Multi-platform analysis of 12 cancer types reveals molecular classification within and across tissues-of-origin

Supplemental Data

Table of Contents

Supplemental Figures

Figure S1. Single-platform Clustering Analysis, Related to Figure 1

(A) mRNA expression clustering of 12 Pan-Cancer tumor types. **(B)** Unsupervised clustering of miRNA-seq data. (1) Schematic of an miRNA primary transcript (pri), the trimmed pre-miRNA (pre), reference miRBase 5p and 3p strands, and potential 5’ and 3’ isomiR variation. The gray triangle indicates the 5p/3p-strand data representation used. (2) From the NMF rank survey [1] the profile of average silhouette width suggests a 15-group solution (gray triangle). (3) Consensus membership heatmap. The generally sharp red-vs-blue colors indicate that most samples were placed into clusters consistently over 200 iterations. (4) NMF consensus clustering. Top to bottom: normalized abundance heatmap for 51 discriminatory miRNAs, silhouette width profile, and covariate tracks showing disease type and sequencing platform. (5) Summary table of group number (c), number of samples (n) and average silhouette width (w) for each group, and for the overall set of 4229 tumor samples. **(C)** Clustering of SCNAs across all tumors. SCNAs in tumors (vertical axis) are plotted by chromosomal location (horizontal axis). Tumors are clustered based on copy number alterations (broad and focal) in regions of reoccuring alterations identified by GISTIC 2.0 analyis of the entire set. The heatmap shows the presence of amplifications (red) and deletions (blue) throughout the genome. Colorstrips on the side show tumor type and integrative cluster membership. **(D)** DNA methylation subtypes. Shown in the heatmap are DNA methylation beta values for 4,923 tumors (columns) of twelve tumor types at 2,204 CpG loci (rows) with no pre-existing methylation in normal tissues (left panel). DNA methylation levels are shown as a color gradient from low (blue) to high (red). The top color bar denotes tumor type. BLCA, Bladder Cancer; BRCA, Breast Cancer; COAD, Colon Adenocarcinoma; GBM, Glioblastoma; HNSC, Head and Neck Squamous Cancer; KIRC, Kidney Renal Cell Clear Cell Carcinoma; LAML, Acute Myeloid Leukemia; LUAD; Lung Adenocarcinoma; LUSC, Lung Squamous Cancer; OV, Serous Ovarian Cancer; READ, Rectal Adenocarcinoma; UCEC, Uterine Endometrial Cancer. Bottom color bars indicate the cluster membership, mutation rate (mutations per Mb, capped at 2^6), segment counts representing the level of somatic copy number alteration (SCNA) (capped at 600/genome), genome ploidy, as well as purity estimated based on DNA methylation and copy number changes for each tumor. **(E)** RPPA based unsupervised clustering using 3467 samples and 131 antibodies. Eight different clusters can be seen, mainly driven by tumor type. Several annotation bars are provided at the top of the map, but they haven’t been used in the clustering. **(F)** Pan-Cancer 12 mutation-based subtypes. The heatmap illustrates the proportion of pathways with at least one mutation in the sample (columns) grouped into one of the eight mutated-programs (rows). Samples are further labeled by tissue-of-origin and (for the BRCA samples) BRCA subtype based on a PAM50 call.

Figure S2. Integrated Platform Analysis, Related to Figure 1

(A) Clustering of pairwise concordance between single-platform subtypes. Row color bar indicates the platform from which a given subtype is derived (gold: methylation, gray: miRNA, pink: mRNA, green: mutation, blue: RPPA, turquoise: SCNA). Column color bar indicates the cancer type composition, where color intensity reflects the proportion of the cancer type within a given single-platform cluster (light blue: BLCA, blue: BRCA, pale green: COAD, green: GBM, pink: HNSC, red: KIRC, beige: LAML, orange: LUAD, light purple: LUSC, purple: OV, yellow: READ, brown: UCEC). **(B)** SuperClusters derived from clusters on individual data types using a novel algorithm that equalizes the weights of contributions from all the data types. The columns contain samples. The rows contain the cluster memberships from different data types. The top annotation bar indicates tumor type (Disease) and was not used in the clustering. The disease color legend is given at the top left. Colors for cluster memberships were matched with disease colors as much as possible. Nine different SuperClusters can be seen, mainly driven by tumor type. Some interesting observations are given in the accompanying supplementary text. **(C)** Consensus clustering of top varying PARADIGM features. (1) Heatmap of top varying IPLs. Samples are arranged in order of the consensus cluster tree. IPLs are clustered using Pearson correlation and average linkage for display. (2) Cancer type composition of PARADIGM clusters. (3) Distribution of cluster membership within each cancer type. **(D)** Unsupervised clustering of binary integrative subtype membership. Blue samples (columns) clustered by different integrative methods (rows) with assignments in clusters shown by blue ticks. Row color bar reflects integrative subtyping methods (red: COCA with mutation, pink: COCA without mutation, yellow: SuperCluster with mutation, gold: SuperCluster without mutation, light blue: PARADIGM). **(E)** Comparison of individual mRNA-seq-based subtypes to the integrative COCA subtypes. Samples were clustered according to mRNA-seq-based clustering. Color bars at the top indicate the tissue type, the integrative COCA cluster membership. Remaining rows correspond to the subtypes defined within each tissue by mRNA-seq subtyping where each color represents a cluster.

Figure S3. Clinical Importance of the COCA Subtypes, Related to Figure 1

(A) Overall survival by tumor type. **(B)** Overall survival by COCA subgroups **(C)** Overall survival by a proliferation signature. **(D)** Overall survival by a mutant TP53 gene signature. **(E)** Overall survival by PIK3CA mutation status. **(F)** Overall survival by TP53 mutation status.

Figure S4. Genomic Determinants of the Integrated Subtypes, Related to Figure 2

(A) Frequency of arm level alterations in integrative clusters. Bar graphs show the frequency of arm level amplifications (red) and deletions (blue). Shown are bonforini corrected chi square p values derived by comparing the frequencies of alterations in each cluster to linerage adjusted frequencies in all other tumors. **(B)** Frequency of amplifications and deletions in regions of reoccurring alterations. Regions of significantly reoccurring amplifications or deletions were identified by GISTIC 2.0 analysis of the entire set. The heatmap is colored by the frequency of all types of amplifications (arm level or focal) in each region. Numbers indicate the bonforini corrected chi square p values derived by compaing the frequencies of alterations in each cluster to linerage adjusted frequencies in all other tumors. 1 is $p \leq .05$, 2 $p \leq .01$ and 3 is $p \leq .001$. **(C)** Dendrogram of average SCNA in each integrative cluster. An SCNA profile for each integrative cluster was generated by averaging the copy number alterations across 24174 genes. Integrative clusters SCNA profiles are clustered by Euclidean distance, using Ward's method. **(D)** Clustered heatmap of the frequency of mutation events in the 127 significantly mutated genes (SMGs) across the COCA subtypes. SMG genes (rows) were clustered according to their frequency pattern across 11 of the COCA subtypes (columns). Higher frequency of mutations in a gene for a particular subtype is shown with darker shading. **(E)** Summary of significantly mutated sub-networks identified by HotNet2**.** Summary of significantly mutated subnetworks ($P < 0.01$) identified by HotNet2 run on individual clusters, including only those subnetworks mutated in ≥ 20% of samples in a given cluster. Proteins (nodes) are positioned closer to the clusters in which they are identified as part of a significant subnetwork by HotNet2. Each protein is colored using a pie chart, where individual wedges represent the relative proportion of mutations in that protein restricted to only the clusters in which the protein is found. Pairs of proteins are connected by interactions (edges) if the pair is found to interact in the iRefIndex, HINT, and Multinet interaction networks.

Figure S5. Expression-based Determinants of the Integrated Subtypes, Related to Figure 3

(A) Subtype-specific patterns in gene-program scores can be used to visualize, annotate and interpret the Pan-Cancer integrative subtypes. This heatmap shows the mean gene-program score for each subtype (red=high and blue=low). The subtypes are well-represented by gene-programs, as shown by the classification accuracy barplot above the heatmap. **(B)** Principal component analysis of the mean gene-program scores. Basal breast cancers are most similar to the squamous subtype, and more similar to squamous, ovarian and uterine cancers than to luminal breast cancers. **(C)** MYC signature heatmap and boxplots. The gene program GP5_Myc targets/TERT (upper black arrow) is correlated with most of the 53 MYC related gene expression signatures in our compendium, as shown in this correlation heatmap, and modestly but significantly correlated to MYC copy number aberrations (panel 2; Kendall's tau =0.154, p-value < 2.2E-16). Interestingly, the MYC copy number alteration data (panel 1, lower black arrow) is not highly correlated with the dominant MYC expression signature block, though it is highly correlated to the chr8q24 expression amplicon in the signature set, which contains the MYC gene (panel 3). **(D)** This figure shows surprising similarities between kidney cancer (KIRC) and luminal breast cancer, a highlight of our subset analysis results. Cox proportional hazards survival analysis applied to KIRC identified gene programs reflecting estrogen signaling (GP7), fatty acid oxidation (GP10), and tumor suppressing miRNA targets (GP3) as significantly associated with patient outcome in kidney cancer, along with expression of the PTEN/MTOR signaling axis. **(E)** DNA methylation-based immune signature predictions in each of the integrated COCA subtypes. An immune cell signature score (*Y*-axis) for each tumor sample was determined using tissue specific DNA methylation discriminating white blood cells from the tissue type of origin for that tumor using the method described in the TCGA colorectal manuscript [2]. Samples were then grouped by their integrative COCA subtypes (X-axis). Boxplots reveal an appreciable variation in immune signatures, reflective of differences in immune-related components, in these different subtypes. Note, because this method relies on tissue-specific methylation in normal samples, estimates could not be derived for AML samples, which are the major members of C13. Shown for Cluster 13 are the results applied to the restricted set of the few solid tumors that occur in this cluster that have extreme immune cell infiltration as can be seen by the predicted higher immune cell fraction. **(F)** Box plots for each of the 9 pathways, showing differentially expressed pathways between COCA clusters. Pathway scores were computed for each sample. The samples were then grouped by the COCA clusters to generate the box plots. The lower edge, middle notch, and upper edge of each box represents the 25th, 50th, and 75th percentile pathway scores, respectively, in the given COCA cluster. The bottom whisker represents the lowest score within 1.5 inter-quartile ranges of the lower quartile, and top whisker represents the highest score within 1.5 inter-quartile ranges of the upper quartile. Any scores outside of that range are shown as outliers.

Figure S6. Multi-platform Determinants of the Integrated Subtypes, Related to Figure 3

(A) Uniquely differential pathways across 5 data platforms and 11 COCA subtypes. In each subtype, pathways are ranked according to their enrichment score (ES); pathways supported by more than one platform are ranked at the top. Their ES score (purple and dark blue squares in the heatmap) is offset by 1 in order to distinguish them from single-platform supported pathways (shades of red). **(B)** Elastic Net features for COCA subgroups. Elastic Net was used to identify features (miRNA, mRNA modules, protein, DNA copy number, and DNA mutation) characteristic of each subtype. The full set of features is available as Data File S3 (syn2486685).

Figure S7. Convergence of Squamous-like Subtype and Features Common to Squamous, Breast Basal, and Ovarian Subtypes, Related to Figure 4 and Figure 5

(A) (1) Putative mutational drivers exhibiting higher mutation frequency among samples of both the majority (HNSC) and minority (mainly LUSC and BLCA) tissues of the C2-Squamous-like cluster as compared to the remaining samples of the Pan-Cancer data set (Qmajority and Q-minority, respectively). Those genes exhibiting a corrected p value <5% in both comparisons are depicted. The 291 high-confidence drivers retrieved from a combinatorial approach based on detecting complementary signals of positive selection have been included in the present analysis [3]. Mutation frequency has been compared with Fisher's exact test and corrected for false discovery rate. Green cells: protein-affecting mutations; upper bar: tumor type. (2) Mutation relation of COCA subtype C2-Squamous-like. Each row represents one significantly mutated gene. Columns represent individual sample mutations for samples in the C2-Squamous-like subtype. Samples are grouped by tumor tissue type as labeled on row 1. Red indicates increased numbers of somatic mutations in a sample. Samples with no mutations are not shown. **(B)** PARADIGM SuperPathway regulatory sub-network defining the 2-Squamous-like integrative subtype. Only features interconnected through hubs with > 15 regulatory interactions are shown. Color of the nodes reflects activation (red) or repression (blue) within the squamous subtype. Edge color denotes interaction type: inhibitory (green) and activating (purple). Node shape reflects feature type: protein (circle), complex (diamond), miRNA or RNA (square), abstract concepts (vee). **(C)** (1) Pathway commonalities between the Squamous and Basal integrative subtype. Only features with significant basalness score interconnected through hubs with > 5 regulatory interactions are shown. Hubs are labeled in bold. Color of the nodes reflects activation (red) or repression (blue) within the basal and squamous subtypes. Edge color denotes interaction type: inhibitory (green) and activating (purple). Node shape reflects feature type: protein (circle), complex (diamond), miRNA or RNA (square), abstract concepts (vee). (2) Pathway commonalities between the Ovarian and Basal integrative subtype. (3) Pathway commonalities between the Squamous, Ovarian and Basal integrative subtype. **(D)** (1) CircleMap of the PARADIGM-SHIFT differences across the Pan-Cancer 12 integrative subtypes. Samples were ordered first by integrative subtype membership (first ring), then by TP53 mutation status (second ring), and finally by P-Shift (outer ring). Red-blue color intensity reflects magnitude (red: positive, blue: negative). Negative P-Shift scores (outer ring blue) predicts LOF. (2) CircleMap of the TP53 PARADIGM-SHIFT predictions and inferred activities of p53 and p63 (dNp63a and TAp63g isoforms). (3) CircleMap showing expression and inferred activities of common TP63 and TP53 targets in the Ovarian, Basal, and Squamous integrative subtypes.

Figure S8. Divergence of the Bladder Cancer Subtype, Related to Figure 6

(A) PARADIGM SuperPathway regulatory sub-network differentiating C2 from C8 bladder cancer cases. Only features interconnected through nodes with > 5 regulatory interactions are shown. Color of the nodes reflects activity (red: high, blue: low) within the C2-Squamous-like relative to the C7-BLCA bladder cancer cases. Edge color denotes interaction type: inhibitory (green) and activating (purple). Node shape reflects feature type: protein (circle), complex (diamond), miRNA or RNA (square), abstract concepts (vee). **(B)** Survival analysis of bladder samples by histology in main COCA groups. Overall survival for bladder cancer samples in three Pan-Cancer subtypes are portrayed by Kaplan-Meier survival plots with a Log Rank test for significance. X indicates censored data from either loss to follow-up or information at last checkup.

