possible to include multiple sources in the same bin using this procedure, which makes the overall random model more conservative. The bin size, *K*, was chosen to be *n**10, where *n* was the number of sources supplied. In this case, *n*=29,23 so bins of 290 and 230 genes were created, respectively. Permutations were performed by permuting within each bin only to create swaps among genes of approximately the same size. Once all genes were swapped with another gene in the same bin, the TieDIE algorithm was repeated and a random influence score was recorded. The influence score of the network determined for the original dataset could then be compared to the background distribution obtained from this permutation analysis.

TieDIE Application to BLCA dataset. For the sources we used the 29 genes identified as significant by MutSig analysis. For the targets, Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) was then used to identify transcription factors having targets with a non-random distribution of EdgeR (Robinson et al., 2010) significance scores. This resulted in the selection of 26 transcription factors including JUN, FOS and MYC/Max. Not surprisingly since RB1 is one of the genes in the SMG, several retinoblastoma-pathway TFs were selected including HDAC1, E2F1, E2F4, and TFDP1. The TieDIE solution was found to be highly significant using a conservative background model determined with constrained permutations (Figure S8.2A). The resulting network (Figure S8.3) contained 103 genes connected by 1,233 interactions (523 HPRD-PPI, 409 regulatory, 301 component; $p < 0.008$). 24 (83%) of the sources were connected by some path in this network to all 26 of the targets, with 56 interconnecting linking genes.

To investigate the specific effects of mutations in histone-modifying genes, we selected 23 genes with known histone-modifying activity and at least 1 non-synonymous mutation, and weighted these genes by their mutation frequency to define the 'source' set. For the targets, we divided the samples into 2 groups: one containing at least 1 non-synonymous mutation in the list of histone-modifying genes (100 samples) and those without (28 samples), and used EdgeR to (Robinson et al., 2010) to rank the genes by differential expression between those groups. Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) was then used to identify transcription factors having targets with a non-random distribution of EdgeR (Robinson et al., 2010) significance scores, which resulted in the selection of 35 transcription factors, weighted by GSEA significance. TieDIE was run on this source and target set, producing a highly significant network (Figure S8.2B) with 107 nodes and 2463 edges (603 HPRD-PPI, 322 regulatory, and 1538 component; $p < 0.001$).

To investigate genes correlated to histone-modification activity, we ran PARADIGM on the samples to produce Inferred Pathway Levels (IPLs): a two-sided t-test was then calculated for each gene using the IPLs, and 12 genes were found to correlate ($p < 0.05$). We re-ran TieDIE using the original source and target sets along with this additional IPL-correlated set of genes, with the algorithm set to find linker genes with high heat in either the source and IPL-correlated sets, or the target and IPL-correlated sets. The resulting network was notably smaller (49 genes, 600 edges), and we ran an additional filtering step by performing a graph traversal from the 3 most mutated histone-modifying genes (EP300, CREBBP, MLL) to the 7 most differentially active transcriptional hubs (TP53, MYC, MAX, MYB, HES1, FOXA2, HSP90AA1). The resulting network contained 24 genes, including 4 of the IPL-correlated input nodes (SP1, HNF4A, FOXA2, CD19), and 66 edges (55 HPRD-PPI, 10 transcriptionally activating, 1 post-transcriptionally activating).

Figure S8.2. Genomic perturbations in bladder cancer

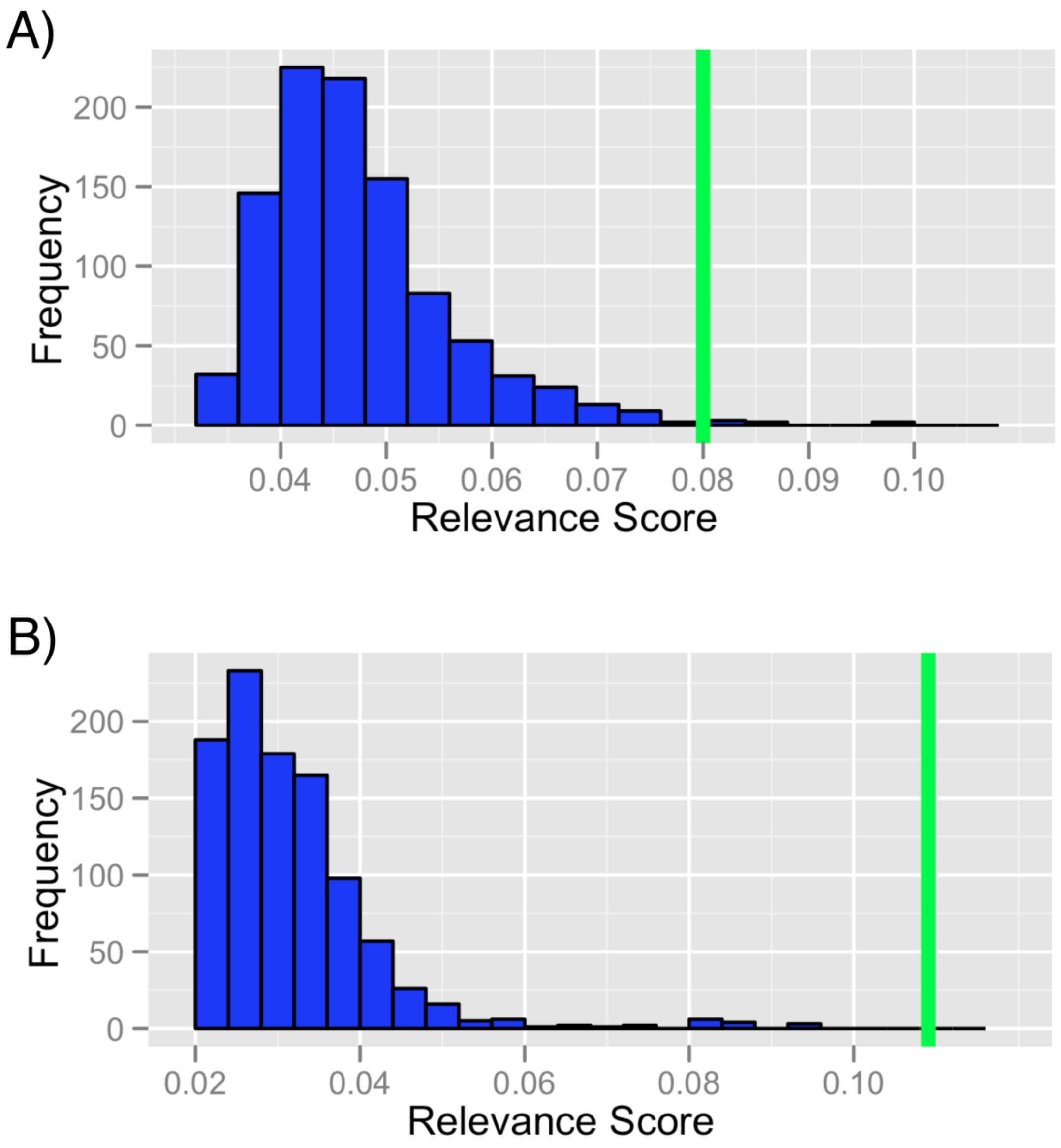


Figure S8.2. Genomic perturbations in bladder cancers are significantly associated with downstream transcriptional changes through known and novel pathway circuitry. **A)** The TieDIE algorithm was used to identify a network connecting the top 29 significantly mutated genes to transcriptional hubs identified from the identification of transcription factors with targets significantly up- or down-regulated in tumors relative to normal controls. These inputs sets are significantly close in pathway space, under 1000 random permutations of the input sets; blue bars are the scores of the permutations, the green line represents the score of the real network. **B)** The TieDIE algorithm was applied to connect 23 mutated histone-modifying genes, weighted by mutation frequency, to transcription factors with differential activity in histone-gene mutated and non-mutated samples. The

Figure S8.2.1 TieDIE network

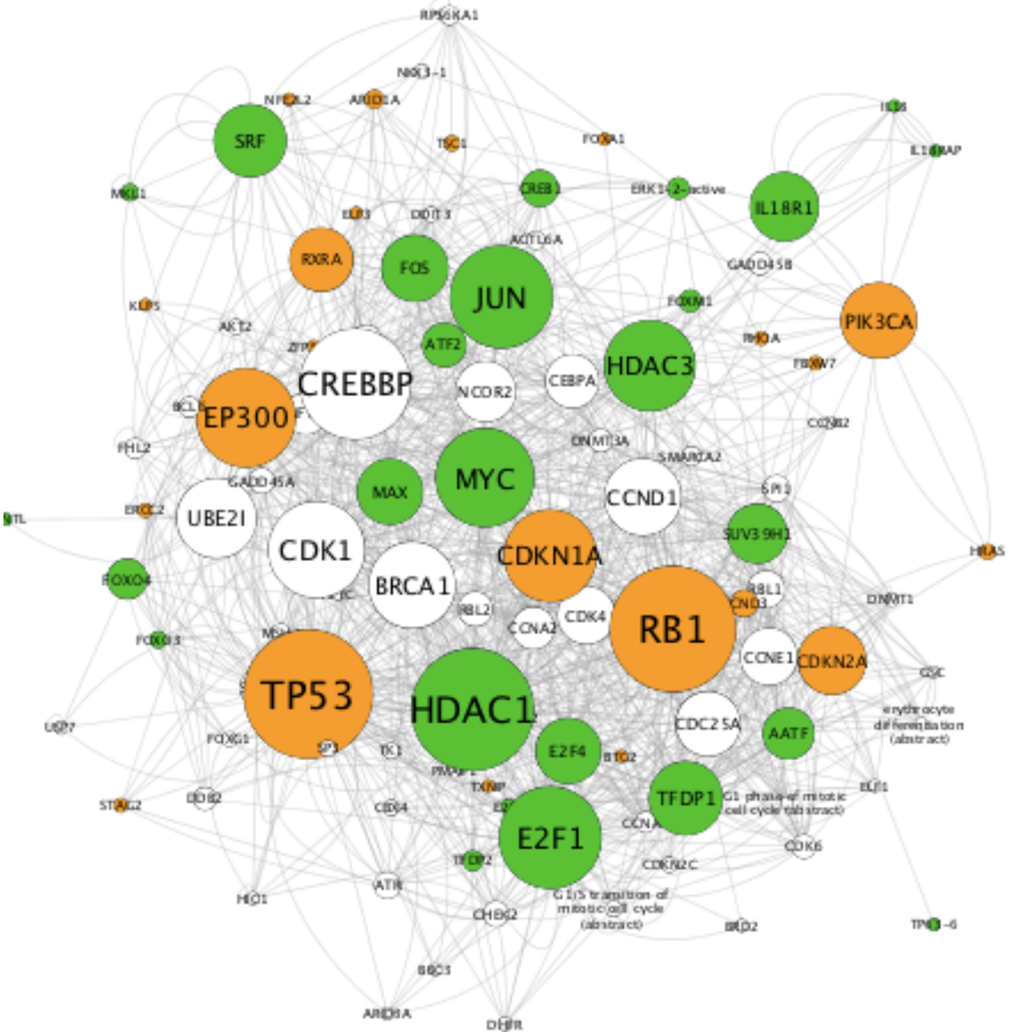


Figure S8.2.1 TieDIE network connecting Significantly Mutated Genes (SMGs) to transcription factors with altered activity in tumor samples. SMGs (orange) are shown as part of a network that connects these mutated genes to transcription factors (green) with altered activity. Solid lines indicate transcriptional regulation and dashed lines indicate protein regulation, and dotted lines indicate HPRD-PPI interactions or component associations. Size of the node reflects the betweeness centrality measure of the gene's position in the network with larger nodes as more "central" to the network solution.