Supplemental Tables and Data Files

Supplemental Tables

- Table S1. COCA assignments for each Pan-Cancer-12 sample across platforms, Related to Table 1
- Table S2. Mutation frequencies of 127 SMGs based on COCA subtypes, Related to Figure 2
  - (A) Top 40 SMGs and their frequencies in COCA subtypes
  - (B) Mutation frequencies of all 127 SMGs across COCA subtypes
- Table S3. HotNet2 sub-network analysis, Related to Figure 2
  - (A) Sub-networks identified by HotNet2 on each integrated subtype
  - (B) Sub-networks identified by HotNet2 on the Squamous subtype (subtype 2)
- Table S4. Gene-program analysis, Related to Figure 3
  - (A) Gene-program and Paradigm pathway annotations of the 11 integrated subtypes
  - (B) Pan-Cancer 12 Gene-programs (GPs), dominant themes, and most representative published signatures
  - (C) Integrated Pan-Cancer 12 subtypes annotated using Gene-programs and the full signature set
  - (D) Selected pathways associated with drug targets or canonical to cancer
  - (E) Summary of significant associations to OS in multivariate Cox Proportional Hazards models of the Pan-Cancer 12 dataset
  - (F) Common significant associations to OS in multivariate Cox Proportional Hazards models of C5: KIRC and C3: Luminal BRCA, in subset models with cancer stage as a covariate
- Table S5. PARADIGM pathway enrichment analysis of the COCA subtypes, Related to Figure 4 and Figure 5
  - (A) PARADIGM pathways overrepresented by EASE analysis in each of the COCA integrated subtypes
  - (B) PARADIGM pathways enriched in Squamous-like, BRCA-Basal, and Ovarian comparisons using EASE scoring
- Table S6. Bladder cancer subtype analysis and correlations with survival, Related to Figure 6
  - (A) Correlation of Bladder classifications to COCA subtypes
  - (B) Correlation of Bladder samples, COCA subtypes, and Bladder histology
  - (C) Multivariate Cox Proportional Hazards model for Bladder samples

Supplemental Data Files

- Data File S1. Mutated-programs determined from the mutated pathways, Related to Figure 1
- Data File S2. Platform-specific COCA features, Related to Figure 1
- Data File S3. Elastic Net COCA features, Related to Figure 1
- Data File S4. 33 published p53 signatures, Related to Figure 5
- Data File S5. Detailed enrichment results for gene programs and selected pathways, Related to Figure 3.

Extended Experimental Procedures and Analysis

Data for the complete TCGA sample set were obtained for the December 22, 2012 Pan-Cancer-12 data freeze from the Sage Bionetworks repository, Synapse (https://www.synapse.org/#!Synapse:syn300013). As part of this article's resource, associated data files are available from a single Synapse project dedicated to this manuscript, available from the page: https://www.synapse.org/#!Synapse:syn2468297, where the provenance of the data is tracked as an explicit web of interlinked results. These data files are stored in formats suitable for convenient downstream programmatic access, including as downloadable delimited tables and as objects for loading into a running R session. The Pan-Cancer clustered heat maps in the figures in this paper can be explored interactively (with zooming and navigation capabilities) at http://bioinformatics.mdanderson.org/main/Pancan12Subtypes:Overview. Interactive views of the datasets are available through Gitools (http://www.gitools.org/datasetsdev/pancancer12) and the UCSC Cancer Genomics Browser (https://genome-cancer.ucsc.edu/proj/site/hgHeatmap/?datasetSearch=PANCAN12).

Section 1: Single-platform Clustering Analysis

The location of the input data and subtype assignments on Synapse for each platform is indicated in the table below.

Platform	Input Data Matrix	Subtype Assignments
mRNAseq	syn1715755	syn1715788
miRNAseq	syn2491366	syn2027079
Somatic Copy Number (SCNA)	syn1710678	syn1712142
Methylation	syn2486658	syn1875816
Reverse Phase Protein Array (RPPA)	syn1759392	syn1756922
Mutated Pathways	syn2495279	syn2492003

1.1 mRNAseq clustering

Data Generation: mRNA sequencing for colon (COAD), rectum (READ), breast (BRCA), endometrial (UCEC), kidney clear cell (KIRC), lung adenocarcinoma (LUAD), lung squamous carcinoma (LUSC), head and neck squamous (HNSC), and bladder (BLCA) was performed at UNC as previously described [2, 4-6]. Acute myeloid leukemia (AML) samples were sequenced at the British Columbia Cancer Agency [7] and ovarian (OV) and glioblastoma multiforme (GBM) samples were sequenced at the Genome Sequencing Centers (GSCs). BAM files for AML, OV and GBM were downloaded from CGHub (www.cghub.ucsc.edu) and converted to FASTQs. Then all samples were processed similarly aligning reads to the hg19 genome assembly using MapSplice [8]. Gene expression was quantified for the transcript models corresponding to the TCGA GAF 2.13, using RSEM4 and normalized within-sample to a fixed upper quartile [9]. Gene level expression data is available at the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/). COAD, READ, UCEC and AML were sequenced on Illumina GAIIx while the other samples were sequenced on Illumina HiSeq. We used a set of 19 colon samples that were sequenced on both platforms to estimate platform differences. A limitation of this approach is that the platform correction was restricted to the 16,116 (out of the 20,531 total) genes expressed in colon, defined as those with 3 or more reads. Upper quartile normalized RSEM data was log2 transformed. Genes with a value of zero were set to the missing value after log2 transformation and genes were filtered if they had missing data in greater than 30% of samples. For the 19 colon samples sequenced on each platform, within each dataset the gene median were calculated. The difference between the GAII platform and the HiSeq platform was calculated and subtracted from the full set of GAII data. The corrected GAII set was merged with the HiSeq data set followed by gene median centering. The platform-corrected input data is made available on Synapse as part of the Pan-Cancer 12 data freeze (syn1715755).

Tissue-Dependent Clustering: Using the platform corrected mRNAseq data, genes were filtered for those present in 70% of samples and then the top 6,000 most variable genes were selected. ConsensusClusterPlus R-package [10] was used to identify clusters in the data using 1000 iterations, 80% sample resampling from 2 to 20 clusters (k2 to k20) using hierarchical clustering with average innerLinkage and finalLinkage and Pearson correlation as the similarity metric. Eleven main groups were identified when 16 clusters were used (Figure S1A). These 11 groups were observed to be stable through the use of 20 clusters (K20) and significant in pairwise comparisons of the 11 main clusters with SigClust [11]. The subtypes were deposited into Synapse (syn1715788).

1.2 miRNAseq clustering

For miRNA-seq data, normalized read count data (reads per million, RPM) for 4229 tumor samples was compiled into an abundance matrix for 5p and 3p mature (processed) miRBase strands, as described in publications for each cancer type [4]. Strands corresponding to miRNAs that had been removed from v18 miRBase (miRNA.dead) were eliminated from the data matrix. Because data matrices for some disease types had been generated using v13 miRBase annotations, and others v16, we filtered the matrix to contain only the subset of 860 miRNA 5p and 3p strands that was present in both miRBase versions. We ranked 5p and 3p strands by RPM variance across the samples, and input into NMF v0.5.02 [1] a data matrix subset consisting of the most variant 25% of strands for unsupervised consensus clustering (R v2.12.0). We used NMF's default Brunet algorithm, with 25 iterations for the rank survey and 200 iterations for the clustering runs, and generated clustering results with between 3 and 25 clusters. We selected a preferred clustering result by considering the profile of average silhouette width [12] of the consensus membership matrix, for which we generated silhouette results by applying the R 'cluster' package v1.14.1 to a default distance matrix generated from the NMF consensus membership matrix. We generated an miRNA abundance heatmap for the NMF result by ordering the columns of the RPM abundance matrix to match the sample order in the NMF output, and then retaining only the rows for the subset of 51 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 [13] we log-transformed and median-centered the 51 miRNA abundance profiles, and then hierarchically clustered only the rows using an absolute centered correlation and average linkage. We visualized the resulting matrix with Java Treeview [14]. See Figure S1B. The input data matrix for miRNAseq clustering is available in Synapse at syn2491366 and the subtype assignments are at syn2027079.

1.3 Somatic copy number (SCNA) clustering

Generation and GISTIC analysis of somatic copy number alteration data from SNP6.0 arrays is described elsewhere [15]. For copy number based clustering, tumors were clustered based on thresholded copy number at reoccurring alteration peaks from GISTIC analysis. Tumors were hierarchical clustered in R based on Euclidean distance using Ward's method. The number of cluster groups was chosen based on cophenetic distances generated from clustering. For comparison of broad and focal alteration between cluster of cluster groups, frequency of alterations in each cluster group was compared to the average frequency of all other groups by chi squared tests with an added Bonferroni correction to control for multiple testings. See Figures S1C and S4A-C. The input data matrix for SCNA clustering is available in Synapse at syn1710678 and the subtype assignments are at syn1712142.

1.4 DNA methylation clustering

We performed a moderate probe-design dependent platform normalization to remove systematic platform bias, and generated a merged dataset on 25,978 probes shared by the HM27 and HM450 platforms. We also used two sets of technical replicates (TCGA-07-0227 and TCGA-AV-A03D) that were repeatedly measured as internal controls (99 and 74 times respectively) across platform and batch to monitor residual batch and platform variations. Any probes with a standard deviation of >0.05 is removed from the clustering analysis, so batch and platform had minimal impact on the clustering. Illumina Infinium DNA methylation probes that overlap with single nucleotide polymorphisms (SNPs) and repeats, or map to sex chromosomes were masked from the analysis.

For the clustering analysis, we focused on CpG loci that are unmethylated in normal tissues. Therefore, we removed probes that showed methylation (median beta value > 0.2) in any of the 12 matched normal tissue types included in the current study. After the aforementioned filters, 11,696 probes remained. As these loci are mostly within CpG islands that remain constitutively unmethylated in normal tissues, we dichotomized the beta values in the tumors at 0.3. Tumors with a beta value of 0.3 or greater are designated methylated and tumors with a beta value of lower than 0.3 are designated unmethylated. The dichotomization greatly ameliorated the effect of tumor sample purity on the clustering, and further removed most residual batch/platform effects that are primarily reflected in small variations near the two ends. We selected the 2,203 probes that were methylated in more than 10% of any of the tumor types or 50% of any of the well-defined subtypes for clustering. We used hierarchical clustering with Ward's method on the Jaccard Distance, a distance measure that best suits binary data. The dendrogram was cut at different levels with the 'cutree' function in R and evaluated for associations with clinical data and the k=19 result was used. See Figure S1D. The input data matrix for methylation clustering is available in Synapse at syn2486658 and the subtype assignments are at syn1875816.

1.5. RPPA clustering

Data Generation: Protein was extracted from all the samples 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 [16-21]. 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 [16]. During the process, the raw spot intensity data were adjusted to correct spatial bias before model fitting. A QC metric [20] 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 [17, 19, 20] using median centering across antibodies (level 3 data). In total, 131 antibodies and 3467 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 [22]. 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 [22].

Six 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 [16, 17]. 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.

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 3467 RPPA Pan-Cancer samples were run in a total of 6 batches. That presented the potential problem of batch effects when trying to merge together all the batches. Batch effects in RPPA data are a known problem, even when critical materials such as the treated glass slides, antibodies, and enzymes are consistently obtained from one manufacturer [23]. To address that problem, we developed a new algorithm, Replicates Based Normalization (RBN), which uses replicate samples run across multiple batches to adjust the data for batch effects. The underlying hypothesis is that any observed variation between replicates in different batches is primarily due to linear batch effects plus a component due to random noise. Given a sufficiently large number of replicates, the random noise is expected to cancel out since it has a mean of zero, by definition. Remaining differences are treated as systematic batch effects. We can compute those effects for each antibody and subtract them out for the slide.

In one of the 6 batches, batch #64, we ran many replicate samples that were common with the other 5 batches. The number of common samples with each batch was 69 or more. We designated batch 64 as the “anchor” batch that would remain unchanged. We then computed the means and standard deviations of the samples in common between batch 64 and each of the other batches. The difference between the means of each antibody in the two batches and the ratio of the standard deviations provided an estimate of the systematic effects between the batches for that antibody (both location-wise and scale-wise). Each data point in the non-anchor batch was adjusted by subtracting the difference in means and multiplying by the inverse ratio of the standard deviations to cancel out those systematic differences.

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. We identified eight robust clusters. The eight clusters and their protein expression patterns are shown in Figure S1E. As expected, most of the clusters are driven by tumor type. A few notable exceptions include basal and Her2 breast samples, which don’t cluster near the luminal breast samples; bladder (BLCA) samples, which cluster mainly with breast basal and Her2 samples; and head and neck (HNSC), lung squamous (LUSC) and lung adenocarcinoma (LUAD) samples that fall into a single cluster. Colon (COAD) and rectal (READ) samples cluster together, indicating that their proteomic profiles are very similar. The RPPA cluster memberships have been used for downstream analysis, such as the Cluster of Clusters Analysis in the main text and are discussed further there. The input data matrix for RPPA clustering is available in Synapse at syn1759392 and the subtype assignments are at syn1756922.

Batch effect and RPPA analyses were supported in part by Grant Numbers U24CA143883 and P30CA016672 from the National Cancer Institute (UT-MD Anderson TCGA Genome Data Analysis Center, and Cancer Center Support Grant, respectively), by a gift from the H.A. & Mary K. Chapman Foundation, and by a grant from the Michael & Susan Dell Foundation honoring Lorraine Dell.

1.6. Mutation clustering

Datasets: We created a compendium of 14,575 previously described pathways (syn1741407). The compendium was assembled from 7 databases – GO (http://www.geneontology.org/), KEGG (http://www.genome.jp/kegg/pathway.html), NCI (http://pid.nci.nih.gov/), Omim (http://www.ncbi.nlm.nih.gov/omim), Reactome (http://www.reactome.org/), BioCarta (http://www.biocarta.com/genes/index.asp), GenMapp (http://www.genmapp.org/). We collected mutation events from Synapse MAF file (syn1729383) for 3269 samples in Pan-Can-12 tumor types. The mutation data in these curated MAF files were adjusted to exclude non-somatic variants, any mutations marked as silent (including events occurring in predicted pseudogenes), and those occurring in genes with only a single event in any one tissue. The per-gene coverage, gene length and mutations/Mbp were extracted from the publicly available output of the MuSiC suite (syn1713813). Finally, small genes with just 1 or 2 observed mutations tend to produce extremely high mutation rates/Mbp. To prevent them from biasing the enrichment results, we excluded mutation/Mbp measurements for genes with less than 3 non-silent mutations or less than 5-fold exon coverage. The added filters can be considered very lenient and served to eliminate only a few extreme outliers (mostly ncRNAs).

Identification of Frequently Mutated Pathways: We collected pathways associated with highly mutated genes across all tumors. To this end, we used mutation rate per megabase as the score for a gene, which avoids any gene size biases. We then took these gene scores as the rank statistic and fed them as input to the Gene Set Enrichment Analysis program [32] to find pathways having a significant number of high scoring genes (i.e. with multiple highly mutated genes). Random permutations of the mutations/Mbp values were used to construct a background empirical distribution for P-value calculations. Multiple hypothesis correction was done by controlling the False Discovery Rate [37]. A pathway was considered positively enriched for mutations if it had enrichment score > 0.75 and FDR< 0.05. Out of the 14,575 total pathways used, 214 met both criteria. To eliminate pathways that were too narrow or too broad in their definition we filtered for enriched pathways that had between 5 and 200 genes, leaving 150 pathways. Finally, to eliminate the significant redundancy inherited from the original compendium, we removed any pathway that had more than 80% overlap with a pathway more highly enriched for mutations. This gave a final set of 96 non-redundant pathways (Data File S1, syn2468302).

Identification of Mutated-Programs: Many of the resulting non-redundant pathways exhibit mutual correlation with respect to their patterns of mutations across the PanCan-12 dataset, suggesting they can be organized into a higher-level organization to capture the relatedness between pathways. To this end, we clustered the 96 pathways using affinity propagation clustering [50]. To identify the best parameters, we performed a parameter sweep by varying the self-similarity quantile from 0 to 0.7 in steps of 0.1 and the damping factor from 0.4 to 0.9 in steps of 0.1 [49] and chose the values that gave the best average ranking on three different cluster validity metrics -average silhouette width, average distance within clusters, and average distance between clusters (fpc R package). This resulted in the identification of 8 groups of pathways that we refer to here as *mutated-programs* that ranged in size from 1-2 to 23-26 constituent pathway members. Note that the list of 96 pathways included below has been grouped into one of the eight mutated-programs identified by the AP clustering step. Not surprisingly, the larger mutated-programs were related to TP53 and the PIK3-family of kinases – in fact, TP53 was present as a pathway member in three of the eight mutated-programs. The matrix of mutated programs alterations across the samples used for subtyping based on mutation data is available as part of the Synapse resource (syn2495279).