The resulting linking set of genes in the full SMG solution contained over a dozen “linker” genes with high betweeness centrality measures (Figure S8.3). While many of the highly central genes by this measure correspond to the starting input set (TP53, RB1, HDAC1, E2F1), several genes were found as linkers that were not part of either the source or target sets used as input to TieDIE that increase the overall connectivity of the solution. These genes are depicted as large white nodes in Figure S8.3. Among these linkers were several cycle-cycle related genes such as CCND1, CDC25A, CDK1 and CDK4. Also among the list was the DNA repair gene, BRCA1. Finally, the linker with the highest centrality was CREBBP, which is a transcriptional co-activitator of several transcription factors that couples chromatin remodeling to transcription factor recognition involving growth, homeostasis, and development. CREBBP was also included in the chromatin-remodeling sub-network included as Figure 3B in the main text. Interestingly, even though 17 samples in the BLCA cohort have non-silent mutations in CREBBP, the level of significance did not reach the cutoff for CREBBP to merit inclusion into the SMG by Mutsig analysis. Therefore, the TieDIE analysis provides an important orthogonal perspective on the significant role of CREBBP and other linking genes in terms of their role in bladder carcinogenesis.

PARADIGM-Shift identification of gain-of-function mutation in NFE2L2 gene in BLCA. PARADIGM-Shift (Ng et al., 2012) predicts the functional impact of a mutation, gain- or loss-of-function, by interrogating the pathway surrounding a mutated gene. PARADIGM is used to score the observed downstream consequences of a gene's activity, as well as what is expected from its regulatory inputs, and the discrepancy between these two scores is used to infer the impact of mutation in the gene of interest. The significance of this ‘shift’ is determined through comparison to a random background simulation that permutes the gene labels while fixing the pathway.

We employed the PARADIGM-Shift algorithm to compare the pathway impact of mutations in NRF2/NFE2L2 (the pathway activation of which is an emerging feature present in many tumors) in BLCA. There were 126 of the white-listed samples with available copy number and expression data to run PARADIGM and PARADIGM-Shift analysis on the mutant versus non-mutant comparison, with 10 mutations annotated for NFE2L2. NFE2L2 mutation neighborhoods were selected in a supervised fashion by selecting features based on a rank ratio of the features determined by two-sided t-test. PARADIGM-Shift (P-Shift) scores for NFE2L2 (reflecting the shift between activity as inferred by up-/down- stream pathway signals) were computed as the difference in activity between two runs of PARADIGM - one in which only upstream regulators are connected (R-run) and one where only downstream targets are connected (T-run). The results of PARADIGM-Shift analysis are shown in Figure S8.4A

We then assessed the accuracy of the models by using the absolute P-Shift score as a classifier to predict NFE2L2 mutation status with 5-fold cross validation. The average AUC over the 5-folds for predicting mutations (against non-mutants) is 0.61, suggesting that PARADIGM-Shift is effective at distinguishing mutants from non-mutants. Comparing the distribution of P-Shift scores between mutants and non-mutants shows an enrichment of positive P-Shift scores in the mutant samples indicative of a gain-of-function (GOF) mutation. The significance of this GOF call was determined by running a background model in which the selected network topology is fixed, but the data is permuted. Under this background model, the GOF call was found to have a z-score of 6.2 (Supplemental Figure S8.4B-C). The entire sample set was used for training to determine the functional impact of mutations of NFE2L2 on the SuperPathway network.

Figure S8.2.2. PARADIGM-Shift Analysis of reveals gain of function mutation in NFE2L2.

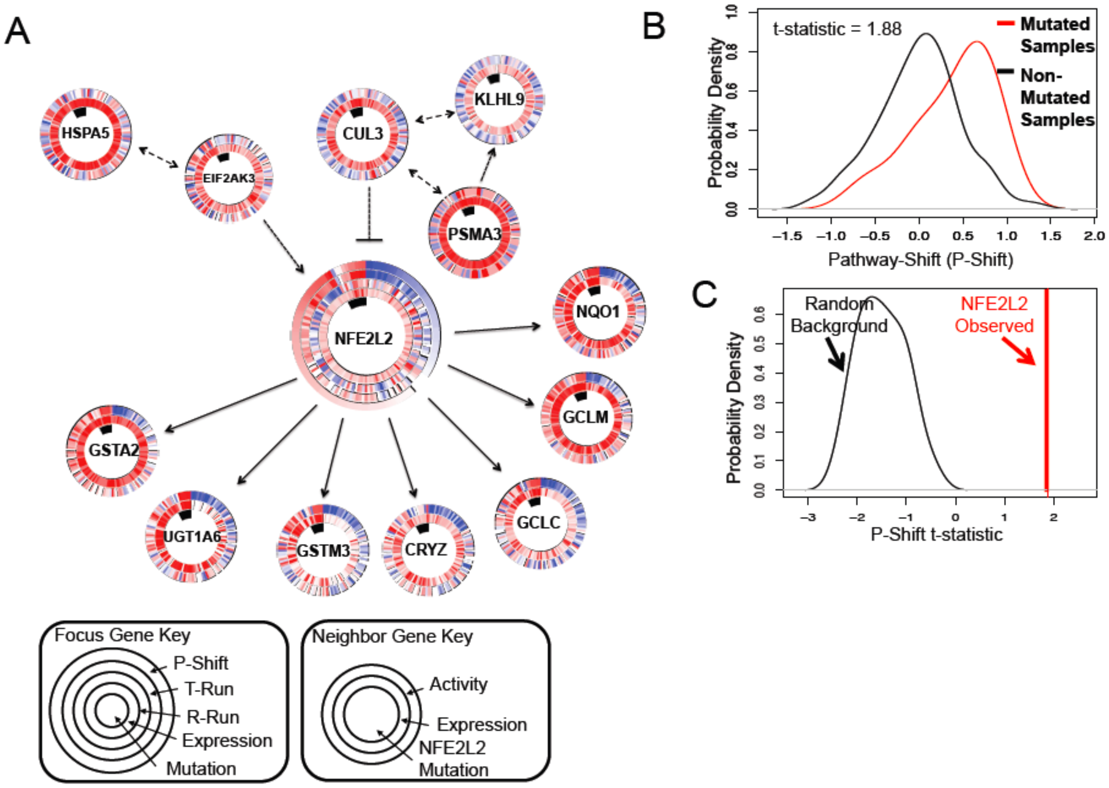


Figure S8.2.2. PARADIGM-Shift Analysis of reveals gain of function mutation in NFE2L2. **A)** Circlemap display of mutation neighborhood selected for NFE2L2 mutations. Solid lines indicate transcriptional regulation and dashed lines indicate protein regulation. Samples are sorted first by the NFE2L2 mutation status, then by PARADIGM-Shift (P-Shift) score; rings represent mutation status (inner ring), expression, inferred activity from upstream, inferred activity from downstream, and the P-Shift score. **B)** Distribution of P-Shift scores for mutated (red) and non-mutated samples (black), yielding a t-statistic of 1.88**. C)** Distribution of t-statistics of the difference in P-Shift scores between non-mutants and mutants, under the permuted background model.

REFERENCES

Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.Bioinformatics 26, 139-140 (2010). PMID: 19910308.

Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102, 15545-15550 (2005). PMID: 16199517.

Vaske, C.J. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26, i237-i245 (2010). PMID: 20529912.

Ng, S. et al. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics 28, i640-i646 (2012). PMID: 22962493.

Hast, Bridgid E. et al. Proteomic analysis of ubiquitin ligase KEAP1 reveals associated proteins that inhibit NRF2 ubiquitination Cancer Res canres. doi:10.1158/0008-5472.CAN-12-4400 (2013). PMID: 23382044.

Text S8.3: Integrated Analysis and Interactive Exploration

To gain greater insight into the underlying system-level phenomena that characterize the development and progression of urothelial carcinoma, we have integrated all of the data types produced by TCGA and described in this paper into a single “feature matrix”. From this single heterogeneous dataset, significant pairwise associations have been inferred using statistical analysis and can be visually explored in a genomic context using Regulome Explorer, an interactive web application (http://explorer.cancerregulome.org). In addition to associations that are inferred directly from the TCGA data, additional sources of information and tools are integrated into the visualization for more extensive exploration (e.g., literature-based associations, molecular interaction databases, miRBase, the UCSC Genome Browser, etc).

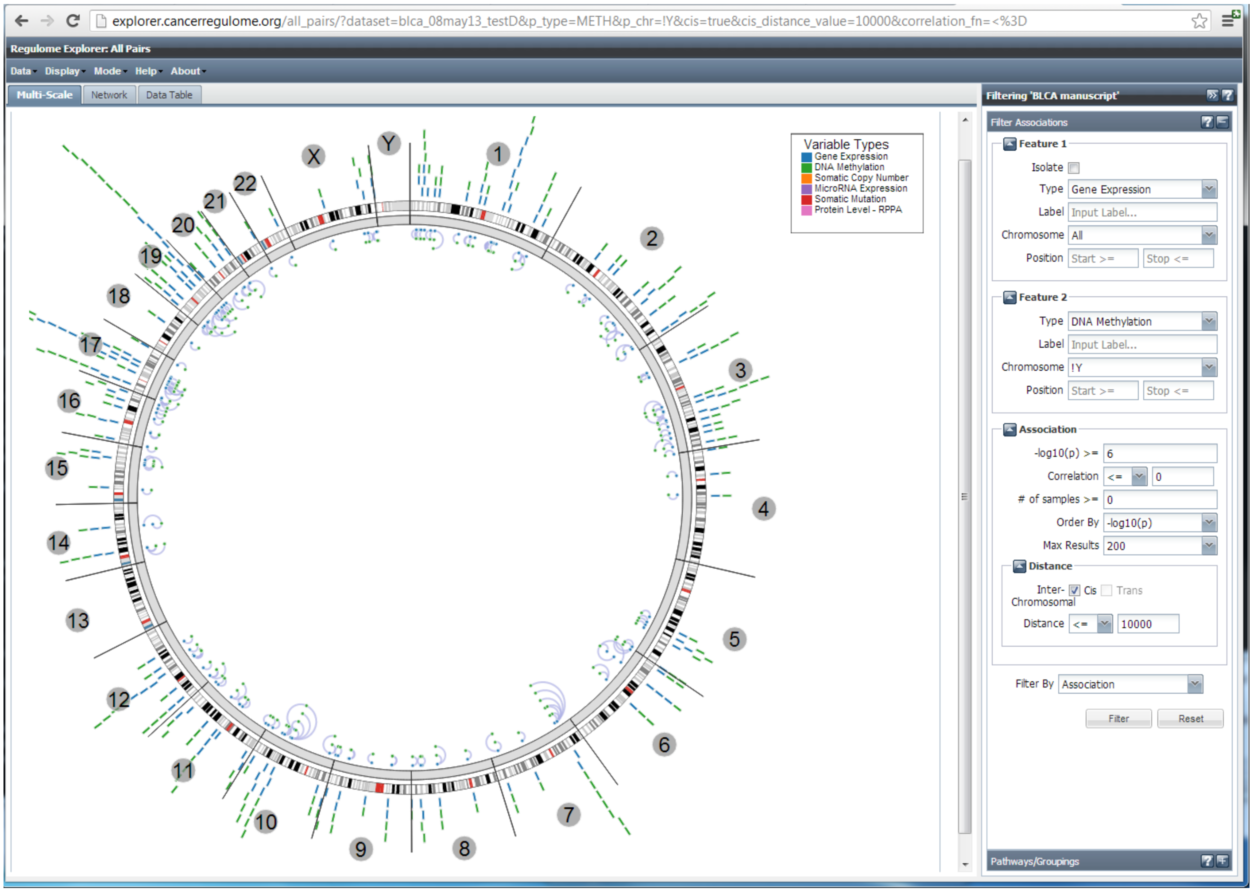
Feature Matrix Construction

A feature matrix was constructed using all available clinical, sample, and molecular data for 131 unique patient/tumor samples. The clinical information includes features such as age, stage, smoking history; while the sample information includes features derived from molecular data such as single-platform cluster assignments, and mutation rates. The molecular data includes mRNA and microRNA expression levels (Illumina HiSeq data), protein levels (RPPA data), copy number alterations (derived from segmented Affymetrix SNP data as well as GISTIC regions of interest and arm-level values), DNA methylation levels (Illumina Infinium Methylation 450k array), and germline and somatic mutations. For each mutated gene, several binary mutation features indicating the presence or absence of a mutation in each sample were generated, depending on the type and position of the mutations. Mutation types considered were synonymous, missense, nonsense and frameshift. Protein domains (InterPro) including any of these mutation types were annotated as such, with nonsense and frameshift annotations being propagated to all subsequent protein domains.