Subtyping PanCancer-12 Samples with Mutated-Programs: We could then compute a vector of mutated-program activity scores for each sample – an activity score was calculated as the proportion of pathways within a mutated-program in which a sample had at least one mutation. The same apcluster approach used for determining the mutated-programs was used in clustering PanCancer samples into subtypes. We again used

Affinity Propagation but this time used to cluster the samples and swept through parameters checking three separate clustering metrics as described above. Some parameter combinations lead to clustering solutions that failed to converge – those, as well as any converged solutions that resulted in more than 60 subtypes, were excluded from consideration. The best PanCancer mutation subtype solution under the three cluster validation criteria previously discussed had 14 subtypes (Figure S1F). A complete list of subtype sample membership is available (Table S1).

Section 2: Integrated Platform Analysis

2.1 Cluster of Cluster assignments (COCA)

Subtype calls from each of the 5 platforms analyzed for subtypes within each data type were used to identify relationships between the different classifications. Subtypes defined from each platform were coded into a series of indicator variables for each subtype. The matrix of 1 and 0s was used in ConsensusClusterPlus R-package [10] to identify structure and relationship of the samples. Parameters for Consensus cluster were 80% sample resampling with 1000 iterations of hierarchical clustering based on a Pearson correlation distance metric.

The distribution of subtypes derived from single tumor type-specific analysis of mRNA-seq data alone was compared to the COCA subtypes to reveal how integrative clustering might redefine even the subtyping within each tumor type. Figure S2E shows the result of comparing each individual tumor type's mRNA-seq subtype against the COCA subtyping. Intriguingly, much of the within-tumor-type cluster relationships are preserved. For example, the luminals split out from the HER2 samples within the breast group.

2.2 Pairwise concordance of single-platform clusters

To evaluate the concordance between the multiple subtypes (clusters) identified by each of the 6 different assay platforms (DNA methylation, mutation, and SCNA; mRNA and miRNA expression; RPPA), we computed the proportion of sample overlap between each pair of platform clusters as the number of shared samples relative to the total number of unique samples in the two groups. The analysis was restricted to the 3527 samples also given Integrated Subtype (cluster of cluster) assignments; and altogether, 80 different platform clusters (16 mRNA, 8 SCNA, 8 RPPA, 19 Methylation, 15 miRNA, and 14 Mutation subtypes) were considered. Of note, the pairwise overlap between clusters from the same assay platform is by default 0; and when comparing between platforms, only samples that were assigned to both assay platform subtypes were considered in the computation. Cancer type composition (proportion) within each cluster was determined and converted to a color scale appropriate to the Pan-Cancer 12 cancer type color designation. The pairwise overlap was then clustered (Pearson correlation, average linkage), and the corresponding heatmap display of the pairwise overlap was generated using the heatmap.plus package in R (Figure S2A).

Hierarchical clustering by pairwise concordance placed 70 of the 80 single-platform clusters into 10 major groups with significant cancer type associations. Single platform clusters not within these 10 major groups include: mRNA clusters 2, 3, 5, 7 and 14 (which has ≤ 10 samples), miRNA cluster 9, Methylation clusters 12 and 16, and Mutation clusters 6 and 7.

As expected, single platform clusters predominantly composed of BRCA samples form two distinct groups, reflecting the basal vs. luminal BRCA subsets. Apart from the clusters containing COAD and READ samples which were expected to group together, the concordance clustering reveals two other groups of mixed tissue clusters: one comprised of LUSC and HNSC cases, and another comprised mostly of LUAD and BLCA cases. While the mixing of LUSC/HNSC cases is seen by all the individual platform subtype analyses, the grouping of LUAD and BLCA cases together was observed only by the miRNA (cluster 5) and methylation (cluster 5 and 6) subtyping.

The distance between these two groups of mixed tissue clusters (as assessed by pairwise concordance) is smaller than that separating the two sets of predominantly BRCA clusters, suggesting that the genomic distinctness of the basal from the luminal BRCA samples may be even greater than the genomic differences between LUSC, HNSC, LUAD and BLCA cases assessed by at least some of the assay platforms.

Generally, within the pre-dominantly single cancer type concordance groups, mRNA and Methylation clusters tend to show stronger 1:1 correspondence with the single cancer type than do other platform clusters. Examples of this are apparent in the luminal-BRCA grouping, where the SCNA, Mutation and miRNA clusters are less concordant and have much more diverse cancer type compositions (mixtures with LUAD, UCEC, LUSC, OV and/or HNSC cancers); the basal-BRCA grouping, where the RPPA and miRNA clusters are more heterogeneous (containing BLCA, UCEC and LUSC cases, respectively); and the KIRC grouping, where RPPA, miRNA and SCNA clusters are more heterogeneous (and includes BRCA and UCEC samples). This latter group contains two methylation clusters – a homogenous Methylation cluster 17 (99.5% KIRC) and a more heterogeneous Methylation cluster 2 (with >10% BRCA and UCEC).

While this stronger correspondence between methylation and mRNA clusters with cancer type may in part be due to the higher k selected for these platform-specific subtype assignments, only 11 of the 16 mRNA subtypes were included in the 10 major concordance groups. Tissue dependency may play a significant role here, as the SCNA and Mutation clusters which show greater cancer type heterogeneity are tissue-independent assay platforms. However, DNA methylation subtyping is similarly tissue-independent, and yet produces some of the most homogenous cancer type clusters. Conversely, the 15 miRNA clusters, all but one of which were included in the 10 different concordance groups, was presumably tissue dependent, and yet exhibits diverse cancer type compositions. Altogether, these observations suggest there may be inherent differences in the biological information represented within each of the platform data types that goes beyond tissue dependency contributing to the differences in concordance and cancer type compositions between the platform-specific clusters.

2.3 SuperClusters

We developed a new clustering algorithm, called SuperCluster, to derive overall subtypes for the samples based on their cluster memberships of different data types [mRNA, protein (RPPA), DNA methylation, miRNA, and copy number variations (CNV)]. The algorithm adjusted the contribution from each data type so that their relative weights were all equal. All the clusters memberships were treated as nominal variables. Disease type was not used for clustering. The results are shown in Figure S2B where nine super clusters can be seen. As expected, the clusters are mainly driven by tumor type. There is good correlation across the platforms, once again mainly due to tumor type, with DNA methylation, mRNA and protein (RPPA) showing excellent correlation. miRNA and CNV are less correlated. Breast (BRCA), endometrial (UCEC), bladder (BLCA) and basal breast don't show any tissue specific copy number signatures. On the other hand, renal (KIRC), GBM and AML show very distinctive tissue signatures across all platforms (GBM miRNA and AML RPPA data were not available). Colon (COAD) and rectal (READ) samples mixed together quite well, indicating that they are very similar to each other. Surprisingly, bladder (BLCA) samples mix in with breast basal and HER2 samples, especially by RPPA and a little by CNV. Some of the lung squamous (LUSC) samples cluster with head and neck (HNSC), whereas other LUSC samples cluster with lung adenocarcinoma (LUAD). The LUSC split seems mainly driven by mRNA and CNV. A comparison of the memberships by SuperCluster and by Cluster of Cluster Analysis is given in the main text.

2.4 PARADIGM integrated pathway analysis

Integration of copy number, mRNA expression and pathway interaction data was performed on the 3531 samples using the PARADIGM software [24]. 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. Expression and gene copy number data was obtained from Synapse (syn1715755 and syn1695369 respectively). The platform-corrected, median-centered mRNA data was rank transformed and discretized prior to PARADIGM analysis.

Pathways were obtained in BioPax Level 3 format, and included the NCIPID and BioCarta databases from http://pid.nci.nih.gov and the Reactome database from http://reactome.org. 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 henceforth referred to collectively as pathway concepts. The resulting pathway structure contained a total of 17365 concepts, representing 7325 proteins, 7813 complexes, 1574 families, 52 RNAs, 15 miRNAs and 586 processes.

The PARADIGM algorithm infers an integrated pathway level (IPL) for each gene that reflects a gene's activity in a tumor sample relative to the median activity across all tumors. An initial minimum variation filter (at least 1 sample with absolute activity > 0.05) was applied, resulting in 13480 concepts (5380 proteins, 6282 complexes, 1407 families, 9 RNAs, 15 miRNAs and 387 processes) with relative activities showing distinguishable variation across tumors.

To identify subtypes implicated from shared patterns of pathway inference, we ran consensus clustering based on the 3370 most varying features (i.e. with variances within the highest quartile). KIRC cases that were suspected chromophobes (syn1768397) were excluded, yielding a sample size of 3941. Consensus clustering was implemented with the ConsensusClusterPlus package in R, after slight modifications to disable graphical outputs. Specifically, median-centered IPLs were used to compute the squared Euclidean distance between samples; and this metric was used as the input to the ConsensusClusterPlus algorithm. Hierarchical clustering using the Ward's minimum variance method (i.e. ward inner linkage option) with 80% subsampling was performed over 1000 iterations; and the final consensus matrix was clustered using average linkage. The number of clusters (k) was selected based on the area under the empirical cumulative distribution function (CDF) curve at a point where the area reaches an appropriate maximum and further separation provides minimal relative change.

Figure S2C1 shows the cancer type composition of each PARADIGM cluster. Interestingly, half of the PARADIGM clusters are comprised primarily of a single cancer type (e.g. clusters 1-LAML, 2-BRCA, 3-BRCA, 8-LUAD, 11-KIRC and 12-GBM), while the other half (4-COAD/READ, 5-OV/UCEC, 6-Mixed Lung/HNSC/BLCA, 7-Mixed LUSC/HNSC/BLCA, 9-Mixed BRCA/Lung and 10-Mixed UCEC/BLCA) are comprised of cancers arising from multiple tissues. Of note, the mixed tissue cluster 5 was comprised of serious ovarian and predominantly serous or mixed histology endometrial cancers (41/51 with 7 missing), and as expected, showed characteristic low p53 and high E2/ERA activity.

Figure S2C3 showed the distribution of cluster membership within each cancer type. Once again, some cancer types (e.g GBM, LAML, KIRC, COAD, OV and READ) were predominantly placed within a single cluster, while others (BLCA, BRCA, HNSC, LUAD, LUSC and UCEC) were split into multiple PARADIGM clusters. Of note, 128 of the 137 basal breast cancers were placed into cluster 9, separate from the remaining breast cancer cases, in line with previous observations of the distinctness of the basal PAM50 subtype from other breast cancers. Interestingly, despite previous TCGA reports of the molecular similarities of basal breast cancers with serous ovarian cancers, the basal cases appear to cluster with a subset of lung squamous and lung adenocarcinomas in this Pan-Cancer 12 analysis. This mixed 9-BRCA/Lung cluster appears to be characterized by low p53 and ER signaling, and high FOXM1 and MYC/Max activity.

The majority of LUSC and HNSC were intermixed and placed into PARADIGM clusters 6 or 7. Interestingly, PARADIGM cluster 6 also contains 15 LUAD cases, all of which were of the squamoid (solid-enriched) LUAD subtype. This accounts for ~19% of the solid-enriched LUAD cases; and suggests that this particular subtype of LUAD may be a necessary but not sufficient condition for assignment into the mixed squamous PARADIGM cluster. Common pathway inference patterns associated these two squamous PARADIGM clusters include high activity of HIF1A/ARNT, Jun/Fos, FOXM1 and TAp63, and low ER signaling. Potential pathway differences between these two clusters may include higher immune activation and lower MYC/Max signaling in the 6-Mixed Lung/HNSC/BLCA cluster.

2.5 Integrative subtypes derived from different methods are highly concordant

We evaluated concordance between 5 integrative subtyping methods: COCA subtypes and SuperClusters derived with and without consideration of the mutation subtypes and PARADIGM subtypes. Unsupervised clustering (based on Pearson correlation) of binary integrative subtype membership of the 3143 cases with assignments from all five integrative subtyping methods reveals that the integrative subtypes are highly concordant. Of note, integrative subtype which includes mutation data closely correlates with their corresponding subtype as determined without taking into account mutation subtyping. See Figure S2D.

There were, however, a few interesting exceptions to the concordance. Super Cluster 7 is comprised of a mix of LUAD, BLCA, BRCA, basal-BRCA as well as a subset of HNSC and LUSC cases. Both other subtype solutions further subset this large group. This may in part be attributed to the lower cluster number (k=9) selected in the Super Cluster solution; but differences in methodology likely played a role as well. OV samples form a monolithic cluster in the two integrative solutions which use all data types (Integrative Subtype 9 and Super Cluster 1). In contrast, PARADIGM places ~15% of UCEC samples together with the OV cases in PARADIGM cluster 5. Other differences between the PARADIGM subtypes and the other integrative solutions included the further sub-division of the predominantly luminal breast (Integrative Subtype 3 and Super Cluster 5) and predominantly squamous (Integrative Subtype 2 and Super Cluster 6) subtypes into two groups (PARADIGM clusters 2 and 3, and PARADIGM clusters 6 and 7, respectively). These latter observations are consistent with results from the comparison with the single platform mRNA subtypes (as described in the previous section).

Section 3: Clinical Importance of the COCA Subtypes

3.1 Survival analysis

Overall survival was calculated for samples using information from the enrollment and follow-up forms available at the Data Portal and downloaded on 6/17/2013. Kaplan-Meier survival plots were performed with the package survival in R. A log-rank test was used to assess significance (Figures S3A-F). For TP53 mutation signature [25] and the proliferation signature [26], samples were rank ordered and divided into thirds (high, medium, and low) for analysis. Mutation status for TP53 and PIK3CA were from taken from synapse (syn1710680).

For the tumor types that did not have a one to one relationship with their COCA groups, we wanted to determine if COCA group status added additional information to a prognostic model for overall survival. We limited analysis to BLCA, BRCA, COAD, HNSC, LUAD, LUSC, and READ in COCA clusters COCA1 – LUAD-enriched, COCA2-Squamous, COCA3-BRCA/Luminal, COCA4-BRCA/Basal-like, COCA7-COAD/READ, and COCA8-BLCA. We estimated the change in log likelihood statistic as we added COCA group and Tumor Type information to the clinical variables tumor size, metastasis status, node status, and age at diagnosis to a Cox Proportional Hazards model. COCA group was still able to add significant information to Tumor Type in predicting overall survival. A chi square test was used to assess statistical significance of the change in log likelihood statistic as we added additional features to the model. See Figure 1E.

3.2 Gene program correlations of COCA clusters with outcome

We used univariate and multivariate Cox Proportional Hazards modeling to assess associations between Gene Program expression levels and patient survival in the PanCancer 12 data compendium. Gene Programs, 22 non-redundant features extracted from ~7000 gene expression signatures, are described in Section 10 and listed in Table S4B. Multivariate models were adjusted for CoCa subtype (model: Surv~Gene_Program_i+subtype) and for CoCa subtype and cancer stage (model: Surv~Gene_Program_i+subtype+Tpath+Npath+Mpath), with a threshold of significance for association of Wald p-value<0.05 after Benjamini-Hochberg multiple testing correction. In models adjusted for stage, Tpath= tumor size, Npath=lymph node status, and Mpath=distant metastases.

Univariate and multivariate subset analyses were performed in each individual PanCancer subtype as well, to identify Gene Programs that specifically associated with survival in each of the 12 subtypes. In addition, the above univariate, multivariate, and subset analyses were performed on the Drug Pathways listed in Table S4D. The survival and multitest software packages in Bioconductor (www.bioconductor.org) were used for these analyses.

3.2.1 Gene programs and drug pathways associated with outcome across 12 cancers

In univariate analysis, the majority of gene programs and drug pathways were associated with patient outcome. However, since we have shown that many of these pathways – say neural signaling for GBM and estrogen signaling for luminal BRCA – are characteristic of a single cancer subtype, and since different cancer types have different natural histories and levels of aggression, one might be concerned that these univariate associations are merely a proxy for cancer type or subtype. To investigate, we constructed a multivariate Cox model with CoCa subtype as a covariate to see which if any gene programs remain associated with patient outcome above and beyond subtype. We found basal signaling (GP_17), squamous differentiation/development (GP6), proliferation/cell cycle, and estrogen signaling (GP_7) retained significance in the multivariate model, as well as immune checkpoint pathways CTLA4 and PD1, among others (see Table S4E). Another natural question concerns cancer stage as a potential confounder. In a multivariate model including integrated subtype *and* cancer stage, estrogen signaling (GP7; pro-survival) retained significance, even after adjustment for multiple testing. Considering uncorrected p-values, in the multivariate model including stage and subtype, Basal signaling (GP17) also associated with decreased OS, and immune pathways PD1_signaling and CTLA4_pathway with increased OS, though these associations did not reach significance after BH correction. Interestingly, GP10_Fatty acid oxidation – a novel gene program related to alternative metabolism of cancer - gains significant association in this model. Overall, these results indicate that despite uneven clinical information and follow-up across cancer types, the data are sufficiently good to produce 'expected' associations such as the link between highly proliferative cancers with basal or squamous phenotypes and poor patient outcomes. These 'expected' results then add confidence to other results consistent with hypotheses currently under active investigation, such as the emerging role of immune checkpoint pathways PD1 and CTLA4 in survival and treatment of multiple cancers, and, we hope, to the potential importance of some of the more novel gene programs such as GP10 (fatty acid oxidation) and GP3 (tumor suppressing miRNA targets) described in Table S4E.

3.2.2. Common signaling pathways associate with patient outcome in KIRC and Luminal BRCA

Subset analysis aims to identify features associated with outcome in a specific cancer without regard to the other cancers in the analysis, thereby making post-hoc comparisons possible. Here we present some surprising similarities between kidney cancer (KIRC) and luminal breast cancer, a highlight of our subset analysis results. Cox proportional hazards survival analysis applied to KIRC identified gene programs reflecting estrogen signaling (GP7), fatty acid oxidation (GP10), and tumor suppressing miRNA targets (GP3) as significantly associated with patient outcome in kidney cancer, along with expression of the PTEN/MTOR signaling axis. *Many of these same pathways were significantly (and canonically) associated with outcome in luminal BRCA* (see Table S4F and Figure S5D with GP7 dichotomized at the median).

Interestingly, in KIRC, high expression levels of these pathways were associated with increased OS, whereas in luminal BRCA, with the exception of GP7_Estrogen, upregulation was associated with decreased OS. These results suggest similar or mirror image paths to malignancy for the two cancers, and, importantly, treatment strategies effective in one type of cancer – such as endocrine inhibitors - might be considered for the other. They also further highlight GP3_Tumor.suppressing.miRNA.targets as a novel gene program that is associated with outcome in multiple cancer types, and likely druggable.

Section 4: Genomic Determinants of the Integrated Subtypes

4.1 MuSiC significantly mutated gene (SMG) analysis on COCA subtypes

For each cluster of cluster (COCA) subtype (syn1889916), we used the significantly mutated gene (SMG) test in Mutational Significance in Cancer (MuSiC) [27] to identify significant genes based on the curated, strictly-filtered, list of Pan-Cancer mutations (syn1729383).

The SMG test in MuSiC assigns mutations to seven categories (AT transition, AT transversion, CG transition, CG transversion, CpG transition, CpG transversion, and indel) and uses statistical tests based on convolution, the hypergeometric distribution (Fisher's test), and likelihood to combine the category-specific binomials to obtain overall P-values. All P-values were combined using the methods described in Dees *et al*. [27]. We collected the SMG test results for each COCA subtype and filtered the SMGs based on tumor tissue gene expression levels and literature curation.

The final list of top 40 SMGs for each subtype is in Table S2A. Due to the small number of samples, in subtypes 11 and 12, no genes were determined to be significantly mutated. The full list of the frequencies within each of the COCA subtypes for all of the 127 SMGs was plotted (Figure S4D) and is available as an excel table (Table S2B).

4.2 Predicted driver genes for COCA subtypes

Putative mutational drivers of the Pan-Cancer data set were retrieved by combining the results of multiple methods aimed to identify genes exhibiting different signals of positive selection. On detail, five methods were used: MuSiC [27] which selects genes mutated more frequently than the expected by the mutation background model; OncodriveFM [19] which identifies genes with a bias towards accumulation of mutations with high functional impact; OncodriveCLUST [28] which detects genes with significantly clustered mutations; ActiveDriver [29] which pinpoints genes whose mutations occur predominantly in protein active sites; and MutSig, whose results take into account several analyses, as the detection of frequently mutated genes according to a background model corrected by several covariates involved in the mutation rate [30]. This resulted in a list of 291 high-confidence putative drivers available at syn1962006.

4.3 HotNet2 sub-networks analysis

We used an updated version of the HotNet algorithm, here referred to as HotNet2, to search for sub-networks of frequently mutated proteins in a large protein-protein interaction network in each of the 11 largest integrated subtypes (1 through 10 and 13). HotNet2 identifies significantly mutated sub-networks using an insulated heat diffusion model that accounts for both the "heat score" on each protein in the given network, and the local topology of the protein's interactions. We downloaded non-synonymous single nucleotide variants (SNVs), indels, and splice-site mutations from Synapse (syn1710680). Copy number aberrations (CNAs) were obtained from GISTIC 2 output via Firehose. This resulted in a final dataset of somatic aberrations in 6989 genes from 3105 samples.

We calculated a heat score for each protein as the number of samples with a non-silent SNV, indel, splice-site mutation, or copy number aberration in the corresponding gene, restricting to genes mutated in ≥ 2% of samples in each subtype. We ran the HotNet2 heat-diffusion approach to identify subtype-specific sub-networks using these scores on the HINT [31] physical protein-protein interaction network that includes 28,497 interactions among 8,269 human proteins.

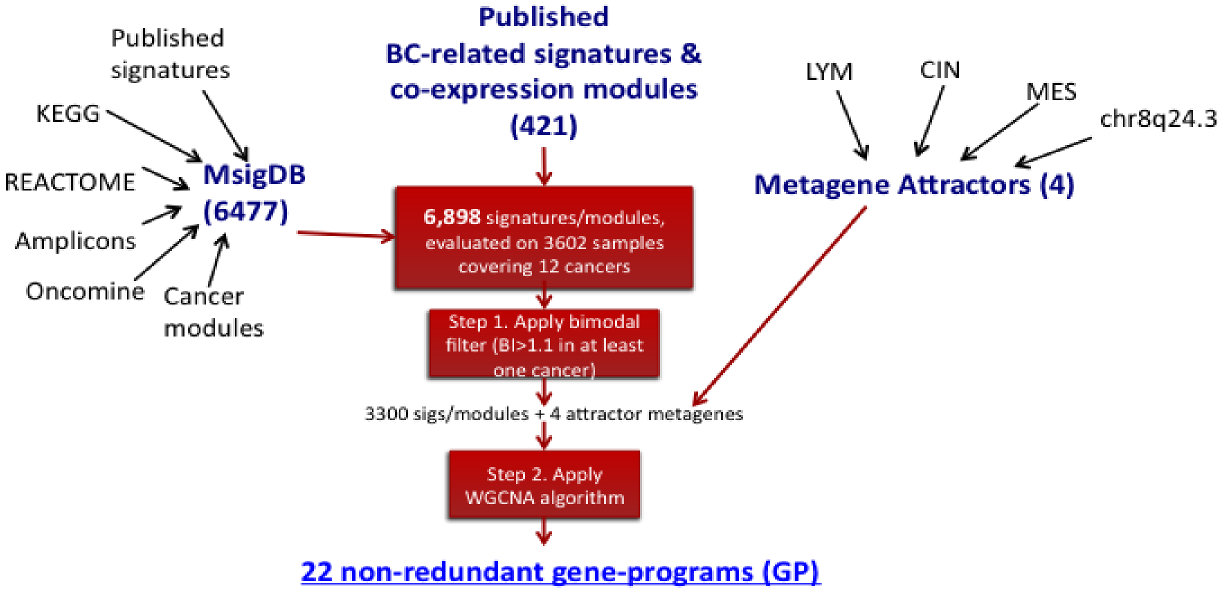
HotNet2 identified four sub-networks from the 546 Squamous subtype samples (Table S3B). The largest and most mutated of these sub-networks includes many well-known cancer genes and tumor suppressors including TP53, CDKN2A, and PTEN, and is mutated in 91.7% of the Squamous samples. The other three sub-networks are distinctive of the Squamous subtype (Figure 4D). The second most mutated sub-network (59.9%) includes oxidative stress response genes NFE2L2, CUL3, and KEAP1, as well as well-known cancer genes CCNE1, FBXW7, and NOTCH1. NFE2L2 is a transcription factor involved in oxidative stress, and is ubiquitinated by CUL3 and sequestered in the cytoplasm by KEAP1. Mutations in NFE2L2, CUL3, and KEAP1 were recently shown to be important for squamous lung cancer. The third most mutated sub-network (37.1% of Squamous samples) includes the ASCOM complex (MLL2 and MLL3) and the putative ASCOM-interacting protein KDM6A. These proteins are involved in histone modifications that promote transcription.

Section 5: Expression-based Determinants of the Integrated Subtypes

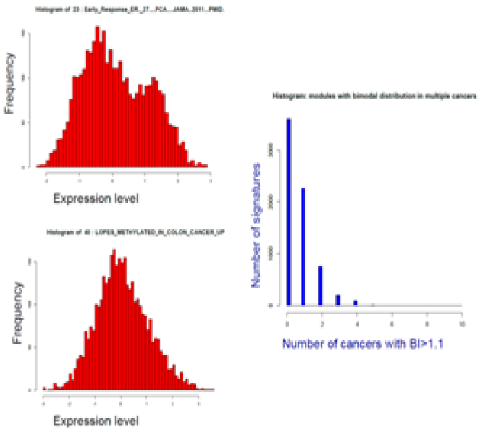
Thousands of molecular signatures have been published representing pathways and co-expression modules purported to associate with cancer biology or patient outcome. These signatures were developed for diverse cancers, using diverse data sources. We wondered whether this compendium of signatures, modules, and pathways could be reduced to a set of non-redundant gene expression programs, and hypothesized that such 'gene-programs' may be useful for investigating cancer heterogeneity beyond tissue of origin.

5.1 Application of WCGNA and a bimodality index filter reduces ~7000 published gene expression signatures, modules and pathways into 22 non-redundant gene-programs

Using multiple approaches, we assembled a compendium of 6,898 gene expression signatures/co-expression module features, which we evaluated on normalized, batch adjusted RNAseq expression profiles from 3602 TCGA samples representing 12 cancer types (Pan-Cancer 12). We obtained 6477 of these features, a collection of published prognostic signatures from multiple cancer types, KEGG or REACTOME pathways, Oncomine pathways, amplicons, and cancer co-expression modules, from the database MsigDB [32] (http://www.broadinstitute.org/gsea/msigdb). An additional 421 signatures and co-expression modules capturing prognostic and predictive signals; intrinsic subtype; and pathway, mutation, micro-environment and amplicon features in breast cancer, were obtained from prior breast cancer publications and scored as previously described [33] either by median expression level (283 modules from 30 publications), (first) principle component analysis (113 signatures from 45 publications), Pearson correlation coefficient to a centroid (21 signatures from 11 publications), or model/Euclidean distance/mean calculations (5 signatures from 2 publications) as appropriate. Pubmed IDs for the 91 publications are available on SAGE Bionetwork's Synapse web portal (http://sagebase.org/synapse/) as files syn1960751, syn1960752, and syn1960753. The dominant attractor metagenes CIN (chromosomal instability), LYM (lymphatic), MES (mesenchymal), and chr8q24.3, evaluated as previously described[33], were included in the analysis as well. The complete collection of constituent signature/module scores for the Pan-Cancer 12 dataset are available on Synapse as well (syn1960760), along with the associated gene lists, and method of evaluation (syn1960751,2,3,8).


Published signatures
KEGG
REACTOME
Amplicons
Oncomine
Cancer modules
MsigDB (6477)
Published BC-related signatures & co-expression modules (421)
LYM
CIN
MES
chr8q24.3
Metagene Attractors (4)
6,898 signatures/modules, evaluated on 3602 samples covering 12 cancers
Step 1. Apply bimodal filter (BI>1.1 in at least one cancer)
3300 sigs/modules + 4 attractor metagenes
Step 2. Apply WGCNA algorithm
22 non-redundant gene-programs (GP)

Our methodology for reducing the signature compendium to a smaller set of non-redundant features is illustrated above. First, we filtered the MsigDB and other published signatures and gene sets on the basis of bimodality, considering only those with a bimodal index (BI) greater than 1.1 in at least one cancer type for further analysis[34] (see inset figure; syn1960763). This filtering, based on the hypothesis that signatures/modules that are bi- or multi-modal within individual or multiple cancers are capturing heterogeneity that might be shared across cancers, was implemented using the R package ClassDiscovery [34].The 3300 bimodal signatures/modules, plus the 4 attractor metagenes, were then aggregated into coherent clusters (gene-programs) using Weighted Gene Correlation Network Analysis (R package WGCNA[35]) with settings power=6 and minModuleSize=8. This algorithm uses topological criteria such as scale-free connectivity (small world networks) and topological overlap metrics to identify highly correlated sets of features by 1) constructing a co-expression network relating features using correlation coefficients, and 2) using 'soft' thresholds and dynamic tree cutting to define gene-programs[35]. WGCNA grouped all but 141 of the remaining 3304 signatures/modules into 22 coherent, non-redundant gene-programs. Gene-program scores for the TCGA samples, evaluated as the first principal component of each group of constituent signatures, are available on Synapse (syn1809701), as are the signature/module members of each gene-program and the correlation coefficients between each signature and each gene-program (syn1960647).



5.2 Gene-programs represent many of the hallmarks of cancer, with most constituent signatures clustering into proliferation and immune modules

The 22 non-redundant gene-programs encompass many of the hallmarks of cancer, including sustained proliferative signaling (GP1_Proliferation); immune system infiltration (GP2-Immune-cells) or disregulation (GP11_Interferon); altered metabolism yielding resistance to stress (GP13_Hypoxia/glycolosis and GP10_Fatty acid oxidation); evading apoptosis (GP21_Anti-apoptosis); epithelial to mesenchymal transition (GP4_MES/ECM); co-opted development and differentiation programs (GP6_Squamous differentiation/development and GP8_FOXO/stemness/ALK), and adhesion and cell-cell communication programs (GP9_Cell-cell adhesion enriched for claudins and E-cadherins, GP14_Plasma membrane cell-cell signaling correlated to contactin (CNTN1), and GP18_Vesicle/EPR membrane coat) contributing to invasion and metastasis; and self-sufficiency in growth signals (GP7_Estrogen signaling, GP15_EGF signaling, and GP5_Myc/TERT), among others (see Table S4B).

Interestingly, more than a third (1158/3304) of the signatures cluster together in the proliferation gene-program. In addition to cell cycle and proliferation pathways, this gene-program includes prognostic signatures for lymphoma; bladder, lung, breast, and ovarian cancer; mesothelioma, adrenocortical cancer, leukemia, Ewings sarcoma, melanoma, and other cancers. This convergence of prognostic signatures from many different cancers into a single gene-program, that also includes canonical proliferation pathways, suggests that many – or perhaps most - prognostic signatures are sensing a common proliferative signal associated with cancer aggression.