Pairwise Statistical Significance

The statistical significance of each pairwise association is assessed using rank-ordered data and a statistical test appropriate to each data type pair, e.g. Fisher's test (categorical-categorical), F-statistic (continuous-continuous) and ANOVA (continuous-categorical).

Figure S8.3 high mRNA expression *vs* copy-number associations



explorer.cancerregulome.org/all_pairs/?dataset=blca_08may13_testD&p_type=METH&p_chr=!Y&cis=true&cis_distance_value=10000&correlation_fn=<%3D

Regulome Explorer: All Pairs

Filtering 'BLCA manuscript'

Filter Associations

Feature 1

Isolate ☐

Type: Gene Expression

Label: Input Label...

Chromosome: All

Position: Start >= Stop <=

Feature 2

Type: DNA Methylation

Label: Input Label...

Chromosome: !Y

Position: Start >= Stop <=

Association

-log10(p) >= 6

Correlation <= 0

# of samples >= 0

Order By: -log10(p)

Max Results: 200

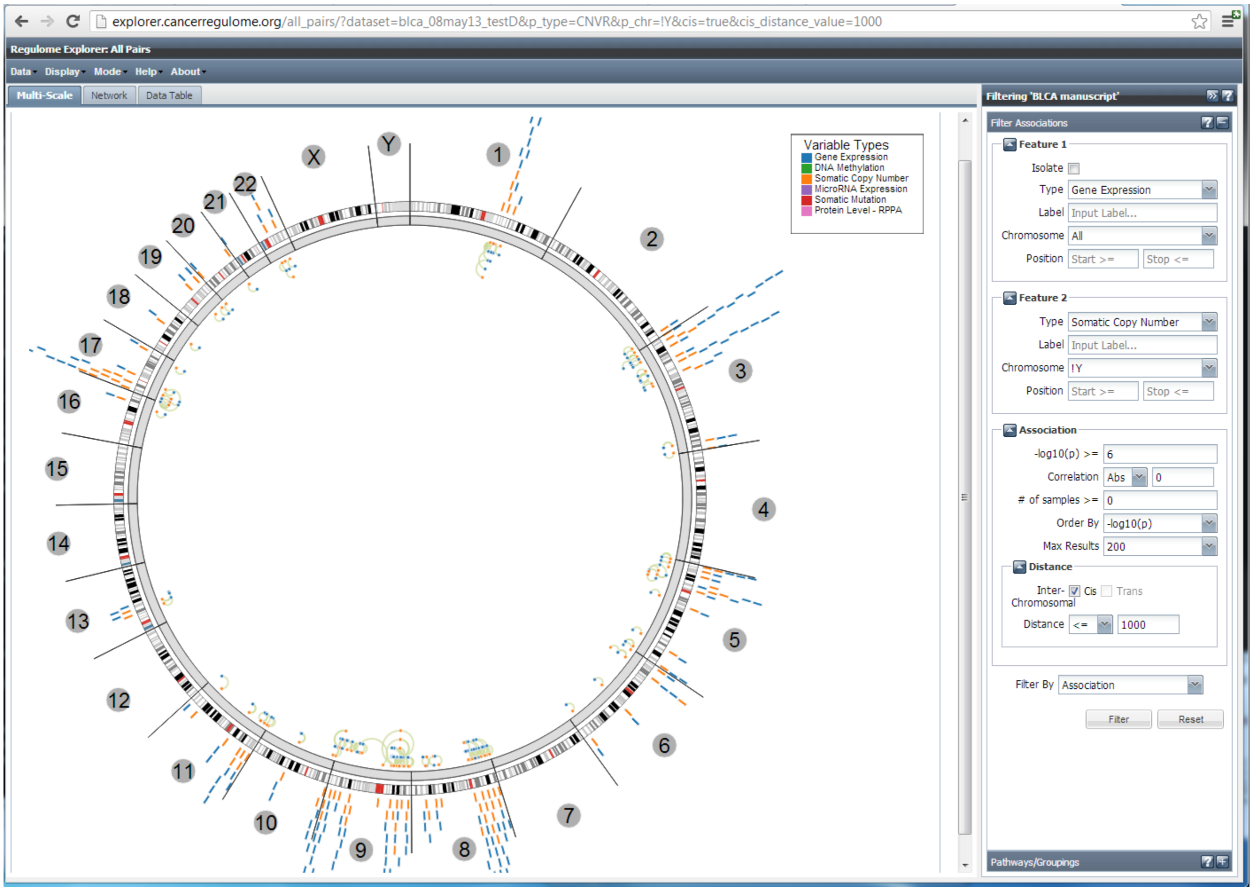
Distance

Inter-Chromosomal ☑ Cis ☐ Trans

Distance <= 10000

Filter By: Association

Figure S8.3.1 high (negative) expression vs methylation associations.



explorer.cancerregulome.org/all_pairs/?dataset=blca_08may13_testD&p_type=CNVR&p_chr=!Y&cis=true&cis_distance_value=1000

Regulome Explorer: All Pairs

Variable Types
🟦 Gene Expression
🟩 DNA Methylation
🟧 Somatic Copy Number
🟪 MicroRNA Expression
🟥 Somatic Mutation
🟪 Protein Level - RPPA

Filtering 'BLCA manuscript'

Filter Associations

Feature 1
Isolate □
Type: Gene Expression
Label: Input Label...
Chromosome: All
Position: Start >= | Stop <=

Feature 2
Type: Somatic Copy Number
Label: Input Label...
Chromosome: !Y
Position: Start >= | Stop <=

Association
-log10(p) >= 6
Correlation: Abs 0
# of samples >= 0
Order By: -log10(p)
Max Results: 200

Distance
Inter-Chromosomal ☑ Cis □ Trans
Distance: <= 1000

Filter By: Association

Team Leader: ***Xiaoping Su*** *XSu1@mdanderson.org* *Team Members: Chandra Pedamallu , Raju Kucherlapati, Semin Lee and Michael Parfenov*

S9.1 Viruses and their integration site detection in TCGA Bladder samples

Text S9.1 Viruses and their integration site detection in TCGA Bladder samples

We detected virus transcripts in 8 out of 122 tumor samples as follows: three tumors with unequivocal CMV, all harboring transcripts encoding RL5A, RNA2.7, RL6, RL8A, RL9A, RNA1.2, UL5, UL22A; one tumor with BK virus; four tumors with HPV6, 16, 45, 56 respectively. None of the two tumors with CMV has evidence of CMV integration into the host genome. In one tumor with BK, BKPyVgp1_agnoprotein and BKPyVgp5_largeTantigen were integrated into GRB14 that is known to interact with a number of receptor tyrosine kinases and signaling molecules. In the tumor with HPV45, HPV45-E1, E6, and E7 transcripts were integrated into *DEC1*, a putative tumor-suppressor gene. In the tumor with HPV56, HPV56-E6, and E7 transcripts were integrated into *NOTCH1,* a key member of the Notch signaling pathway that plays a central role in virus–mediated host cellular network perturbations. In the same tumor, HPV56-E1, E6, and E7 transcripts were also integrated into *SEC16A*, a gene involved in the assembly of endoplasmic reticulum exit sites and endoplasmic reticulum-to-Golgi protein transport. In the tumor with HPV16, HPV16-E6, E4, L1 transcripts were integrated into BCL2L1. In the tumor with HPV6, there is no evidence of HPV6 integration into the host genome. Of note, two tumors with HPV6 or HPV56 have >50% squamous tissue, and the tumor with HPV45 might be cervical in origin.

METHODS

Mapping/Alignment: We first performed quality checks on sequencing data using the HTSeq package (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html). The raw paired-end (PE) reads in FASTQ format were then aligned to the human reference genome, GRCh37/hg19, using MOSAIK (Hiller, et al., 2008) alignment software. MOSAIK works with PE reads, and uses both a hashing scheme and the Smith-Waterman algorithm to produce gapped optimal alignments and to map exon junction-spanning reads with a local alignment option for RNA-seq data. VirusSeq (Chen et al., 2013) was used to detect both viruses and theirs integration sites in Bladder RNA-seq data.

Virus detection from RNA-Seq: VirusSeq started by computationally subtracting human sequences, followed by generating a set of nonhuman sequences (e.g., viruses) on RNA-Seq. Once raw PE reads from RNA-Seq were aligned to the human genome reference, any read with more than a half read length mapped to the human reference genome was removed along with its paired mate in this subtraction step. Thus, a set of nonhuman sequences was generated after human sequence subtraction. In the second step, VirusSeq determined whether the nonhuman sequences matched with any known viral sequences by searching a comprehensive database that includes all known viral sequences (Genome Information Broker for Viruses; GIB-V, http://gib-v.genes.nig.ac.jp/), and quantified virus representation by a measure of the virus genome coverage (or overall count of mapped reads) to determine the existence of viruses in human samples with an empirical cutoff. Any virus with an overall count of mapped reads below the cutoff was treated as nonexistent. We used 1000 as the cutoff for the overall count of mapped reads within a virus genome.

Identification of virus integration sites: The genomes of viruses detected in the previous steps were concatenated into a single genome named chrVirus with related annotation of each virus in refFlat format. A new hybrid reference genome named hg19Virus was built by combining hg19 and chrVirus. All PE reads without computational subtraction were mapped to this reference (hg19Virus). If the PE reads were uniquely mapped with one end to hg19 and the other to chrVirus, the read pair was reported as a discordant read pair. All discordant reads were then annotated by using the genes and viruses defined in the curated refFlat file. VirusSeq then clustered the remaining discordant read pairs that support the same integration (fusion) event (e.g., HPV56-NOTCH1), and selected them as fusion candidates. The cluster size was constrained by the library insert size (fragment length) distribution after excluding the sizes of introns if mapped reads were located across adjacent exons. VirusSeq implemented a dynamic clustering procedure to accurately determine the exact fusion junction between a human gene and a virus. In order to remove outliers within a cluster, VirusSeq implemented the robust “extreme studentized deviate” multiple-outlier procedure (Rosner, 1983). Meanwhile, an *in-silico* sequence was generated using the consensus of reads within discordant read clusters for each fusion candidate to help the PCR primer design, which facilitates quick PCR validation.