Also interesting is the presence of two dominant immune signals, showing a clear divergence in immune signaling between T cell/B cell adaptive and innate immune activation (the large gene-program GP2_Immune T cell/B cell), representing 504/3304 (15%) constituent signatures) and interferon responsive signaling (GP11_Interferon). We hypothesize that GP2 reflects lymphocytic infiltrate, which is associated with a positive outcome and chemo-response in many cancers. A Spearman correlation coefficient of 0.967 between GP2 and the expression ESTIMATE model of immune infiltrate (syn1901530) supports this hypothesis. GP11-Immune-IFN may be more tumor cell intrinsic and related to inflammation. See Table S4B for a complete list of gene-programs annotated by theme, number of constituent signatures, and representative constituent signatures.

5.3 Gene-program and signature-compendium based annotation of Pan-Cancer integrated subtypes

Predictive accuracy and visualization: To assess whether the variability in gene-program expression patterns was sufficiently coherent with the integrated subtype classification to render a gene-program based annotation reasonable, we built classifiers to predict integrative subtypes from the gene-program scores. To evaluate the accuracy of these classifiers, we performed standard five-fold cross-validation. For each fold, we trained a Linear Discriminant Analysis (LDA) predictor [36] which learns a linear projection that maximizes separation between class centroids, while minimizing the scatter of samples within each class. Test data was classified according to the closest centroid under this projection, and the results aggregated into a single confusion matrix. With an average accuracy of 90%, we concluded that despite the low dimensionality of the gene-programs compared to that of the data used to derive the integrated subtypes, gene-program expression patterns may reasonably be used for annotation (see Figure S5A).

Pan-Cancer subtype interpretation: Encouraged by the high predictive accuracy of the model, we used the gene-programs, supplemented by the full set of signatures, as an aid in characterizing and interpreting the 11 COCA subtypes. The unique pattern of features characterizing each COCA subtype, identified through application of t-tests comparing samples in each subtype to those in all other subtypes (with Benjamini Hochberg [37] multiple testing adjustment), is summarized in Table S4C, with more detailed results in syn2227907. An overview of the expression-based determinants of the integrated subtypes can be seen in the clustered heatmap of Figure 3 of the main text. As expected, the gene programs *GP6-squamous differentiation/development*, *GP13-neural signaling* and *GP20-TAL-1-leukemia/erythropoiesis* were the most highly expressed in the C2-Squamous-like, C10-GBM and C13-LAML subtypes, respectively. As well, *GP7_Estrogen signaling* was highest in the C3-BRCA/luminal cases, whereas *GP17_basal signaling* had its highest levels in the C4-BRCA/basal cases. Intriguingly, the *GP7_Estrogen signaling* program, along with *fatty acid oxidation* (*GP10*), *tumor suppressing miRNA targets* (*GP3*) and the *PTEN/MTOR signaling program*, were found to be significantly associated with patient outcome in kidney cancer using a Cox proportional hazards survival analysis (Figure S5D; Table S4F). Higher levels of the *Estrogen signaling program* (GP7) were associated with better prognosis. Consistent with the higher frequency of elevated HER2 protein levels in bladder, colorectal and serous endometrial cancers [38], the *HER2-amplified* gene signature appeared elevated in the C8-BLCA, C7-COAD/READ and C6-UCEC subtypes, as well as the C3-BRCA/luminal subtype. The latter contained the HER2-positive breast cancers. Surprisingly, the *GP5-MYC targets/TERT* gene program and *MYC-amplified* signature, although significantly correlated, appeared to show different expression patterns across the COCA subtypes. *GP5-MYC targets/TERT* was significantly upregulated in the C2-Squamous-like, but not the C9-OV subtype, even though ovarian cases show the highest expression of the *MYC amplified* signature. Other notable enrichments include *GP2_Immune-cell upregulation* in KIRC, C2-Squamous-like and C1- LUAD-enriched subtypes, and *high GP11-Interferon* but *low GP2 immune expression* in C9-OV, with C4-BRCA/Basal expressing high levels of the *CTLA4 immune checkpoint pathway*. Those signatures may reflect varying amounts of immune infiltrate in the tumors, as has been estimated by DNA methylation-based analysis of the Pan-Cancer-12 dataset (Figure S5E). In any case, these immune cell-associated gene programs may be pertinent to emerging treatment strategies based on immune modulation.

Additional highlights of our findings are as follows:

- Basal breast cancer is its own subtype, different from other breast cancers and from all other cancers and subtypes. Basal breast cancer is characterized by high basal and proliferation signatures; low estrogen signaling; loss of function PTEN alterations and P53 mutations; high PI3K mutation and signaling; BRCA1-like gene expression; downregulated TGFB1 targets; high HDAC expression; high expression of DNA mismatch repair pathways; high MYC signaling; and high expression of signatures prognostic for risk of relapse in lung and brain but low expression for risk of relapse in bone signatures. These cancers have an interesting immune profile, in that they over-express immune checkpoint pathways CTLA4, and interferon controlled genes and pathways, but do not seem to over-express T cell/B cell signatures to the same degree.
- Squamous lung and squamous head and neck cancers are more similar to one another than to adenocarcinomas. Squamous lung and HN cancers share many common features, including upregulated mesenchymal pathways; high expression levels of squamous development and differentiation signatures and pathways including WNT; and upregulated protein kinase signaling (MAPKs), and EGF signaling. Squamous cancers appear to be hypoxic with strongly upregulated glycolysis pathways, and highly proliferative with highly expressed DNA repair pathways including RB and ATR. HDACs are highly expressed in these cancers relative to other subtypes, as are signatures reflecting cell-cell adhesion using claudins and Ecadherins and plasma membrane signaling. These cancers also express strong immune signals manifesting as high expression levels of B cell and T cell immune signatures, as well as the immune check-points PD1 and CTLA4, proteins targeted by drugs currently being tested in clinical trials that have shown promise in melanoma. Relative to lung cancers in the C1-LUAD-enriched subtype, squamous cancer expression profiles suggest poor survival, low differentiation levels, and a causative agent other than smoking.
- Basal breast cancer and squamous cancers of the lung or head and neck have many features in common. A principle component analysis of gene-programs and drug pathways shows that basal breast cancers are more similar to squamous cancers than to any other integrated subtype, including luminal breast cancers. Squamous cancers have extremely high 'basal' signaling as determined from basal breast cancer subtype signatures, nearly as high as that seen in basal breast cancer (Figure S5B). In addition, both subtypes share common upregulation of P53 mutation signatures, proliferation and DNA repair pathways, MYC signaling and VEGF signaling. They also both over-express immune check-point pathways like PD1 and CTLA4, as well as interferon responsive immune genes and pathways.
- Bladder is the most heterogeneous cancer, with a subset clustering with adenocarcinomas of the lung. Bladder cancers fall into three main subtypes, a bladder-only subtype characterized by high claudin and Ecadherin mediated cell-cell adhesion; high MYC targets/TERT, high ERBB2, ZEB1, and APC target expression; and low ALK and immune signaling (62%); a sarcoma subtype dominated by HNSC and LUSC (26%); and the subtype dominated by lung adenocarcinoma, characterized by high T cell and B cell immune activation, high cell-cell adhesion, low expression of DNA damage/repair genes, and low mutation and CN aberration rates (8%). This latter subgroup expresses signatures associated with smoking-related cancers.

5.4 Comparing selected gene-programs and drug pathways to genomic data

We used copy number alteration (CNA) data from SNP6.0 arrays analyzed using the GISTIC algorithm (see Supplemental Section 1 for details) to validate the gene programs that appeared to be dominated by chromosomal amplicon signatures. We found that the expression level of the mRNA based gene program GP22_16q22-24 was highly correlated to copy number estimates of genes from this region of the genome (Kendall's tau~0.63; linear regression p-value<2E-16). A similar correspondence was observed between expression of GP19_1Q and copy number estimates of genes from chr1q (Kendall's tau 0.58, 0.7; p-value<2E-16).

Though the gene program GP5_MYC targets/TERT is most highly correlated with MYC expression signatures (see Table S4B), rather than amplicons on chromosome 8q.24 where MYC is located, we were interested in assessing correspondence between GP5 and MYC CNA data. As shown in Figure S5C, GP5_MYC targets/TERT is modestly but significantly correlated to MYC copy number aberrations (panel B; Kendall's tau =0.154, p-value < 2.2E-16). To investigate to what extent GP5 is representing the 53 MYC-related expression signatures in the compendium, we generated the clustered heatmap of the pairwise correlation coefficients between signatures shown in panel S5.D1 Since GP5_MYC targets/TERT (S5.C1, top arrow) clusters with the majority of MYC signatures, and the MYC CNA data (S5.C1, bottom arrow) does not cluster closely with any of the expression signature blocks, we concluded that GP5 is a fair representation of MYC network expression as captured in the signature compendium, and that downstream effects must dominate. The signature that best corresponds to MYC amplification in the Pan-Cancer 12 data set is chr8q24 (S5.C3) from MsigDB, the chromosomal neighborhood of MYC. As MYC is an important if complicated oncogene, we added this signature to the 'drug target' pathways in Table S4D, with the acknowledgement that should a MYC inhibitor be developed it is not clear whether the expression signatures or the amplicon would be more likely to function as response biomarkers.

Section 6: Multi-platform Determinants of the Integrated Subtypes

6.1 miRNAs characteristic of COCA clusters

We analyzed miRNA-seq data for 10 COCA clusters. GBM did not have miRNA sequencing, and so cluster 10 of mainly GBM samples was removed from the analysis along with the small clusters 11 and 12. For each COCA cluster, we generated two expression matrices: one for samples in that cluster, and one for the samples in the other 9 clusters. The expression matrix counts were normalized to millions of reads aligned to miRNAs for each sample. We used a Wilcoxon rank-sum test in R v3.0.2 to find miRNAs that were differentially expressed between each COCA cluster and the remaining samples. The p-value and estimate of the median difference between the cluster samples and the remaining samples were reported for each miRNA. A Benjamini Hochberg correction was run across the multiple miRNA tests within each cluster analysis. To highlight significantly differentially expressed miRNAs that may be biologically functional, we filtered the results by removing miRNAs that had an adjusted p-value of greater than 0.05 or a mean normalized read count below 100 in both expression matrices [39]. Genes that were potentially targeted by miRNAs were extracted from a database of functionally validated miRNA targets, miRTarBase v4.5 [40]. See Data File S2 [syn2468318 (tab A)].

6.2 Gene-level SCNA frequencies of COCA clusters

For each COCA cluster, significantly higher frequencies of copy number alterations in all genes were determined by one tailed Fisher's exact tests using thresholded gene level copy number from GISTIC 2.0 analysis. In these tests, the occurrence of amplifications or deletions in each COCA was compared against those in all other COCAs combined. COCAs were tested either for increased frequencies of high level alterations (estimated changes of 2 or more copies) or for higher occurrences of alterations of any level. See Data File S2 [syn2468318 (tabs C-E)].

6.3 DNA methylation characteristics of COCA clusters

We performed a probe-design dependent platform normalization to remove systematic platform bias, and generated a merged dataset on 25,978 probes shared by the HM27 and HM450 platforms. In brief, the probes were split by the number of CpGs in the probe sequence, as that differentially affects technical variation in the DNA methylation measurement, and LOESS normalization is done within each probe group. We also used two sets of technical replicates (TCGA-07-0227 and TCGA-AV-A03D) which were repeatedly measured as internal controls (99 and 74 times respectively) across platform and batch to monitor residual batch and platform variations. Any probes with a >0.1 standard deviation in beta values in either set was removed. In addition, any DNA methylation probes that overlap with single nucleotide polymorphisms (SNPs) and repeats, or map to sex chromosomes were masked from the analysis. We also exclude any probe that failed in more than one fifth of the samples in any COCA cluster. We focus on the probes that are unmethylated in all normal tissue types (median beta value <0.2) only, to remove pre-existing tissue-specific DNA methylation from the analysis. 11,648 probes remain after the above filters. 9,380 of the 11,648 probes show hypermethylation (defined as beta value >0.3) in at least 10% of the samples in at least one COCA group and are retained in the analysis.

For each of these 9,380 probes, we compare the hypermethlation (beta value >0.3) frequency in tumors in each COCA cluster to the rest with Fisher's Exact test. 99.5% of the probes map to +/- 1500bp of gene promoters, and the corresponding gene(s) to the probes are listed in the resulting p value file. If a probe is mapped to the +/- 1500 bp region of more than one genes multiple entries are created differing only in gene names. See Data File S2 [syn2468318 (tab F)].

6.4 Differentially expressed proteins between COCA clusters

We used the LIMMA [41] package in R to determine differentially expressed proteins between the COCA clusters. We performed a one vs. all other clusters comparison for each COCA cluster where the number of samples with protein data was ≥ 10, which was true for 10 clusters (cluster numbers 1-10). The results for all 181 proteins are shown in Data File S2 [syn2468318 (tab G)]. The table has proteins in rows, and the following columns for each cluster obtained from LIMMA:

1. Log fold-change: Indicating the $\log_2$ fold-change between the mean value of the protein in the given cluster vs. the mean value of the protein in all other clusters.
2. Average expression: The mean expression of the protein in the given cluster.
3. *t*-statistic: The statistic used for computing *P*-values for each protein, based on a variant of the *t*-test [41].
4. *P*-value: The *P*-value resulting from the statistical test.
5. Adjusted *P*-value: The *P*-value after adjusting for multiple hypothesis testing.
6. *B* value: The B value is obtained from LIMMA fitting and represents $\log_2$ odds ratio for the differentially expressed protein.

We computed pathway scores for each sample using 9 different pathways based on protein data. Proteins belonging to the 9 pathways were identified using prior knowledge and literature. Their expression values were then added together or subtracted, depending on whether the protein activated or inhibited the pathway, respectively. The resulting sum provided a relative pathway score, which could be compared across samples to determine which samples had activated or suppressed pathway relative to the other samples. The protein memberships for each pathway are given in Data File S2 [syn2468318 (tab H)] and the pathway scores for each sample are given in Data File S2 [syn2468318 (tab I)].

We then grouped the samples by the COCA clusters and obtained box plots for each cluster, for each pathway. The results are shown in Figure S5F. Some interesting observations (with potential explanations) are that the apoptosis pathway is activated in the squamous-like cluster 2, but suppressed in GBM cluster 10, suggesting that GBM tumor cells are dying by necrosis rather than apoptosis. There is high DNA damage in endometrial cluster 6, and ovarian cluster 9, probably due to MSI-high samples in endometrial [5] and high copy number variations in ovarian cancer [42]. Breast basal and ovarian samples also tend to have mutations in BRCA1/2 genes, resulting in high DNA damage. RTK pathway is activated in GBM, likely due to EGFRv3 and its downstream consequences. PI3K/Akt pathway is activated in GBM and endometrial cancers, probably due to mutations in the PTEN gene.

We performed LIMMA analysis again, this time using the pathway scores as input instead of individual proteins to determine differentially expressed pathways between the COCA clusters. The same methods were used as mentioned before for individual proteins, and the pathway results are presented in Data File S2 [syn2468318 (tab J)].

6.5 Uniquely differential pathways by COCA subtype

We used a compendium of pathways (syn1741407) assembled from 7 publicly available databases - GO (http://www.geneontology.org/), KEGG (http://www.genome.jp/kegg/pathway.html), NCI (http://pid.nci.nih.gov/) ,Omim (http://www.ncbi.nlm.nih.gov/omim), Reactome (http://www.reactome.org/), BioCarta (http://www.biocarta.com/genes/index.asp), and GenMapp (http://www.genmapp.org). We limited the compendium to pathways containing between 5 and 200 genes, for a total of 4923 entries. We looked for distinctive pathways in each of the 11 major COCA subtypes by performing Gene Set Enrichment Analysis [32] in five data types– mRNA, copy number (CN), mutation, methylation, and miRNA. We used an individual enrichment gene ranking for each platform-COCA subtype pair. The ranking was computed by comparing the samples in the respective COCA cluster to the samples in all other clusters by the following methods:

1. mRNA – the absolute value of SAM d-statistic (syn2347490, [43]) (14686 genes covered)
2. mutation – negative log of a Fisher exact test p-value computed on mutation status for a set of 291 mutation driver genes defined in [3].
3. CNV – negative log of a Fisher exact test p-value computed on frequency of CN alterations (24162 genes covered).
4. methylation - negative log of a Fisher exact test p-value computed on hypermethylation status for 9,380 DNA probes (7850 genes covered).
5. miRNA – log of the product of a Wilcoxon rank sum test p-value (based on miRNA expression) and Spearman correlation for each miRNA-mRNA gene pair. Only genes with correlation < - 0.3 and rank sum test p-value <0.05 for the corresponding miRNA were considered (1879 genes meet the criteria across all COCA subtypes).