Table S9.1 Viral Integration

SampleID	Discordant_reads	Virus	Viral_Transcript	HostGenes	Integrated_Site
A3I6-01A	5	HPV16	HpV16gp5_E4	BCL2L1	exon2
A3I6-01A	7	HPV16	HpV16gp8_L1	BCL2L1	intron2
A3I6-01A	3	HPV16	HpV18gp1_E6	BCL2L1	intron2
A20V-01A	60	HPV45	HpV45gp1_E6	DEC1	intron1
A20V-01A	351	HPV45	HpV45gp2_E7	DEC1	intron1
A20V-01A	62	HPV45	HpV45gp3_E1	DEC1	intron1
A3B4-01A	82	HPV56	HpV56gp1_E6	NOTCH1	exon27
A3B4-01A	11	HPV56	HpV56gp2_E7	NOTCH1	exon27
A3B4-01A	10	HPV56	HpV56gp1_E6	SEC16A	exon2
A3B4-01A	227	HPV56	HpV56gp2_E7	SEC16A	exon3
A3B4-01A	19	HPV56	HpV56gp3_E1	SEC16A	exon3
A3B4-01A	36	HPV56	HpV56gp3_E1	SEC16A	intron1

A3IT-01A	72	BK	BKPyVgp1_agnoprotein	GRB14	3Prime
A3IT-01A	48	BK	BKPyVgp5_largeTantigen	GBR14	3Prime

Table S9.2 Pathology Review

SampleID	Virus	Path_Review
A1AF-01A	CMV	Clean
A3JZ-01A	CMV	Clean
A3SN-01A	CMV	Clean
A3IT-01A	BK	Clean
A3I6-01A	HPV16	Clean
A20V-01A	HPV45	Questionable squamous, HPV associated (no feedback from TSS), per path report --> bladder tumor but also extensive involvement of ectocervix, endocervix, endometrium and myometrium. Seriously consider cervical primary
A3B4-01A	HPV56	>50 squamous in 3 reviews, HPV associated
A3N6-01A	HPV6	>50 squamous in 3 reviews, HPV associated

References

1. Hiller, L.W. et al. Whole-genome sequencing and variant discovery in C. elegans. *Nat Methods* **5**, 183-8 (2008).

2. Chen, Y. et al. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* **29**, 266-7 (2013).

3. Rosner B. Percentage points for a generalized ESD many outlier procedure. *Technometrics* 25(2):165-172 (1983).

*Team Lead: **Rehan Akbani** RAkbani@mdanderson.org Team Members: Gordon Mills, and John Weinstein*

S10.1: RPPA analyses:

Text S10.1: RPPA analyses

RPPA experiments and data processing

Protein was extracted using RPPA lysis buffer (1% Triton X-100, 50 mmol/L Hepes (pH 7.4), 150 mmol/L NaCl, 1.5 mmol/L MgCl2, 1 mmol/L EGTA, 100 mmol/L NaF, 10 mmol/L NaPPi, 10% glycerol, 1 mmol/L phenylmethylsulfonyl fluoride, 1 mmol/L Na3VO4, and aprotinin 10 ug/mL) from human tumors and RPPA was performed as described previously [1-5]. Lysis buffer was used to lyse frozen tumors by Precellys homogenization. Tumor lysates were adjusted to 1 µg/µL concentration as assessed by bicinchoninic acid assay (BCA) and boiled with 1% SDS. Tumor lysates were manually serial diluted in two-fold of 5 dilutions with lysis buffer. An Aushon Biosystems 2470 arrayer (Burlington, MA) printed 1,056 samples on nitrocellulose-coated slides (Grace Bio-Labs). Slides were probed with 179 validated primary antibodies followed by corresponding secondary antibodies (Goat anti-Rabbit IgG, Goat anti-Mouse IgG or Rabbit anti-Goat IgG). Signal was captured using a DakoCytomation-catalyzed system and DAB colorimetric reaction. Slides were scanned in CanoScan 9000F. Spot intensities were analyzed and quantified using Microvigene software (VigeneTech Inc., Carlisle, MA), to generate spot signal intensities (Level 1 data). The software SuperCurveGUI[3,5], available at http://bioinformatics.mdanderson.org/Software/supercurve/, was used to estimate the EC50 values of the proteins in each dilution series (in log2 scale). Briefly, a fitted curve ("supercurve") was plotted with the signal intensities on the Y-axis and the relative log2 concentration of each protein on the X-axis using the non-parametric, monotone increasing B-spline model [1]. During the process, the raw spot intensity data were adjusted to correct spatial bias before model fitting. A QC metric [5] was returned for each slide to help determine the quality of the slide: if the score is less than 0.8 on a 0-1 scale, the slide was dropped. In most cases, the staining was repeated to obtain a high quality score. If more than one slide was stained for an antibody, the slide with the highest QC score was used for analysis (Level 2 data). Protein measurements were corrected for loading as described [3,5,6] using median centering across antibodies (level 3 data). In total, 179 antibodies and 127 samples were used. Final selection of antibodies was also driven by the availability of high quality antibodies that consistently pass a strict validation process as previously described [7]. These antibodies are assessed for specificity, quantification and sensitivity (dynamic range) in their application for protein extracts from cultured cells or tumor tissue. Antibodies are labeled as validated and use with caution based on degree of validation by criteria previously described [7].

Two RPPA arrays were quantitated and processed (including normalization and load controlling) as described previously, using MicroVigene (VigeneTech, Inc., Carlisle, MA) and the R package SuperCurve (version-1.3), available at http://bioinformatics.mdanderson.org/OOMPA [1,3]. Raw data (level 1), SuperCurve nonparameteric model fitting on a single array (level 2), and loading corrected data (level 3) were deposited at the DCC.

References:

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

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

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

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

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

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

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

Data normalization

We performed median centering across all the antibodies for each sample to correct for sample loading differences. Those differences arise because protein concentrations are not uniformly distributed per unit volume. That may be due to several factors, such as differences in protein concentrations of large and small cells, differences in the amount of proteins per cell, or heterogeneity of the cells comprising the samples. By observing the expression levels across many different proteins in a sample, we can estimate differences in the total amount of protein in that sample vs. other samples. Subtracting the median protein expression level forces the median value to become zero, allowing us to compare protein expressions across samples. All of the 127 samples were processed in a single RPPA batch.

Reference:

1) Neeley ES, Kornblau SM, Coombes KR, Baggerly KA (2009). Variable Slope Normalization of Reverse Phase Protein Arrays. Bioinformatics, 25(11), 1384-1389.

Unsupervised clustering

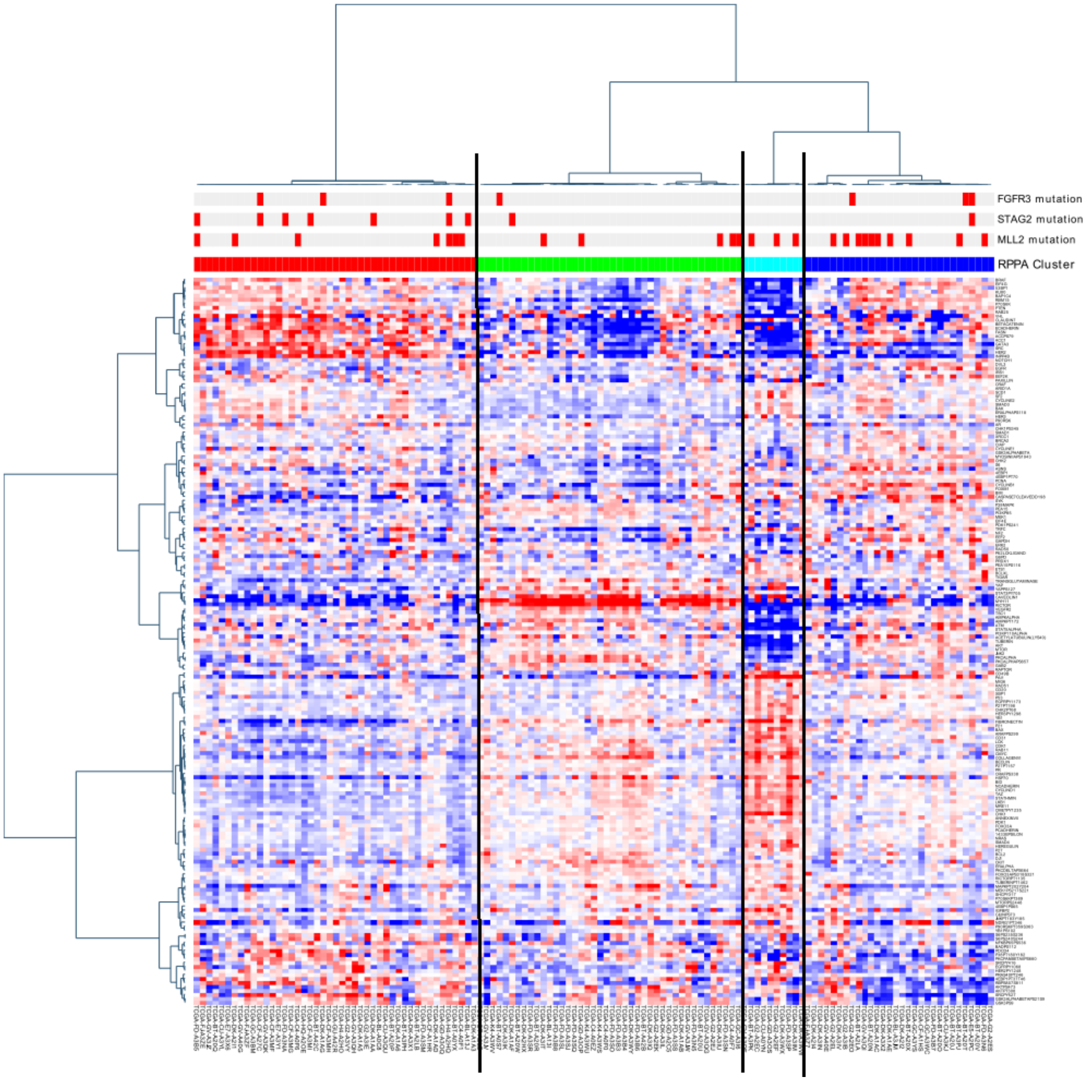
We performed unsupervised clustering on the protein expression data. Pearson correlation was used as the distance metric and Ward was used as the linkage algorithm in an bootstrapped ensemble clustering algorithm that merges the output of hierarchical and k-means clustering. The method clustered the samples by bootstrapping and counted how frequently two samples are in the same cluster. The resampling analysis identified four robust sample clusters. The four clusters and their protein expression patterns are shown in Supplemental Figure SXX.

The red cluster has elevated levels of Her2, E-cadherin, claudin 7, SRC and Gata3, indicating a hormonally responsive subtype. It also has several FGFR3 and STAG2 mutants. There is little difference between the four clusters in MLL2 mutants. The green cluster is enriched in caveolin, MYH11 and rictor, indicative of a previously identified “reactive” RPPA signature found in breast cancer [1]. It is depleted in FGFR3 and STAG2 mutations. The cyan cluster shows very low Her2, SRC, and Gata 3 levels, indicating low hormonal responsiveness. It has low caveolin and MYH11, but high collagen VI and fibronectin that is indicative of a second reactive signature. It has low levels of AMPK-alpha and phospho-AMPK and no mutations in FGFR3 and STAG2. The blue cluster shows low hormonal and reactive signatures, as well as low levels of AKT. It has some FGFR3 mutants.

References

1) The Cancer Genome Atlas (TCGA) Research Network. Comprehensive molecular characterization of human breast cancers. Nature. 2012 Oct 4;490(7418):61-70

Figure S10.1: Unsupervised hierarchical clustering of 127 samples and 179 antibodies, showing 4 RPPA clusters.



The red cluster has elevated levels of Her2, E-cadherin, claudin 7, SRC and Gata3, indicating a hormonally responsive subtype. It also has several FGFR3 and STAG2 mutants. There is little difference between the four clusters in MLL2 mutants. The green cluster is enriched in caveolin, MYH11 and rictor, indicative of a previously identified "reactive" RPPA signature found in breast cancer [1]. It is depleted in FGFR3 and STAG2 mutations. The cyan cluster shows very low Her2, SRC, and Gata 3 levels, indicating low hormonal responsiveness. It has low caveolin and MYH11, but high collagen VI and fibronectin that is indicative of a second reactive signature. It has low levels of AMPK-alpha and phospho-AMPK and no mutations in FGFR3 and STAG2. The blue cluster shows low hormonal and reactive signatures, as well as low levels of AKT. It has some FGFR3 mutants

References

1) The Cancer Genome Atlas (TCGA) Research Network. Comprehensive molecular characterization of human breast cancers. Nature. 2012 Oct 4;490(7418):61-70

Team Lead: ***Michael Ryan*** *mryan@insilico.us.com* *Team Member: John Weinstein*

S11.1: Splicing Analysis

Text S11.1: Methods and Materials: splicing Analysis

SpliceSeq[1] was used to analyze the BLCA RNASeq data for transcript splicing variation. SpliceSeq aligned reads to splice graphs representing all protein coding isoforms of human genes in Ensembl. Read totals and normalized read totals were calculated for each exon and splice. For all potential splice events, a percent spliced in (PSI) value was also calculated. The PSI is the ratio of reads indicating the inclusion path vs. the reads indicating the exclusion path. If, for example, a skip of exon 4 is evaluated, then reads of exon 4 and junction reads that cross exon 4 would be include reads and reads of the junction that splices out exon 4 would be exclude reads. A PSI of 20% would indicate that the exon is included in approximately 20% of the transcripts in a sample and spliced out in 80% of transcripts. Differential analysis of groups of samples was performed in SpliceSeq by calculating t-test p-values on the individual PSIs from each group and then calculating a Benjamini–Hochberg FDR q_value. Splice Events with a delta PSI of > 20% and a q_value of < .01 are reported. Results were further filtered to include only genes with an average expression level > 2 RPKM in both groups and which had PSI values for > 90% of group members. For further details on SpliceSeq methods, see: http://bioinformatics.mdanderson.org/main/SpliceSeqV2:Methods. The type of splice events detected include exon skip (ES), retained intron (RI), alternate donor (AD), alternate acceptor (AA), mutually exclusive exon (ME), alternate promoter (AP), and alternate terminator (AT). HG19 exon coordinates are provided for exons involved in each splice event.

Several sub-groups of BLCA samples were evaluated for differential splicing patterns. No substantive differences were observed for comparisons of tumor disease stage, patient smoking status, gender, age, or lymph node status. The group comparison that did show many significant splicing events was tumor vs. adjacent normal. Splicing pattern analysis is sensitive to tissue type differences so only tumor samples with an abs_purity score of > 90% were included in the tumor / normal analysis. Splicing events identified in this analysis have been included as a separate supplemental spreadsheet, BLCA_SpiceSeq_Normal_V_PureTumor.xlsx.

Discussion

A total of 116 splice events on 111 genes were identified as differentially spliced in BLCA tumor samples as compared to adjacent normal tissue. Fifteen (15) of the differentially spliced genes have been associated with tumor suppression or tumor growth and metastasis. Genes with structural function (adhesion, cellular matrix, actin related) were prevalent in the splice event list totaling 16 of the 111 genes.

Table S11.1: Cancer related genes from the list of alternatively spliced BLCA genes compared to adjacent normal.

Gene Symbol	Splice Event Type	Exons	dPSI	p-value	q_value	Gene Function
ABI1	ES	5	-0.25	3.80E-05	0.0081	Mediates signal transduction from Ras to Rac. May play a role in the progression of several malignancies including melanoma, colon cancer and breast cancer,
CDKN2C	AP	3.1	-0.34	5.40E-07	7.00E-04	Tumor supressor. Cyclin-dependent kinase inhibitor that regulates growth.
CXCL12	AT	5.2	0.46	3.89E-06	0.0023	Plays a role in tumor growth and metastasis.
FBLN1	AT	22	-0.22	3.77E-05	0.00815	Tumor suppressor. Implicated in cellular transformation and tumor invasion.
FN1	ES	40.2	0.21	3.36E-06	0.0023	Fibronectin is involved in cell adhesion and migration processes including metastasis
GNE	RI	13.2	-0.47	1.67E-06	0.00145	Implicated in cell adhesion, tumorigenicity and metastatic behavior of malignant cells.
MLTK	AT	14	-0.32	2.13E-07	5.00E-04	Pro-apoptotic. Isoform 1 may role in cancer development.
MORF4L2	AP	1	-0.23	1.41E-06	0.0013	Component of histone acetyltransferase. Activates transcription of oncogenes and tumor suppressors.
SEPT9	AP	1	0.20	3.65E-05	0.008	Cell cycle control. Candidate ovarian tumor suppressor.
SPAG9	ES	30	-0.52	2.56E-05	0.00645	Mediates c-Jun-terminal kinase signaling. May play a role in tumor growth and development.
TACC2	ES	12	0.36	5.58E-07	7.00E-04	Centrosome- and microtubule-interacting proteins that are implicated in tumorigenesis.
TNFSF12	AT	13	0.31	6.26E-07	7.50E-04	TNF ligand. Cytokine that can induce apoptosis.
TPM1	AP	1	-0.61	2.34E-07	4.50E-04	Tumor supressor that promotes structural stability and controls growth.
TSC2	ES	27:28.1	0.24	2.22E-05	0.00615	Implicated as a tumor suppressor. Negatively regulates mTORC1 signaling.
URGCP	AP	3	-0.22	1.52E-06	0.00135	Cell cycle. Regulates cyclin D1. Implicated in hepatocellular carcinoma and gastric cancer.

One particularly interesting member of this list is the alternate termination event on MLTK. MLTK is reduced in expression in the tumor cells and the alternate termination event indicates a reduction in tumor samples of MLTK-β and a corresponding increase in the relative concentration of MLTK-α. MLTK is a mixed-lineage kinase that plays a role in activating MAPK pathways which in turn regulate cell proliferation, differentiation, stress response, and cell death. The MLTK-α isoform has been shown to induce proliferation and malignant cell transformation. MLTK is also implicated in actin organization and MLTK-α has been shown to disrupt actin stress fibers and   induce dramatic morphological changes.

Table 11.1.2: Structural protein genes with splice events in BLCA tumor vs. adjacent normal comparison.

Gene Symbol	Splice Event Type	Exons	dPSI	p-value	q_value	Gene Function
ABI1	ES	5	-0.25	3.80E-05	0.0081	Abelson-interactor adaptor protein. Regulate actin polymerization and cytoskeletal remodeling.
ARHGAP6	AP	1	-0.26	1.84E-05	0.0056	GTPase activator. Regulates signaling with actin cytoskeleton. Role in cell morphology.
CALD1	ES	8.3:9	-0.70	1.16E-09	1.00E-04	calmodulin- and actin-binding protein that regulates muscle contraction
FBLN1	AT	22	-0.22	3.77E-05	0.00815	Role in cell adhesion and migration along protein fibers within the extracellular matrix
FBLN2	ES	11	-0.56	1.89E-07	5.00E-04	Extracellular matrix protein. Binds various extracellular ligands and calcium
FBLN5	AP	2	-0.26	4.19E-06	0.00235	Extracellular matrix protein. Promotes adhesion. Possible role in vascular development and remodeling.
FLNA	ES	30	0.48	5.70E-09	1.00E-04	Aactin-binding protein that crosslinks actin filaments and links actin filaments to membrane glycoproteins.
FN1	ES	40.2	0.21	3.36E-06	0.0023	Fibronectin, a glycoprotein involved in cell adhesion and migration.
GNE	RI	13.2	-0.47	1.67E-06	0.00145	Regulates the biosynthesis of N-acetylneuraminic acid (NeuAc) involved in cell adhesion.
INF2	ES	22	0.20	9.97E-06	0.00415	Formin proteins - may function in polymerization and depolymerization of actin filaments.
MLTK	AT	14	-0.32	2.13E-07	5.00E-04	Component of a protein kinase signal transduction cascade. Can induce disruption of actin stress fibers.
PPP1R12A	ES	26	-0.58	5.15E-06	0.0027	myosin-binding subunit of myosin phosphatase. Regulates the interaction of actin and myosin.
SVIL	ES	21	-0.65	8.53E-08	2.50E-04	Tightly associated with both actin filaments and plasma membranes.
TAGLN	RI	1.2	0.20	5.28E-05	0.0095	Actin cross-linking/gelling protein found in fibroblasts and smooth muscle.
TPM1	AP	1	-0.61	2.34E-07	4.50E-04	Tropomyosin. Actin-binding proteins involved in contractile systems and cytoskeleton.
VCL	ES	19	-0.58	9.22E-07	9.50E-04	Cytoskeletal protein associated with cell-cell and cell-matrix junctions. Anchors F-actin to the membrane.

Adjacent normal tissue is not a perfect control to use in finding tumor specific splicing changes particularly in epithelial cancers because it is likely to be a heterogeneous mix of tissue and may display field effects. The identified differential splicing is likely to contain both tissue specific splicing differences and tumor specific splicing differences. Some confidence that many of these splicing events are tumor specific is provided by the prevalence of literature association to cancer. The structural nature of many of the identified splicing events may relate to the focus of this study on muscle invasive bladder cancer requiring changes in cell adhesion and mobility.

PKM

An important alternative splicing event in observed in the development of tumors is the shift from the PKM1 to the PKM2 isoform of pyruvate kinase M. The PKM2 isoform induces aerobic glycolysis (Warburg effect) which produces less energy than oxidative phosphorylation but provides advantages for proliferating cells. The SpliceSeq isoform analysis of all BLCA RNASeq samples showed an average expression ratio of 3% PKM1 to 97% PKM2. In the samples with matched adjacent normal, the tumor samples showed a nearly twofold increase of PKM expression but all of the increased expression was the PKM2 isoform while PKM1 expression remained constant or decreased. PKM2 is expressed in normal tissue at varying levels depending on tissue type. The BLCA adjacent normal tissue showed an average ratio of 24% PKM1 to 76% PMK2. In normal vs. pure tumor group analysis, the PKM mutually exclusive exon event showed a dPSI of 18%, p_value of < .003 and q_value of < .03.

CD44

The transmembrane CD44 protein has a large set of variable exons in the center of its transcript that are expressed in complex combinations in normal tissue. CD44 has many roles but a primary role is cell adhesion and altered splicing of CD44 has been strongly associated with metastasis. For example, inducing expression of the variable exons v6-7 in a parental pancreatic carcinoma cell line was sufficient to enable the cells to become metastatic. In the BLCA samples, 5 of 16 samples with matched adjacent normal showed very significant increase in expression of variable exons v2-6 (SI 6.2) and increase to a lesser degree of v1 (SI 3.0). Two other samples showed a modest increase in exons v7-9. The remaining samples showed similar expression patterns for the variable exons in tumor compared to adjacent normal.

Figure 11.1: SpliceSeq compartive splicing analysis of CD44 in TCGA_BL_A13J showing dramatically increased expression of variable exons v2-9 in tumor vs. adjacent normal tissue.

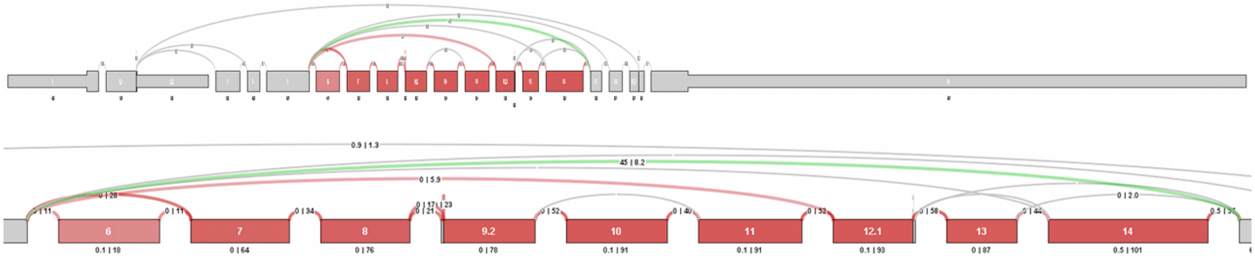
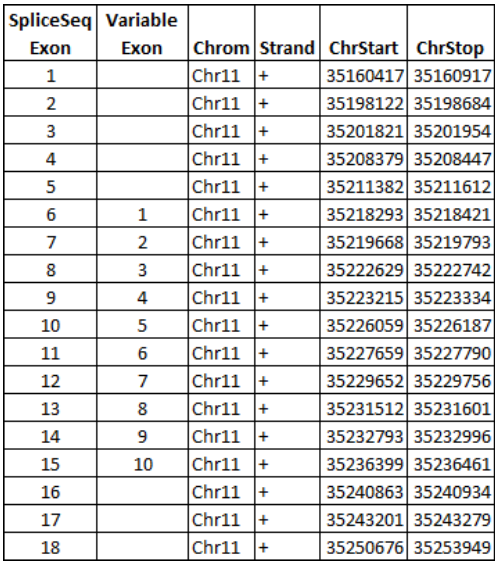


Table 11.1.3: CD44 Definition of variable exons.



SpliceSeq Exon	Variable Exon	Chrom	Strand	ChrStart	ChrStop
1		Chr11	+	35160417	35160917
2		Chr11	+	35198122	35198684
3		Chr11	+	35201821	35201954
4		Chr11	+	35208379	35208447
5		Chr11	+	35211382	35211612
6	1	Chr11	+	35218293	35218421
7	2	Chr11	+	35219668	35219793
8	3	Chr11	+	35222629	35222742
9	4	Chr11	+	35223215	35223334
10	5	Chr11	+	35226059	35226187
11	6	Chr11	+	35227659	35227790
12	7	Chr11	+	35229652	35229756
13	8	Chr11	+	35231512	35231601
14	9	Chr11	+	35232793	35232996
15	10	Chr11	+	35236399	35236461
16		Chr11	+	35240863	35240934
17		Chr11	+	35243201	35243279
18		Chr11	+	35250676	35253949

References

1. Ryan MC, Cleland J, Kim R, Wong WC, Weinstein JN. SpliceSeq: A Resource for Analysis and Visualization of RNA-Seq Data on Alternative Splicing and Its Functional Impacts. *Bioinformatics*, 10.1093, 2012.