The enriched pathways were filtered based on 3 criteria- number of genes in a pathway that were actually ranked, enrichment score, and False Discovery Rate. To avoid enrichment driven by high singleton gene scores, we only considered pathways with at least 5 ranked genes. The only exception was mutations, where the threshold was relaxed to 3 due to the small number of ranked genes. Enrichment scores (ES) were filtered to those exceeding 0.5 - general processes known to be impacted by cancer (i.e. regulation of cell cycle) were observed to produce ES values at or above that threshold. Finally, we computed FDR by adjusting the pathway permutation-based p-values to account for multiple hypotheses testing [37]. Gene scores based on negative log p-values (all platforms except mRNA) tended to produce some genes with identical or very similar scores, particularly for p-values towards the two ends of the spectrum. As a consequence, permutation-based pathway p-values/FDR were somewhat biased towards higher values, resulting in an increased rate of false negatives. To counter that, we set a more lenient FDR cutoff of 0.2 for all platforms except mRNA. No such effects were observed with the SAM-based mRNA gene scores – the mRNA FDR cutoff was therefore set at 0.1.

As we are mainly interested in pathways that uniquely identify a particular COCA subtype, we discarded all pathways that were significantly enriched (within a particular data type) in more than one COCA cluster. The remaining pathways were compared across all five platforms – we identified 11 pathways that were supported by more than one data platform in the same COCA cluster (Figure S6A, purple and dark blue colors). To address the redundancy among uniquely identified pathways, we performed the following steps:

1. We ranked the pathways across all subtypes and all platforms by ES and FDR, with the 11 multiple-support pathways placed at the top of the ranking.
2. We went down the ranked list and compared each pathway in turn against all pathways above it – if the overlap between the queried pathway and any of the higher ranked ones was higher than 25% of either pathway's size, the lower ranked pathway was discarded.

This produced a final list of 261 unique differential pathways (Figure S6A). Of those, 103 were enriched in mRNA, 103 in CNV, 68 in mutation, 25 in methylation, and 1 in miRNA.

In addition to the GSEA analysis, we used the PARADIGM algorithm [24] to identify pathway-based biomarkers differentially activated within each COCA subtype relative to all other subtypes. A complete list of enriched pathways and ‘regulatory hubs’ with more than 15 downstream targets reflective of differentially activated pathways and biomarkers for each COCA subtype is provided (Table S5A). Activated pathway characteristics found by enrichment and sub-network analyses are summarized in Table S4A. Consistent with the gene program analysis, the C1-LUAD-enriched, C2-Squamous-like and C5-KIRC subtypes show higher immune pathway activation. Also consistent was the elevated HIF1A signaling observed in C2-Squamous-like and C5-KIRC subtypes, in which the *GP12-Hypoxia/Glycolysis* gene program expression also appeared highest (Figure 3, Table S4A). We found MYC pathway activation in the C2-Squamous-like, C4-BRCA/Basal, C7-COAD/READ and C9-OV subtypes, which also show high GP19- and/or MYC-amplified signature expression, likely due to ch8q24 amplification. MYC signaling also appeared elevated in the C13-AML subtype despite the relatively low frequency of 8q amplification, suggesting an alternative, amplification-independent mechanism for upregulation of MYC in those tumors [44]. Additional pathway characteristics activated across multiple integrative subtypes include higher p38 pathway activation in the C1-LUAD-enriched, C2-Squamous-like, C7-COAD/READ and C13-LAML subtypes, and relative p63-activation in the C2-Squamous-like and C7-BLCA subtypes that showed elevated GP9-squamous differentiation/development gene program expression.

The *GP7-estrogen signaling* gene program and relative *FOXA1/ER signaling* activation were observed only in the C3-BRCA/luminal subtype, not the C4-BRCA/basal, C6-UCEC or C9-OV subtypes (the latter two of which are often ER+). That finding is consistent with the Pan-Cancer-12 proteomic analysis, in which the downstream components of the hormone signaling pathway were elevated only in the luminal/ER+ BRCA cases [38]. In addition to FOXA1/ER signaling, other subtype-specific pathways showing relative activation included: *PI3K signaling* in C3-BRCA/luminal (with PIK3CA mutation as the top SMG in the subtype; frequency 40%), PLK1 in C4-BRCA/basal, caspase-associated pathways in C13-LAML and AKT signaling in C10-GBM cases (for which the PI3K/AKT protein pathway scores were the highest).

6.6 Elastic Net analysis to identify integrative determinants of COCA subtypes

We performed a supervised training and testing classification analysis with the goals of 1) determining the predictive power of a COCA classifier and 2) identifying the distinctive features for each COCA class selected by the classifier. We used the subset of the COCA sample set (n=1,851) that had mRNA, miRNA, protein/RPPA, DNA copy number, and mutation data. This excludes GBM which does not have RPPA data and AML which does not have microRNA data. A combined matrix for all data was created with a total of 8,070 features. Gene expression modules and signatures (n=6,920) were calculated for the mRNA expression data [33] and also included the 24 Gene Programs discussed in the main text. For copy number, we selected 428copy number aberrations (CNA) that had been previously identified to be altered in all cancer [45] and in breast cancer [46] and defined their start and stop genomic coordinates. Then using the segmented data from the copy number data, we assigned the continuous values from the segments to the CNAs. Non-silent mutation calls for the 127 significantly mutated genes were scored as 1 for mutant or 0 for wild type. RPPA protein data (n=131) and miRNA expression data (n=440) were used as continuous variables. Data was row and column standardized prior to analysis. The data was split into a 70% “training” data set and a 30% test set balanced for COCA subtypes using the R package ‘sampling’ [47]. The training set was then used for an Elastic Net predictor approach using the R package ‘glmnet’ [48]. Using the 70% training data set, we trained an Elastic Net predictor for each COCA subtype versus all other samples, using 10-fold cross validation to optimize algorithm specific parameters (alpha: The Elastic Net mixing parameter; lambda: Regularization parameter), the optimized algorithm was used to fit a model to the entire training set, this model building procedure resulting in 9 different predictors. We applied these predictors to the 30% test set samples (not a pure test set since these were used in the original COCA classification scheme), where there was 95% accuracy in assigning COCA subtype. Next we examined the features selected by Elastic Net algorithm for each COCA classifier, which are provided in Data File S3 (syn2486685).Using the features identified in each analysis, we combined them to make a large Elastic Net feature heatmap that visually displays the features characteristic of each COCA group (Figure S6B).

Section 7: Convergence of Squamous-like Subtype and Features Common to Squamous, Breast Basal, and Ovarian Subtypes

7.1 Pathway biomarkers of integrative subtypes

IPLs differentially activated between the mixed squamous Integrative Subtype 2 and the other subtypes were identified using the t-test and Wilcoxon Rank Sum test with Benjamini-Hochberg(BH) FDR correction. Only features deemed significant (FDR corrected $p<0.05$) by both tests and with an absolute difference in group means > 0.05 were selected. Differentially activated IPLs were then filtered by connectivity within the pathway structure, such that only interconnected features (at least 1 interaction of any kind) was retained. Pathways enriched among these selected features were assessed using the EASE score with BH FDR correction; and sub-networks were constructed to identify regulatory hubs based on interconnectivity (i.e. >15 outgoing regulatory edges) and visualized using Cytoscape. Similar analysis was performed to identify pathway biomarkers of the basal Integrative Subtype 4 as well as the other 9 Integrative Subtypes. A complete list of significant enriched pathways and regulatory hubs for each of the 11 Integrative Subtypes are provided in Table S5A.

7.1.1 Squamous enriched Integrative Subtype 2

Pathway enriched analysis identified 72 pathways as significantly enriched among the interconnected differentially activated IPLs in the squamous enriched Integrative Subtype 2 (Table S5A). A notable hub among these enriched pathways centers on the p63 isoform, Np63, known to initiate epidermal K5/K14 keratinization and epithelial differentiation under mesenchymal induction, but also linked to basal stem/progenitor cell function in other organs (e.g. breast, urogenital tract) as well as the development of squamous and adenosquamous carcinomas where it is never mutated but often overexpressed by genomic (3q27-29) amplification. Interestingly, 9 of these enriched pathways are linked with immune signaling, suggesting this may be an important biological process distinguishing the mixed-squamous Integrative Subtype. Consistent with this implication, subnetwork analysis independently identified 7 regulatory hubs (of 41 total) relating to immune function that were significantly activated within this subtype (Figure S7B). Also of note, 9 pathways relating to cell adhesion, motility and invasion (e.g. integrin, Rac, Rho and BMP signaling) were also identified as significantly enriched; and subnetwork analysis identified RAC and Rho GTPases as activated regulatory hubs in this squamous enriched subtype (Figure S7B). This is consistent with previous report of the ability of squamous carcinomas to collectively migrate and invade during the metastatic process. Other pathway features distinguishing this squamous enriched cluster includes: activation of proliferation-related pathways (e.g. MYC/Max, FOXM1, PLK1, MYB), AP-1 and HIF1A signaling.

7.1.2 Basal enriched Integrative Subtype 4

Pathway enrichment analysis identified 104 pathways as significantly enriched among the interconnected differentially activated IPLs in the basal enriched Integrative Subtype 4 (Table S5A). Surprisingly, despite the large number of enriched pathways, subnetwork analysis was able to identify only 26 differentially activated regulatory hubs with >15 downstream targets. Of note, p53 and ATM signaling appear significantly repressed in this subtype, in keeping with PARADIGM-inferred impairment of DNA damage response. Consistent with the reported roles of p63 in both mesenchymal-epithelial induction and the maintenance and replenishment of basal stem/progenitor cells in skin and other organs, both Np63 and TAp63 isoforms of p63, along with links to basal keratins (K5/K14) and integrin-4 were significantly activated in this Integrative Subtype 4.

Altered immune signaling was also implicated as a distinguishing feature of this basal enriched subtype. Interestingly, 3 of the 4 immune related hubs (RELA/p50, STAT1 and STAT5 complexes) were activated, while the STAT6 complex was repressed, suggesting that pro-inflammatory immune responses, but not lymphocyte-mediated adaptive immunity, are significantly activated within this subtype. Other pathway features of note include the activation of proliferation related pathways (through MYC/Max, FOXM1, PLK1, MYB, E2F1 hubs), an inferred response to hypoxia, as well as down-regulation of FOXA1/ER signaling. These latter findings are consistent with previous TCGA analyses comparing basal versus luminal breast cancer subtypes.

7.2 Pathway Commonalities between Integrative Subtypes

7.2.1 Squamous and basal integrative subtypes

Our pathway biomarker analyses implicate activation of the dNp63 and TAp63 regulatory hubs as distinguishing features of both the squamous and the basal integrative subtypes. Since similar hub activation may integrate multiple upstream signals to yield diverse downstream consequences, we compared the squamous and basal subtypes for pathway biomarkers directly linked to these two regulatory hubs.

Figure 5A shows the Cytoscape plot of the regulatory subnetwork and pathway biomarker differences linking the dNp63 and TAp63 hubs for squamous and/or basal subtypes. 39 of these features, including the dNp63 and TAp63 hubs, were commonly activated (or repressed) in both squamous and basal subtypes relative to the other subtypes, and were thus truely shared pathway features. With 1 exception, all of these shared biomarkers showed more significant differential activation in the squamous vs. all others comparison. Interestingly, 12 of the p63-linked pathway biomarkers activated within the squamous subtype were repressed in the basal subtype (e.g. ITGA3), highlighting potential p63 signaling differences despite having commonly activated p63 hubs. As well, 21 p63-linked pathway biomarkers were unique to the squamous subtype, while only 2 were unique to the basal subtype. Altogether, these results suggest that despite significant p63 hub activation in both the squamous and basal subtypes, p63 signaling is differentially and more strongly activated within the squamous integrative subtype.

Looking beyond p63 and across the entire SuperPathway, in order to systematically identify pathway commonalities between the squamous and basal integrative subtypes, we performed a linear fit of the squamous differential score (i.e. difference in mean activity between the squamous subtype and all other cases) onto the basal differential score (i.e. difference in mean activity between the basal subtype and all other cases) using all ~13.3K varying IPLs. A basalness score (b) was computed as the orthogonal projection of the squamous differential score onto the basal differential score. Features with basalness scores at least 2 standard deviations from the mean were defined as significant and were filtered by consistent activation (or repression) in both subtypes relative to the other cases. These significant common pathway features were then filtered by connectivity within the SuperPathway structure such that only interconnected features (through 1 interaction of any kind) was retained. Pathway enrichment analysis among these interconnected features was performed using the EASE score with BH-FDR correction; and the largest regulatory subnetworks within the SuperPathway structure linking these features were identified and visualized in Cytoscape (Figure S7C1).

Our analysis identified 777 features as comprising a significant basalness score. Among these, 685 were commonly activated (or repressed) within the squamous and basal subtype relative to all other cases, where 365 were interconnected through at least 1 interaction within the SuperPathway structure. 71 pathways were significantly enriched among these 365 interconnected features, including pathways related to cell cycle, p63 signaling and immune function (Table S5B). Subnetwork analysis identified 13 regulatory hubs with more than 5 direct downstream targets; and consistent with the pathway enrichment analysis, the majority of these hubs converge upon regulation of proliferation (FOXM1, MYB, PLK1, MYC/Max, E2F/DP), immune function (STAT1 and STAT5) as well as p63 signaling (TAp63g and dNp63a tetramer complexes). Interestingly, XBP1, a known estrogen-inducible gene, appears to be significantly activated within these subtypes despite their having lower FOXA1 and E2/ER-alpha hub activity, suggesting alternative means of XBP1 activation potentially due to its chaperone function and role in endorecticulum stress responses. Lower inferred DNA damage response activity may be an additional pathway commonality between these two subtypes, as it was present as a hub with reduced activity and emerged by enrichment analysis along with p53, RB and ATR related pathways.

7.2.2 Basal and ovarian integrative subtypes

To systematically identify pathway commonalities between the ovarian and basal integrative subtypes, we computed the basalness score as the orthogonal projection of the ovarian differential score (i.e. difference in mean activity between the ovarian and all other subtypes) onto the basal differential score. Features with significant basalness score were identified and filtered as described above; and pathway enrichment and subnetwork analyses of the selected features were performed.

In this ovarian-basal subtype comparison, 781 features with significant basalness score were identified. Of these, 665 were commonly activated (or repressed) in the ovarian and basal subtypes relative to the other cancers. This yielded 461 interconnected features. Although pathway enrichment identified 80 significantly enriched pathways, the subnetwork analysis was able to confirm the presence of only 11 regulatory hubs with more than 5 direct downstream targets (Table S5B and Figure S7C2). Altogether, our analysis suggested that key pathway commonalities between the basal and ovarian subtypes include activation of proliferation pathways (e.g. MYC/Max, FOXM1, E2F/DP and PLK1) and impairment of p53 signaling. Interestingly, despite previous TCGA reports of activated HIF1A/ARNT signaling in ovarian and basal BRCA, the activity of this hub appears repressed relative to the other subtypes, most likely due to the presence of VHL-mutated KIRC cases with high HIF1A activity in this Pan-Cancer 12 analysis. Once again, XBP1 activity appears to be activated, despite the lack of FOXA1/ER signaling activation within these two integrative subtypes.