2. Ryan MC, Zeeberg BR, Caplen NJ, Cleland JA, Kahn AB, Liu H, and Weinstein JN. SpliceCenter: a suite of web-based bioinformatic applications for evaluating the impact of alternative splicing on RT-PCR, RNAi, microarray, and peptide-based studies. *BMC Bioinformatics*, 9:313, 2008.

3. David CJ, Manley JL. Alternative pre-mRNA splicing regulaiton in cancer: pathways and programs unhinged. *Genes Dev.* 24: 2343-2364, 2010.

4. Christofk HR, Vander Heiden MG, Harris MH, Ramanathan A, Gerszten RE, Wei R, Fleming MD, Schreiber SL, Cantley LC. The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* 452:230–233, 2008.

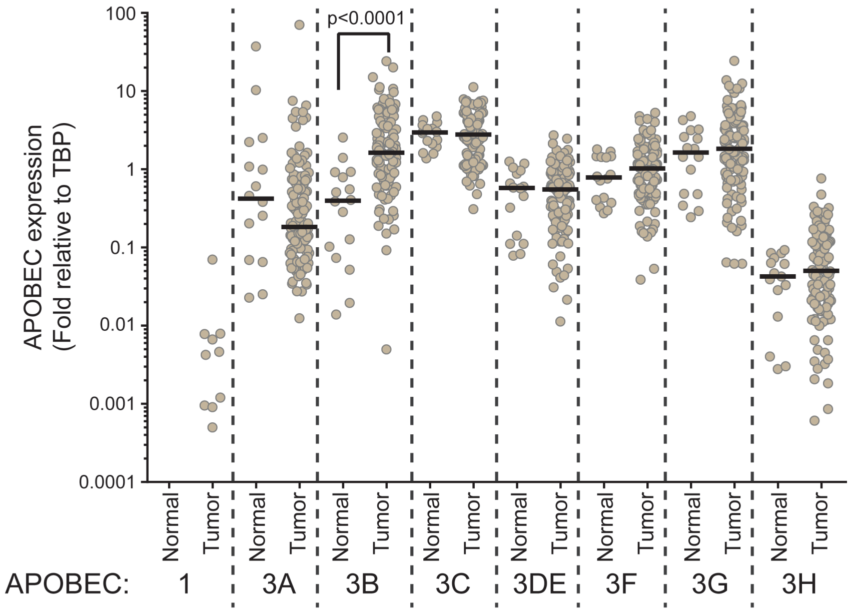
5. Gunthert U, Hofmann M, Rudy W, Reber S, Zoller M, HaussmannI, Matzku S, Wenzel A, Ponta H, Herrlich P. A new variant of glycoprotein CD44 confers metastatic potential to rat carcinoma cells. *Cell* 65: 13–24, 1991.

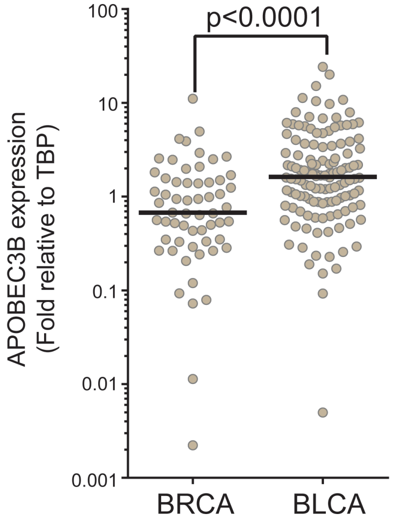
6. Cho Y, Bode A, Mizuno H, Choi B, Choi H, and Dong Z. A Novel Role for Mixed-Lineage Kinase-Like Mitogen-Activated Protein Triple Kinase α in Neoplastic Cell Transformation and Tumor Development. *Cancer Research* 64; 3855, 2004.

Team Leader: Dmitry A. Gordenin ***gordenin@niehs.nih.gov*** *Team Members: Steven A. Roberts, Leszek J. Klimczak, David Fargo*

S12: Analysis of APOBEC Mutagenesis

Figure S12.1 Expression analysis of APOBEC enzymes

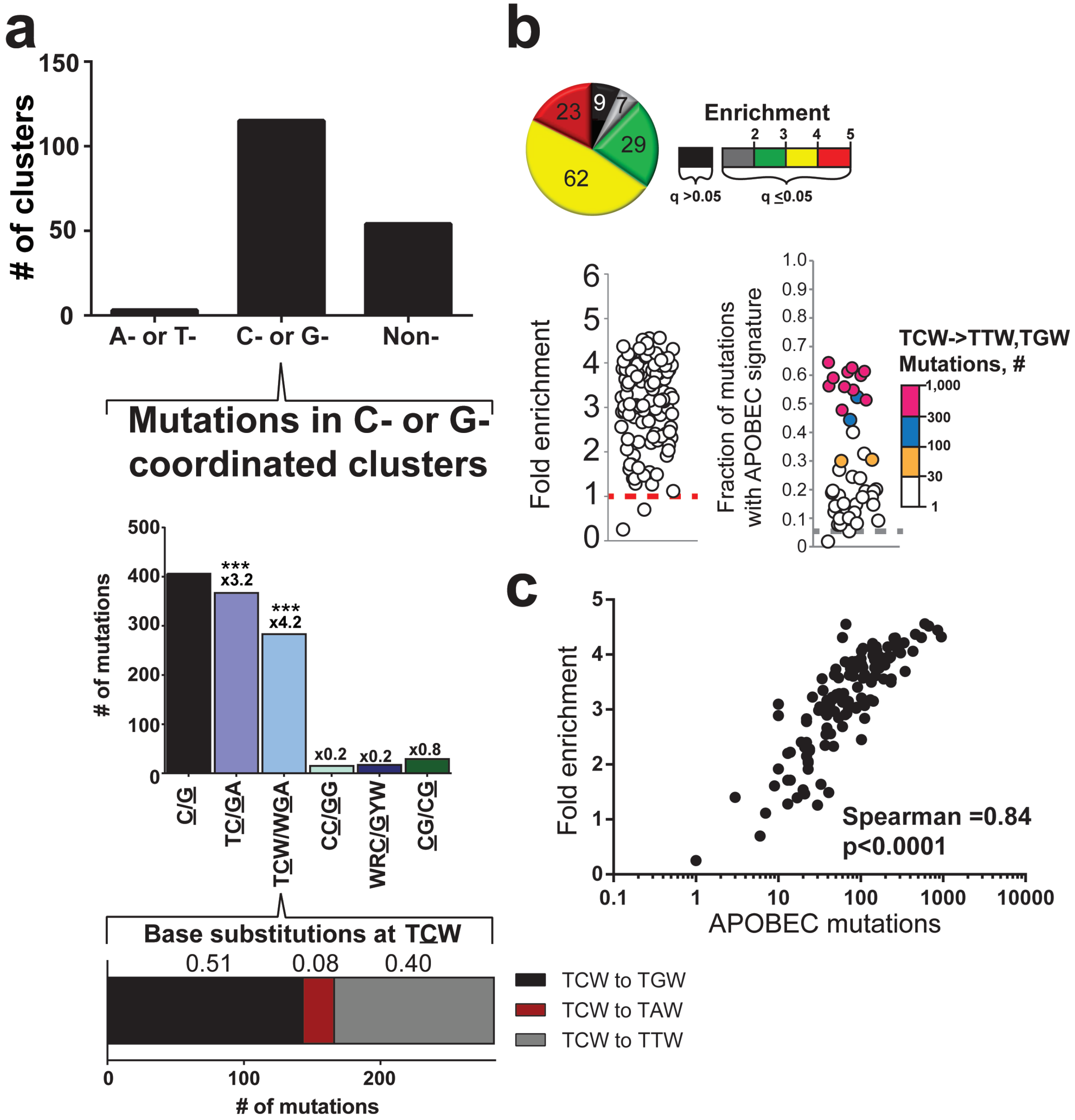




Top panel. mRNA expression levels of APOBEC3B in 130 BLCA samples are mostly higher than the median level of APOBEC3B expression in 16 matched normal samples available in TCGA. Other APOBECs in tumor samples are expressed at the levels comparable with normal tissue samples. APOBEC expression in tumors was compared to expression in matched normal by Mann-Whitney.

Bottom panel. Median APOBEC3B mRNA expression levels in BLCA TCGA samples is higher than in BRCA samples analyzed in[1] based on comparison by Mann-Whitney.

Figure S12.2 APOBEC mutagenesis pattern in bladder cancer exomes.



Note: Here and below the term APOBEC without the gene-designating suffix is used to indicate a subclass of APOBECs with TC (mutated nucleotide underlined) specificity. This specificity is different from Activation-Induced Cytosine Deaminase (AID) (WRC, R – stands for either A or G) or from APOBEC3G (CC)

a. ***Top panel.*** APOBEC mutagenesis pattern in mutation clusters. APOBEC cytidine deaminases display strong preference to ssDNA over dsDNA, therefore they tend to cause clusters of mutations in positions of ssDNA accidentally formed at double-strand breaks (DSBs) or at uncoupled replication forks[2]. Since only one of two DNA strands is present in ssDNA, several cytidine deamination events occur in the same DNA strand and thus mutations within a cluster are strand-coordinated (i.e., mutations within a single cluster occur only in cytosines or only in guanines). A high excess of C- or G-coordinated clusters over A- or T-coordinated clusters agrees with an APOBEC mutagenesis pattern.

Middle panel. Mutation clusters were identified and enrichment of the APOBEC mutation signature was calculated as in[3] assuming that an exome contains approximately 1% of the mutations in a whole-genome. In agreement with APOBEC mutagenesis, mutation events in C- and G-coordinated clusters are highly enriched with the APOBEC mutation motif TCW and are depleted for other known C-containing motifs targeted by mutagenic factors (W is A or T, R is A or G, Y is T or C). (***Bonferroni-corrected q value < 0.0001, as determined by a one-tailed Fisher's exact test comparing the ratio of the number of cytosine mutations at a specified motif and the number of cytosine mutations not in the motif to the analogous ratio for all cytosines within a sample fraction of the genome. Complementary DNA sequences and mutations are included.)

Bottom panel. The major pathway of mutagenesis caused by cytidine deamination occurs via uracil-DNA-glycosylase generation of abasic sites, followed by error-prone trans-lesion synthesis placing either A or C across from the lesion. This results in approximately equal numbers of C→T and C→G changes with very few C→A mutations[2,4]. This pattern is clearly seen in TCW motifs of mutation clusters.

b. High presence of the APOBEC mutagenesis pattern in the vast majority of BLCA samples. 121 out of 130 samples show statistically significant (q≤0.05 after FDR-correction) enrichment with the APOBEC mutagenesis pattern (***pie chart***), with fold enrichment reaching 4.5x over expected for random mutagenesis at cytosines (***left graph***). Fold enrichment is calculated as the frequency mutated cytosines involve C→T and C→G changes at the TCW motif divided by the frequency cytosine bases occur in the DNA sequence TCW. Complementary DNA sequences and mutations are included. Up to 64% of all exome mutations can carry the APOBEC signature with up to 955 APOBEC mutations in a single exome (***right graph***).

c. The value of fold enrichment with the APOBEC mutation pattern highly correlates with the mutation load caused by this type of mutagenesis, which further supports genome-wide mutagenesis by APOBEC (p<0.0001 by non-parametric Spearman Correlation).

References for Supplementary Methods S12.

1. Burns, M.B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494**, 366-70 (2013).
2. Roberts, S.A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Molecular cell* **46**, 424-35 (2012).
3. Roberts, S.A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature genetics* **45**, 970-6 (2013).
4. Chan, K. *et al.* Base damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous environmental agent. *PLoS genetics* **8**, e1003149 (2012).

*Team Leader: **Rehan Akbani** RAkbani@mdanderson.org Team Members: Nianxiang Zhang, Thomas C. Motter, Tod D. Casasent, John N. Weinstein*

S13 Batch Effects

Text S13.1: Batch Effects

We used hierarchical clustering and Principal Components Analysis (PCA) to assess batch effects in the bladder cancer data sets. Four different data sets were analyzed: miRNA sequencing (Illumina HiSeq), DNA methylation (Infinium HM450 microarray), mRNA sequencing (Illumina HiSeq), and SNPs (GW SNP 6). 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). Detailed results and batch effects analysis of other TCGA data sets can be found at: http://bioinformatics.mdanderson.org/tcgabatcheffects

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 four data sets follow.

miRNA (RNA-seq Illumina HiSeq)

Figures 1-3 show clustering and PCA plots for miRNA seq data. miRNAs with zero values were removed and the read counts were $\log_2$-transformed before generating the figures. The figures show a small batch effect by the tissue source site ILSBio. However, the magnitude of batch effects wasn't too great, so we didn't think that it warranted batch effects correction for the type of analyses done in this paper. The trade off with batch effects correction algorithms is the possibility of losing important biological variation in the data, along with the technical variation.

Figure S13.1. Hierarchical clustering for miRNA expression from miRNA-seq data

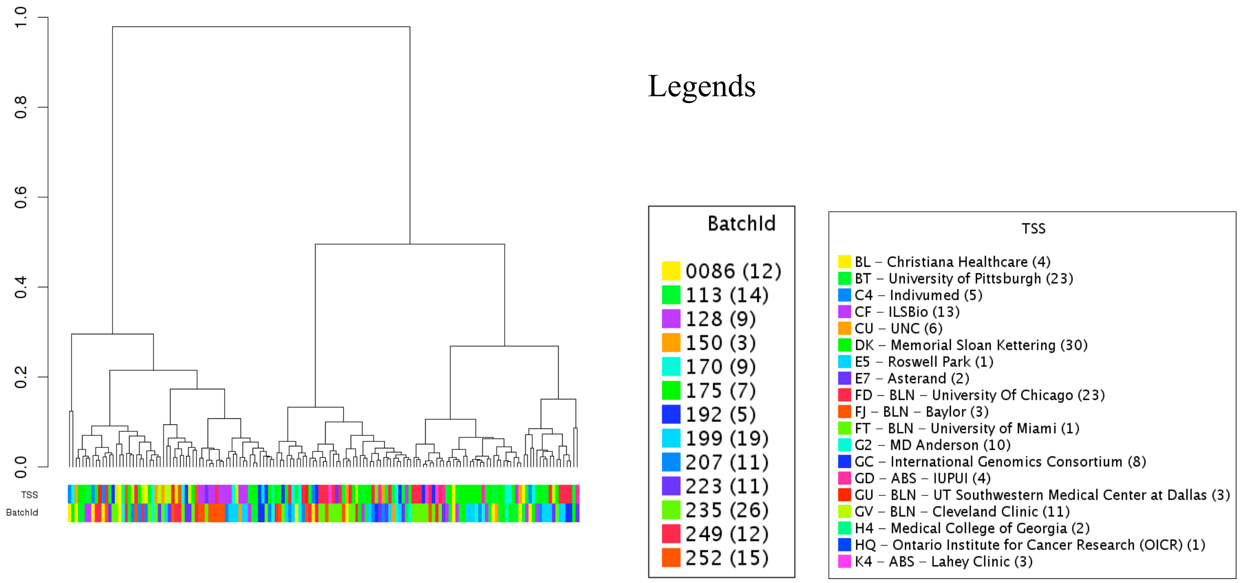


Figure S13.2. PCA: First two principal components for miRNA expression from miRNA-seq data with samples connected by centroids according to batch ID

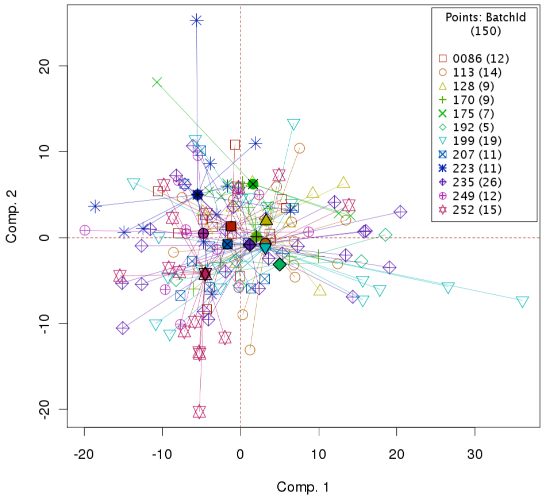
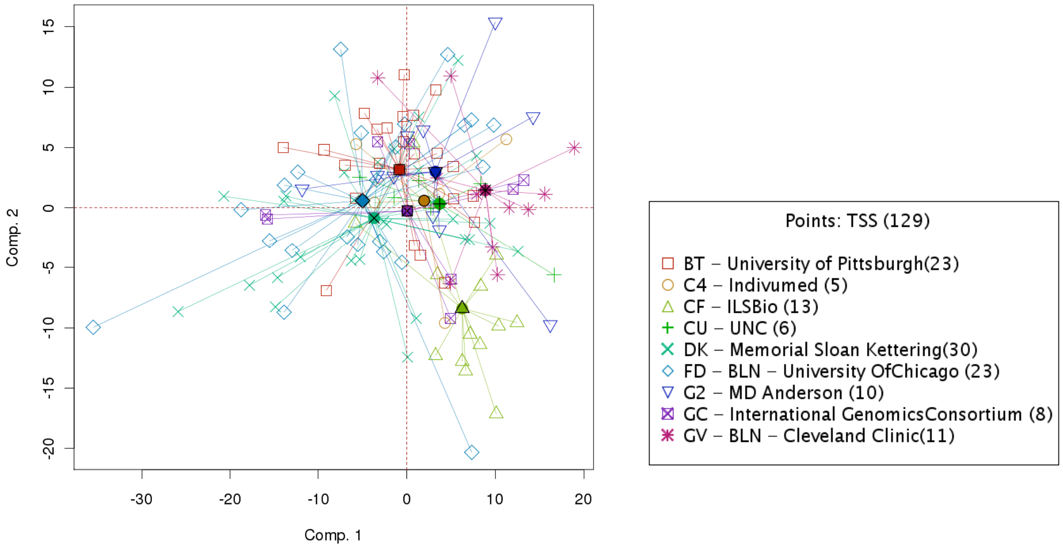


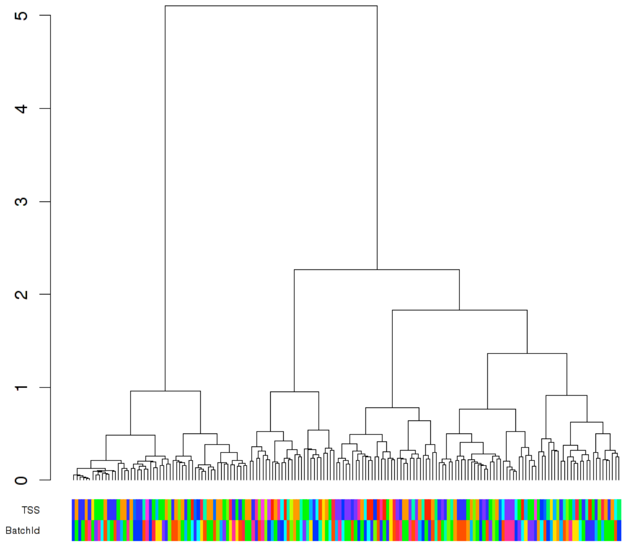
Figure S13.3. PCA: First two principal components for miRNA expression from miRNA-seq data, with samples connected by TSS-wise centroids.



DNA Methylation (Infinium HM450 microarray)

Figures 4-6 show clustering and PCA plots for the Infinium DNA methylation platform. None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.

Figure S13.4. Hierarchical clustering plot for DNA methylation data.



Legends

BatchId

🟨 0086 (12)
🟩 113 (14)
🟪 128 (9)
🟥 150 (3)
🟩 170 (9)
🟧 175 (7)
🟩 192 (5)
🟩 199 (19)
🟦 207 (11)
🟦 223 (11)
🟧 235 (26)
🟦 249 (12)
🟥 252 (15)
🟦 279 (18)

TSS

🟨 BL – Christiana Healthcare (4)
🟩 BT – University of Pittsburgh (23)
🟦 C4 – Indivumed (5)
🟪 CF – ILSBio (13)
🟥 CU – UNC (6)
🟧 DK – Memorial Sloan Kettering (30)
🟩 E5 – Roswell Park (1)
🟦 E7 – Asterand (6)
🟦 FD – BLN – University Of Chicago (30)
🟥 FJ – BLN – Baylor (4)
🟧 FT – BLN – University of Miami (1)
🟩 G2 – MD Anderson (10)
🟩 GC – International Genomics Consortium (9)
🟦 GD – ABS – IUPUI (4)
🟪 GU – BLN – UT Southwestern Medical Center at Dallas (3)
🟥 GV – BLN – Cleveland Clinic (11)
🟨 H4 – Medical College of Georgia (2)
🟩 HQ – Ontario Institute for Cancer Research (OICR) (3)
🟦 K4 – ABS – Lahey Clinic (5)
🟪 MV – University of Minnesota (1)

Figure S13.5. PCA for DNA methylation, with samples connected by centroids according to batch ID.

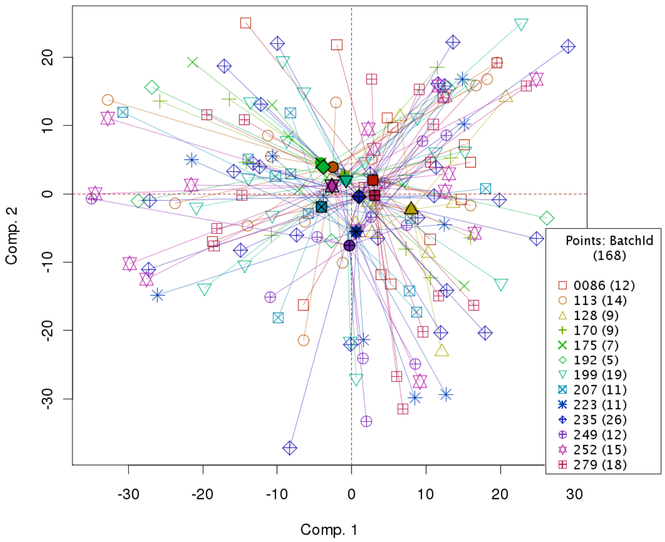
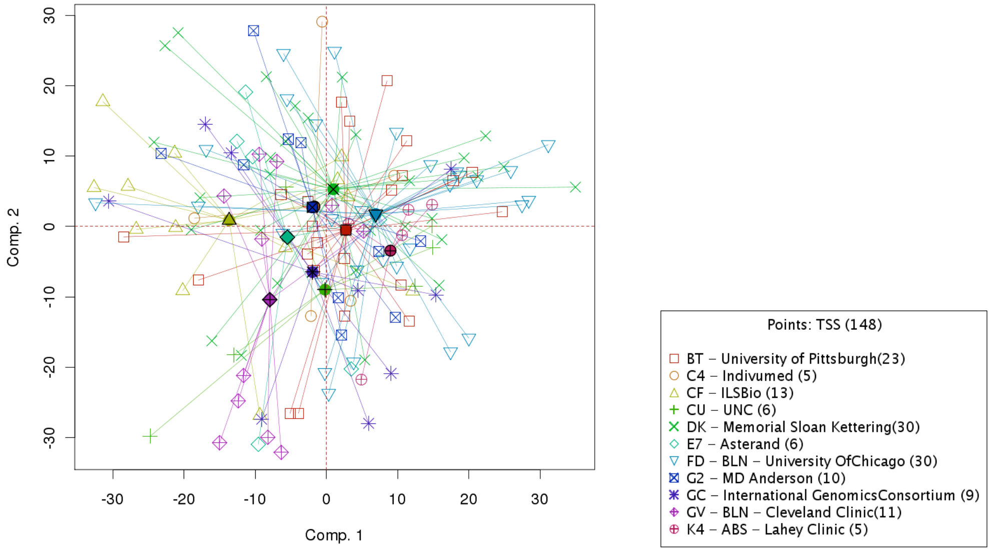


Figure S13.6. PCA for DNA methylation, with samples connected by centroids according to TSS



RNASeqV2 (RNA-Seq Illumina HiSeq)

Figures 7-9 show clustering and PCA plots for the RNA-seq platform. Genes with zero values were removed and the values were $\log_2$-transformed before generating the figures. Once again, the TSS ILSBio showed small batch effects, but not enough to warrant batch effects correction for the type of analyses done in this paper.