7.2.3 Squamous, basal and ovarian integrative subtypes

A comparison of the squamous-basal and the ovarian-basal pathway commonalities revealed 342 pathway biomarkers that were shared between all three of these subtypes. 197 of these common pathway biomarkers were interconnected within the SuperPathway structure. Pathway enrichment analysis (Table S5B) and subnetwork analysis (Figure S7C3) independently implicate activation of proliferation related pathways (e.g. E2F/DP, MYC/Max, FOXM1 and PLK1) as well as XBP1 signaling as pathway commonalities between all three integrative subtypes.

Lower inferred DNA damage response activity may also be an additional pathway commonality. However, we note that p53 was not identified as a common pathway hub, despite the high prevalence of p53 mutations within these three integrative subtypes. This is due to the high inferred p53 activity within the squamous subtype in contrast to the low inferred activity in the basal and ovarian subtypes (relative to all other cases). Potential reasons for this discrepancy include: p53 activity assessed on a relative scale with respect to other cancer types which have greater impairment in p53 function, inferred activity not taking into account the actual mutation status, and compensatory activation of p53-linked downstream targets through an independent squamous subtype specific mechanism such as p63.

7.3 PARADIGM-Shift analysis of p53 mutations within integrative subtypes

To assess the pathway impact of p53 mutations, we applied the PARADIGM-SHIFT [49] algorithm to the Pan-Cancer 12 dataset. p53 mutations were subdivided into two classes: truncating (nonsense or frameshift) and non-truncating (missense or inframe). Pathway impact was predicted by comparing the calculated P-Shifts (i.e. discrepancies in downstream and upstream signal) for samples with TP53 truncating mutations and samples with wild-type TP53. Overall, PARADIGM-SHIFT predicts a significant (p = 0.006) loss-of-function (LOF) shift for TP53 truncating mutations. A CircleMap representation of the p53 PARADIGM-SHIFT scores across the Pan-Cancer 12 Integrative Subtype is shown in Figure S7D1.

The CircleMap demonstrates several key observations. Samples with truncating mutations show markedly lower expression levels of TP53 (3rd ring) in comparison to the non-truncating cases, a phenomenon which may be attributed to nonsense-mediated decay (NMD). Despite the overall predicted LOF associated with p53 truncating mutations (relative to wild type), the P-Shift scores (outer ring) among p53 truncating mutations appears highly variable across the different integrative subtypes. This variation appears to be predominantly driven by differences in the downstream signal (5th ring), rather than upstream signal which appears uniformly negative (blue in the 4th ring). The integrative subtype ordering of the CircleMap was selected by identifying the clusters most consistent with the LOF signal among samples with TP53 truncating mutations. This was done by performing a GSEA on the samples, first ranking the samples with truncating TP53 mutations by P-Shift scores then looking for enrichment of the different integrative subtypes on either tail. The table below shows the ranking of integrative subtypes most consistent with LOF signal among TP53 truncating mutants by GSEA.

CofC	GSEA Score	Adjusted P-Value
10	-0.72	0.003
9	-0.66	< 0.001
4	-0.47	0.03
6	-0.46	0.24
3	-0.43	0.11
7	-0.41	0.25
8	-0.40	0.50
1	0.31	0.0021
2	0.37	< 0.0013
13	0.79	0.0024
5	0.92	0.081

GBM, OV and basal clusters show the strongest signal for LOF, with predominantly negative P-Shift scores (blue in outer ring) tracking with the truncating mutants. Within these clusters, there is also a prediction for LOF in the non-truncating samples, which were not used in the PARADIGM-SHIFT comparison. On the other end of the spectrum are the LAML and KIRC clusters with predominantly positive P-Shift scores (red in outer ring) with few p53 mutations indicating less predicted LOF. Interestingly, the Squamous subtype also falls on the end of the spectrum with less predicted LOF, but also harbor a large number of mutations in TP53.

7.3.1 Pathway impact of p53 mutations within the squamous, basal and ovarian subtypes

To best complement our pathway commonalities analysis, we focused upon the squamous, basal, and OV subtypes which all have high frequency of p53 mutations (Figure S7D2). Consistent with those findings, p53 and p63 (both isoforms) inferred activities are significantly higher in the squamous cluster compared to the OV and basal clusters. We note that although there is less predicted LOF associated with p53 truncating mutations in the squamous subtype, p53 inferred activity is lower within the mutated cases (relative to wild-type). There is also significantly less LOH (i.e. samples with both a TP53 mutation as well as a loss of a copy represented by the blue in the GISTIC ring) among the squamous subtype compared to basal and OV subtypes (p = < 0.001). In addition, there is a trend for higher activation of p63 (both isoforms) within the p53 squamous mutants relative to wild type squamous cases. Both these independent observations support the hypothesis that TP53 downstream targets are more activated in the squamous versus basal and OV and that this difference may be due to lower frequency of LOH and squamous-specific TP63 compensation for TP53 loss.

Indeed, when we compared the expression and activity of the common targets of TP63 and TP53 as annotated in our overall PARADIGM pathway in the OV, basal, and squamous clusters (Figure S7D3), most of the targets show higher levels of expression and inferred activity in the squamous samples. However, this elevated expression and activity appears irrespective of p53 mutation status, making it difficult to distinguish compensatory p63 signaling in the cases of TP53 loss from general squamous-specific p63 activation.

To further evaluate the differences in p53 signaling within the OV, basal breast cancer and squamous-like subtypes, we searched for p53 related gene expression signatures within the compendium of ~7000 published signatures assembled for the gene program analysis (syn1960760). 33 signatures containing the term p53 in its name were found. Of note, two signatures were annotated as representative of p53 and p63 common targets. Unsupervised clustering (Pearson correlation, average linkage) of the 748 samples with known p53 mutation status within the ovarian, basal breast, and squamous-like subtypes based on the expression of these 33 p53-related gene signatures were performed using the heatmap.plus package in R and shown in Figure 5D.

Section 8: Divergence of the Bladder Cancer Subtype

8.1 Gene program and pathway characteristics distinguishing C2-Squamous-like from C8-BLCA bladder cancers

Gene programs and PARADIGM IPLs differentially activated between the bladder cancers in the C2-Squamous-like subtype (n=20) and in the C8-BLCA subtype (n=74) were identified using the t-test and Wilcoxon Rank Sum test with Benjamini-Hochberg(BH) FDR correction. Only features deemed significant (FDR corrected p<0.05) by both tests were selected.

Unsupervised clustering (Pearson correlation, average linkage) of the 11 gene programs showing significant differential expression between the C2 and C8 bladder cases was performed; and a heatmap where samples were first ordered by subtype, then unsupervised clustering (Pearson correlation, average linkage) based on gene program expression within subtype, was generated using the heatmap.plus R package and shown in Figure 6D.

For the PARADIGM pathway marker analysis, differentially activated IPLs were then filtered by connectivity within the pathway structure, such that only interconnected features (at least 1 interaction of any kind) was retained. Pathways enriched among these selected features were assessed using the EASE score with BH FDR correction; and sub-networks connected through regulatory nodes with >5 outgoing downstream targets was constructed and visualized using Cytoscape. Figure S8A shown below contained the complete interconnected regulatory subnetwork; and a zoomed-in view of the immune-related subnetwork highlighted by the dotted red box was displayed in Figure 6E.

Overall, results from these analyses are highly consistent, and points to the C2 bladder cases as sharing the squamous-like subtype characteristics of higher proliferation (GP1, FOXM1), higher hypoxia signaling (GP12, HIF1A/ARNT), higher squamous differentiation signals (GP6, p63), higher MAPK signaling (GP16, p38), lower estrogen signaling (GP7, FOXA1), and higher immune activation (GP2, GP11, STATs and IL/JAK complexes), relative to the C8-BLCA subtype.

8.2. Analysis of Bladder samples across Pan-Cancer COCA subtypes

Bladder subtypes were one of the most diverse tumor types with samples (n=120) clustering in 7 of the 13 COCA subtypes. The majority of the samples fell into three main COCA groups including 10 in COCA1 – LUAD-enriched, 31 in COCA2 – Squamous, and 74 in COCA8 – Bladder. Correlation with histology measurements show that the bladder samples within COCA2- Squamous cluster did indeed have some evidence of squamous features though many had less than 50% squamous differentiation. As one of the most diverse tumor sets in this pan-cancer set, we looked at survival differences between the three main groups the bladder samples in our pan-cancer set. The squamous and LUAD-like bladder samples had a significantly worse overall survival compare to the Bladder-enriched group (Figure S8B).

The classifications from the Pan-Cancer analysis were compared to subtypes derived from the TCGA Bladder Analysis Working Group [50]. For 112 samples in common, there was high correlation between the classifications. In particular, Bladder classes I and II were highly similar with COCA8 while Bladder III was enriched for the COCA2 (Table S6A). While bladder cancer samples with >50% squamous features were excluded from the TCGA Bladder Analysis Working Group, Bladder class III was enriched for bladder samples with some squamous features present.

References

1. Gaujoux, R. and C. Seoighe, *A flexible R package for nonnegative matrix factorization.* BMC Bioinformatics, 2010. **11**: p. 367.
2. Cancer Genome Atlas, N., *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.
3. Tamborero, D., et al., *Comprehensive identification of mutational cancer driver genes across 12 tumor types.* Sci Rep, 2013. **3**: p. 2650.
4. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
5. Cancer Genome Atlas Research, N., et al., *Integrated genomic characterization of endometrial carcinoma.* Nature, 2013. **497**(7447): p. 67-73.
6. Cancer Genome Atlas Research, N., *Comprehensive genomic characterization of squamous cell lung cancers.* Nature, 2012. **489**(7417): p. 519-25.
7. Cancer Genome Atlas Research, N., *Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia.* N Engl J Med, 2013. **368**(22): p. 2059-74.
8. Wang, K., et al., *MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.* Nucleic Acids Res, 2010. **38**(18): p. e178.
9. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.
10. Wilkerson, M.D. and D.N. Hayes, *ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking.* Bioinformatics, 2010. **26**(12): p. 1572-3.
11. Liu, Y., Hayes, D. N., Nobel, A. & Marron, J. S. , *Statistical Significance of Clustering for High Dimensional Low Sample Size Data.* Journal of the American Statistical Association, 2008. **103**(483): p. 1281-1293.
12. PJ, R., *A Graphical Aid to the Interpretation and Validation of Cluster Analysis.* Journal of Computational and Applied Mathematics, 1987. **20**: p. 53-65.
13. de Hoon, M.J., et al., *Open source clustering software.* Bioinformatics, 2004. **20**(9): p. 1453-4.
14. Saldanha, A.J., *Java Treeview--extensible visualization of microarray data.* Bioinformatics, 2004. **20**(17): p. 3246-8.
15. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration.* Nat Genet, 2013. **45**(10): p. 1134-1140.
16. Tibes, R., et al., *Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells.* Mol Cancer Ther, 2006. **5**(10): p. 2512-21.
17. Hu, J., et al., *Non-parametric quantification of protein lysate arrays.* Bioinformatics, 2007. **23**(15): p. 1986-94.
18. Hennessy, B.T., et al., *Pharmacodynamic markers of perifosine efficacy.* Clin Cancer Res, 2007. **13**(24): p. 7421-31.
19. Gonzalez-Perez, A. and N. Lopez-Bigas, *Functional impact bias reveals cancer drivers.* Nucleic Acids Res, 2012. **40**(21): p. e169.
20. Coombes K, N.S., Joy C, et al, in *SuperCurve: R package*. 2011.
21. Liang, J., et al., *The energy sensing LKB1-AMPK pathway regulates p27(kip1) phosphorylation mediating the decision to enter autophagy or apoptosis.* Nat Cell Biol, 2007. **9**(2): p. 218-24.
22. Hennessy, B.T., 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, 2010. **6**(4): p. 129-51.
23. Neeley, E.S., et al., *Variable slope normalization of reverse phase protein arrays.* Bioinformatics, 2009. **25**(11): p. 1384-9.
24. Vaske, C.J., et al., *Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.* Bioinformatics, 2010. **26**(12): p. i237-45.
25. Troester, M.A., et al., *Gene expression patterns associated with p53 status in breast cancer.* BMC Cancer, 2006. **6**: p. 276.
26. Parker, J.S., et al., *Supervised risk predictor of breast cancer based on intrinsic subtypes.* J Clin Oncol, 2009. **27**(8): p. 1160-7.

27. Dees, N.D., et al., *MuSiC: identifying mutational significance in cancer genomes.* Genome Res, 2012. **22**(8): p. 1589-98.
28. Tamborero, D., A. Gonzalez-Perez, and N. Lopez-Bigas, *OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes.* Bioinformatics, 2013. **29**(18): p. 2238-44.
29. Reimand, J. and G.D. Bader, *Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers.* Mol Syst Biol, 2013. **9**: p. 637.
30. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.
31. Das, J. and H. Yu, *HINT: High-quality protein interactomes and their applications in understanding human disease.* BMC Syst. Biol., 2012. **6**.
32. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
33. Fan, C., et al., *Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures.* BMC Med Genomics, 2011. **4**: p. 3.
34. Wang, J., et al., *The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data.* Cancer Inform, 2009. **7**: p. 199-216.
35. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
36. Hastie, T., Tibshirani, R. & Friedman, J, *The Elements of Statistical Learning*. 2001: Springer.
37. Benjamini, Y.H., Y, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* J Roy Stat Soc B Met, 1995. **57**: p. 289-300
38. Akbani, R., et al., *A pan-cancer proteomic perspective on The Cancer Genome Atlas.* Nat Commun, 2014. **5**: p. 3887.
39. Mullokandov, G., et al., *High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries.* Nat Methods, 2012. **9**(8): p. 840-6.
40. Hsu, S.D., et al., *miRTarBase: a database curates experimentally validated microRNA-target interactions.* Nucleic Acids Res, 2011. **39**(Database issue): p. D163-9.
41. Smyth, G., *Linear Models for Microarray Data (LIMMA).* 2014.
42. Cancer Genome Atlas Research, N., *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-15.
43. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
44. Salvatori, B., et al., *Critical Role of c-Myc in Acute Myeloid Leukemia Involving Direct Regulation of miR-26a and Histone Methyltransferase EZH2.* Genes Cancer, 2011. **2**(5): p. 585-92.
45. Beroukhim, R., et al., *The landscape of somatic copy-number alteration across human cancers.* Nature, 2010. **463**(7283): p. 899-905.
46. Weigman, V.J., et al., *Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival.* Breast Cancer Res Treat, 2012. **133**(3): p. 865-80.
47. Guillaume Chauvet, Y.T., *A fast algorithm for balanced sampling.* 2006. **21**(1): p. 53-62.
48. Jerome Friedman, T.H., Robert Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 2010. **33**(1): p. 1-22.
49. Ng, S., et al., *PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis.* Bioinformatics, 2012. **28**(18): p. i640-i646.
50. Cancer Genome Atlas Research, N., *Comprehensive molecular characterization of urothelial bladder carcinoma.* Nature, 2014. **507**(7492): p. 315-22.