Figure S13.7 Hierarchial clustering for mRNA expression from RNA-seq data

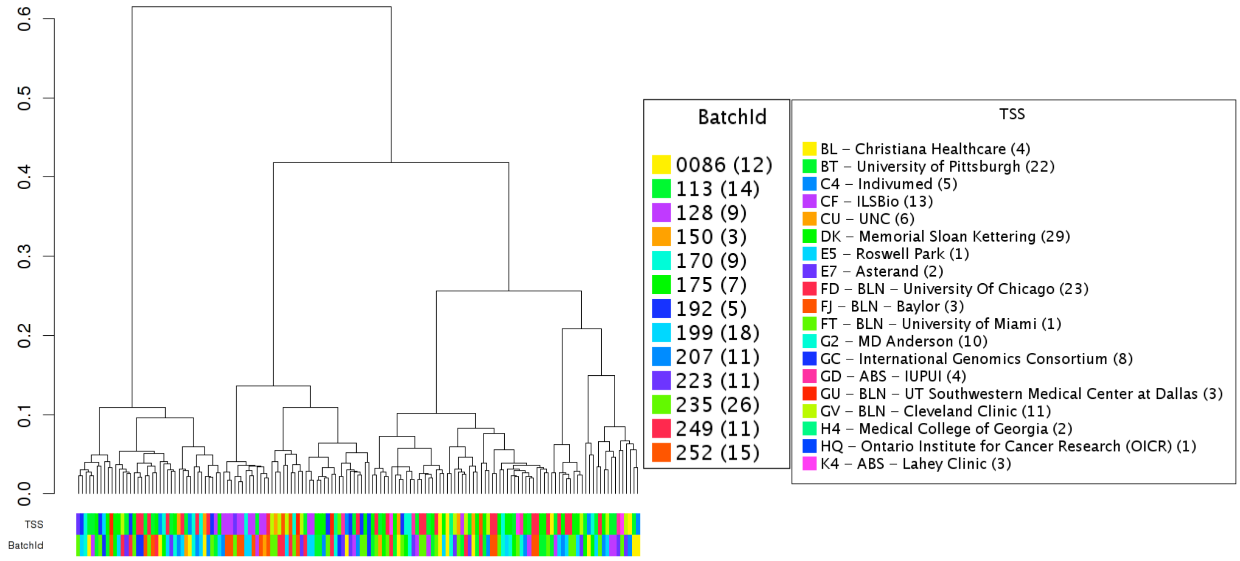


Figure S13.8 PCA: First two principal components for RNA-seq with samples connected by centroids according to batch ID

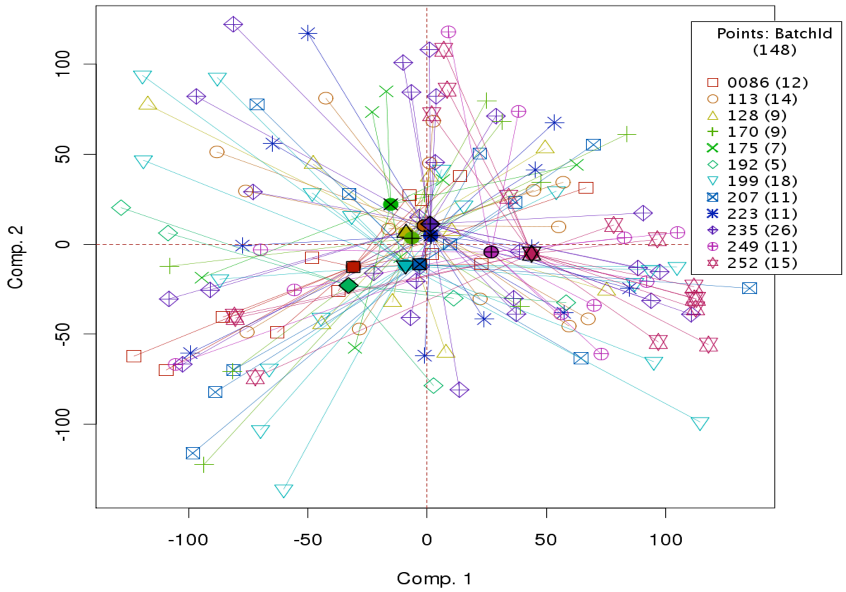
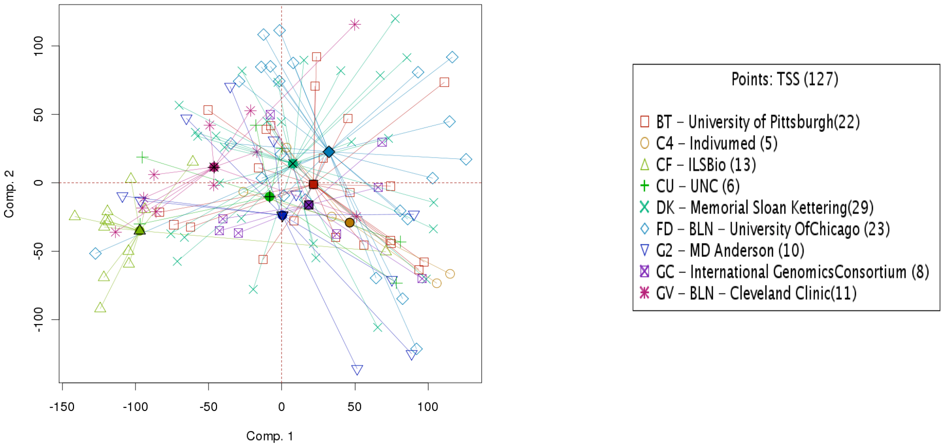


Figure S13.9. PCA: First two principal components for RNA-seq, with samples connected by centroids according to TSS



SNP (SNP 6)

Figures 10-12 show clustering and PCA plots for the Genome Wide SNP 6 platform. Segment values were mapped to gene values for the analysis using Hg19. Batch 235 stood out from the rest, because samples in that batch had very few copy number aberrations. Consequently, we left those samples in the dataset, considering them to reflect actual biology, rather than batch effects.

Figure S13.10 Hierarchical clustering for SNP6 data

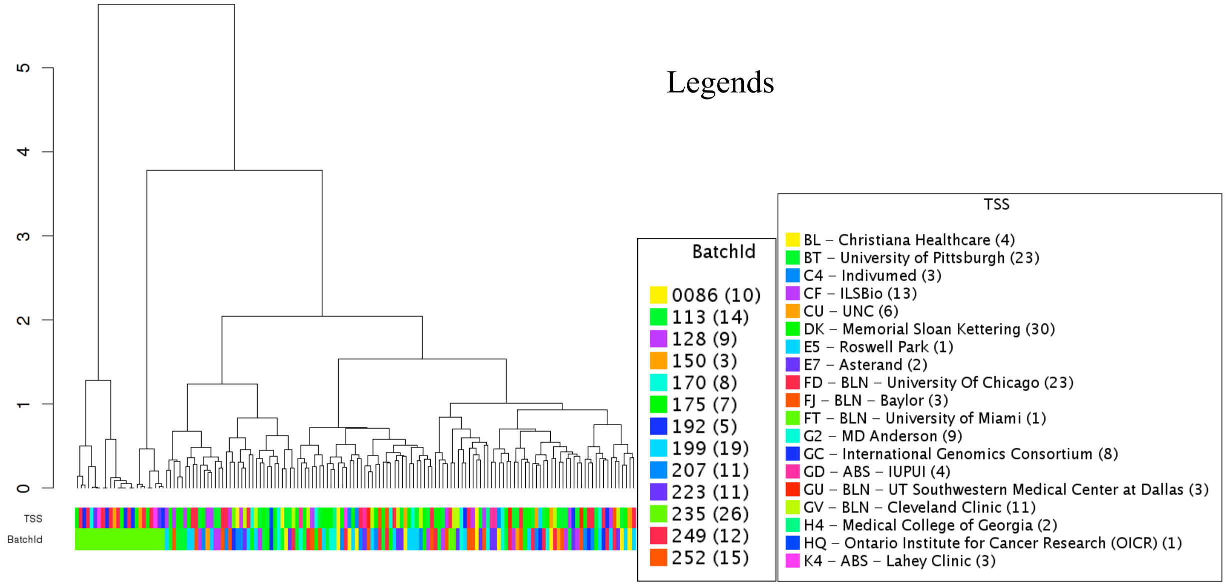


Figure S13.11. PCA: First two principal components for SNP6, with samples connected by centroids according to batch ID

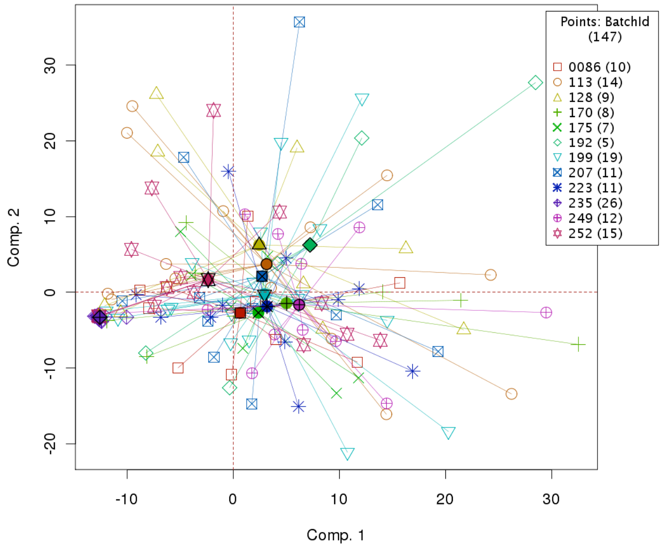
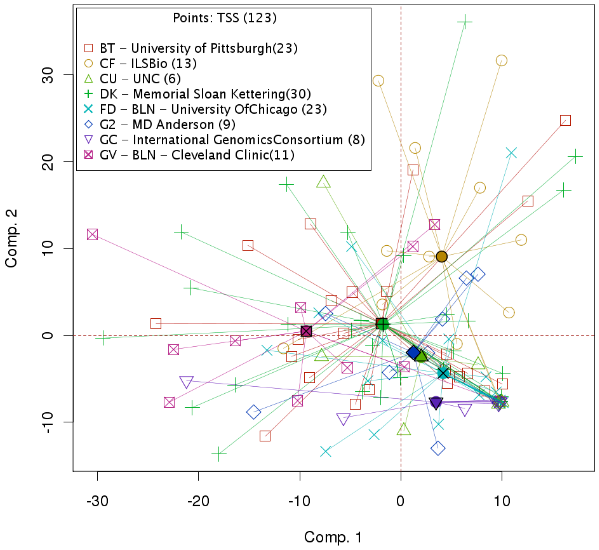


Figure S13.12. PCA: First two principal components for SNP6, with samples connected by centroids according to TSS



Conclusions

Batch effects were analyzed in four different data sets. miRNA and mRNA data showed a small batch effect in samples from the tissue source site ILSBio. However, the batch effects weren’t considered strong enough to warrant algorithmic batch effects correction, since that often removes useful biology along with the batch effects. In SNP 6 data, batch 235 stood out from the rest because the samples in that batch had little copy number variation. That was thought to reflect real biology rather than batch effects, so no correction was applied. DNA methylation data didn’t show any major batch effects.