TCGA Glossary

Cancer Types (Pan-Cancer 12)

AML **–** Acute Myeloid Leukemia

BLCA – Bladder Urothelial Carcinoma

BRCA – Breast Invasive Carcinoma

COAD – Colon Adenocarcinoma

GBM – Glioblastoma Multiforme

HNSC – Head and Neck Squamous Cell Carcinoma

KIRC – Kidney Renal Clear Cell Carcinoma

LUAD – Lung Adenocarcinoma

LUSC – Lung Squamous Cell Carcinoma

OV – Ovarian Serous Cystadenocarcinoma

READ – Rectal Adenocarcinoma

UCEC – Uterine Corpus Endometrial Carcinoma

Center Types

BCR – Biospecimen Core Resource Center

DCC – Data Coordinating Center

CGHub – Cancer Genomics Hub

GCC – Genome Characterization Center

GDAC – Genome Data Analysis Center

GSC – Genome Sequencing Center

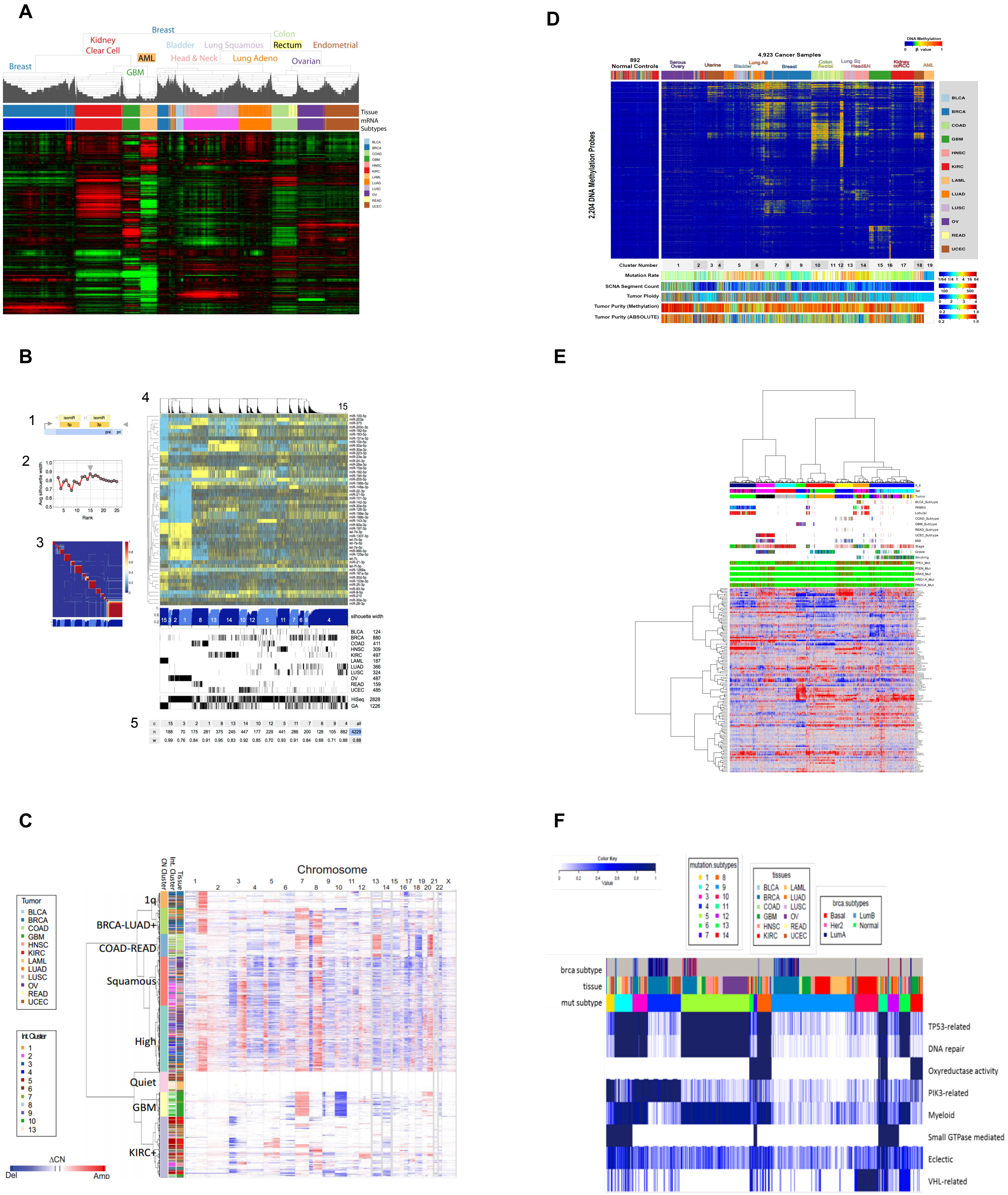
TSS – Tissue Source Site

Technology Platform Names

A list of sequencing and array technology platforms and can be found here: https://tcga-data.nci.nih.gov/tcga/tcgaPlatformDesign.jsp

Supplemental Figure

Figure S1. Single-platform Clustering Analysis, Related to Figure 1



Supplemental Figure

Figure S2. Integrated Platform Analysis, Related to Figure 1

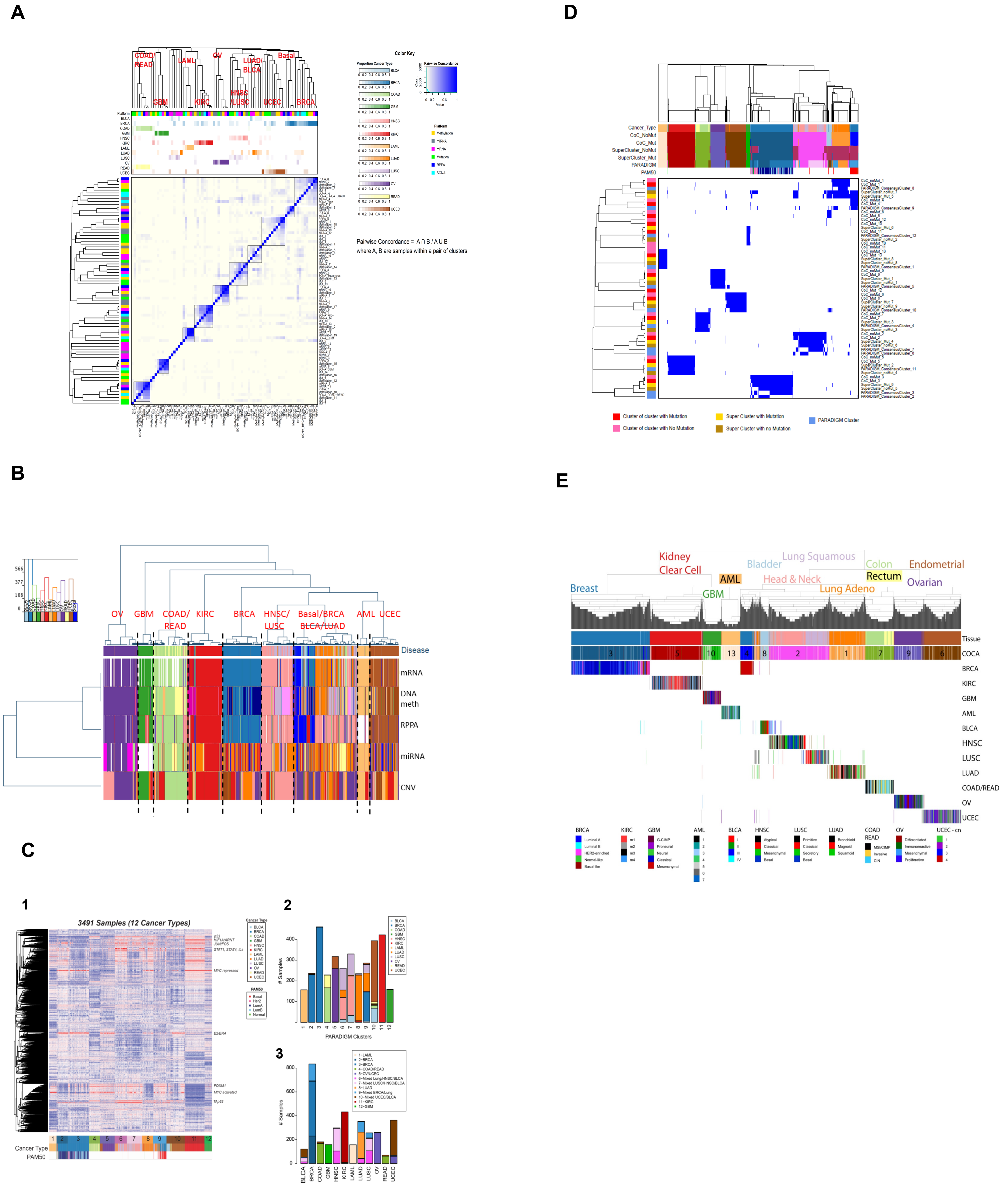
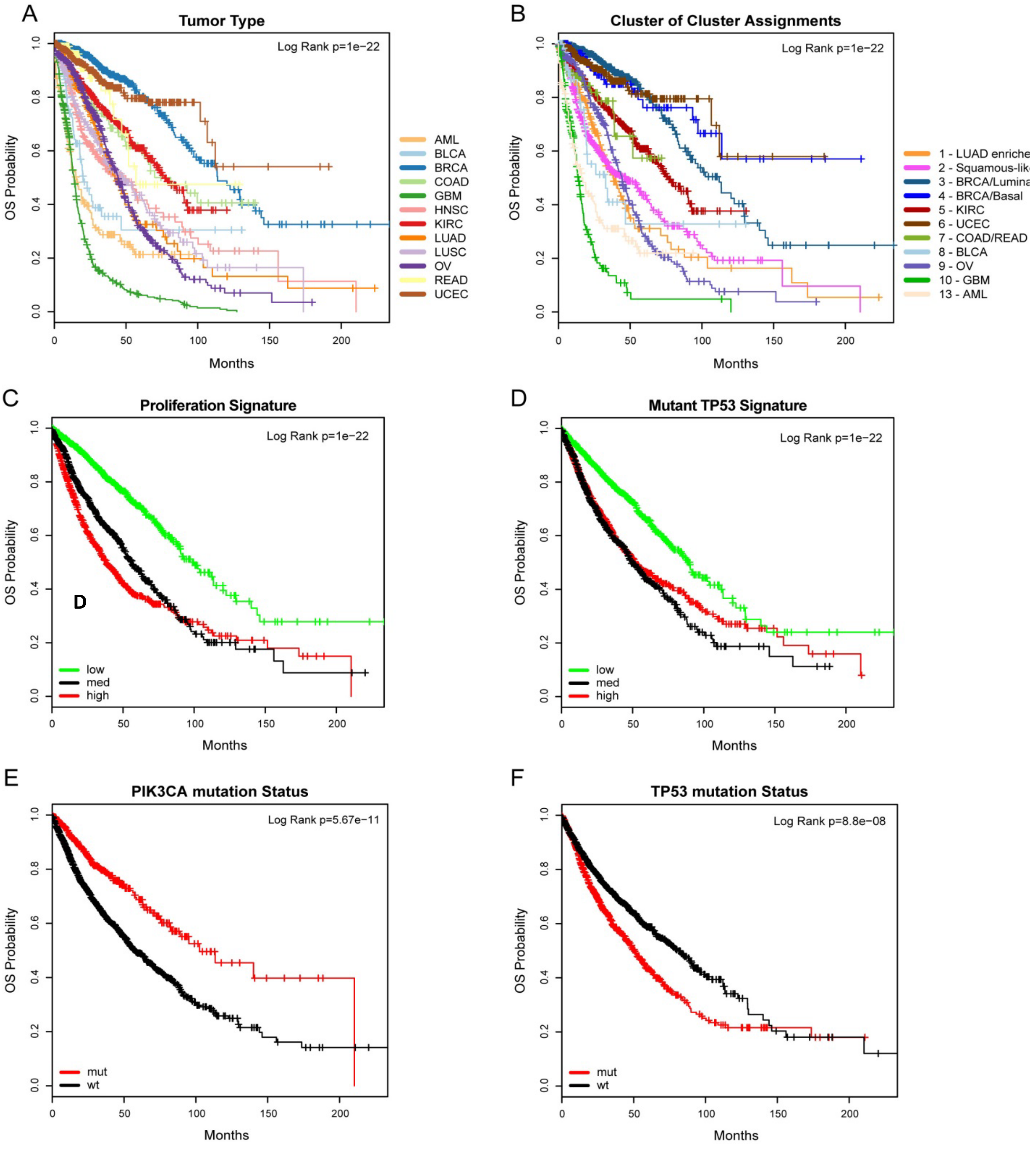
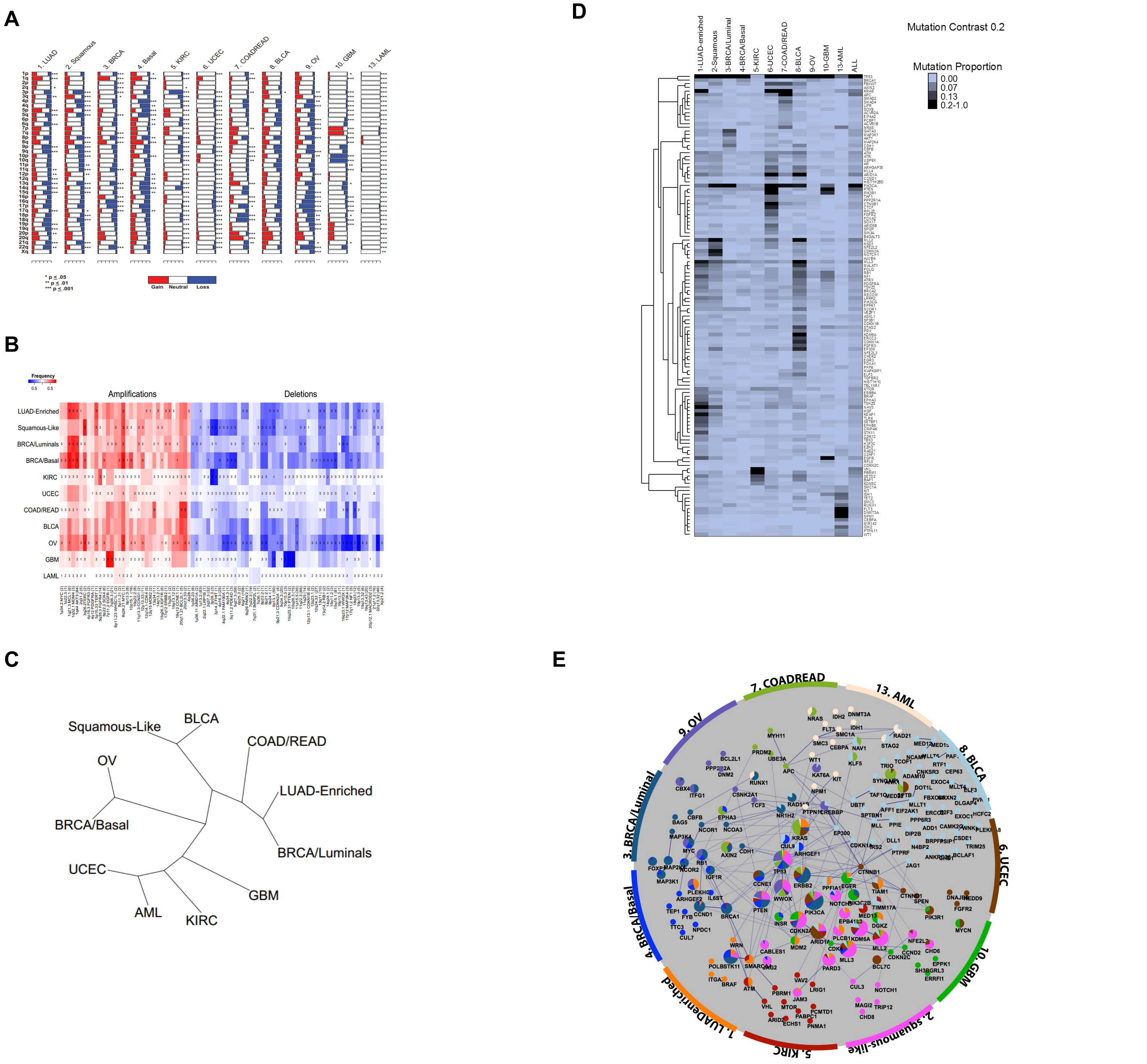


Figure S3. Clinical Importance of the COCA Subtypes, Related to Figure 1



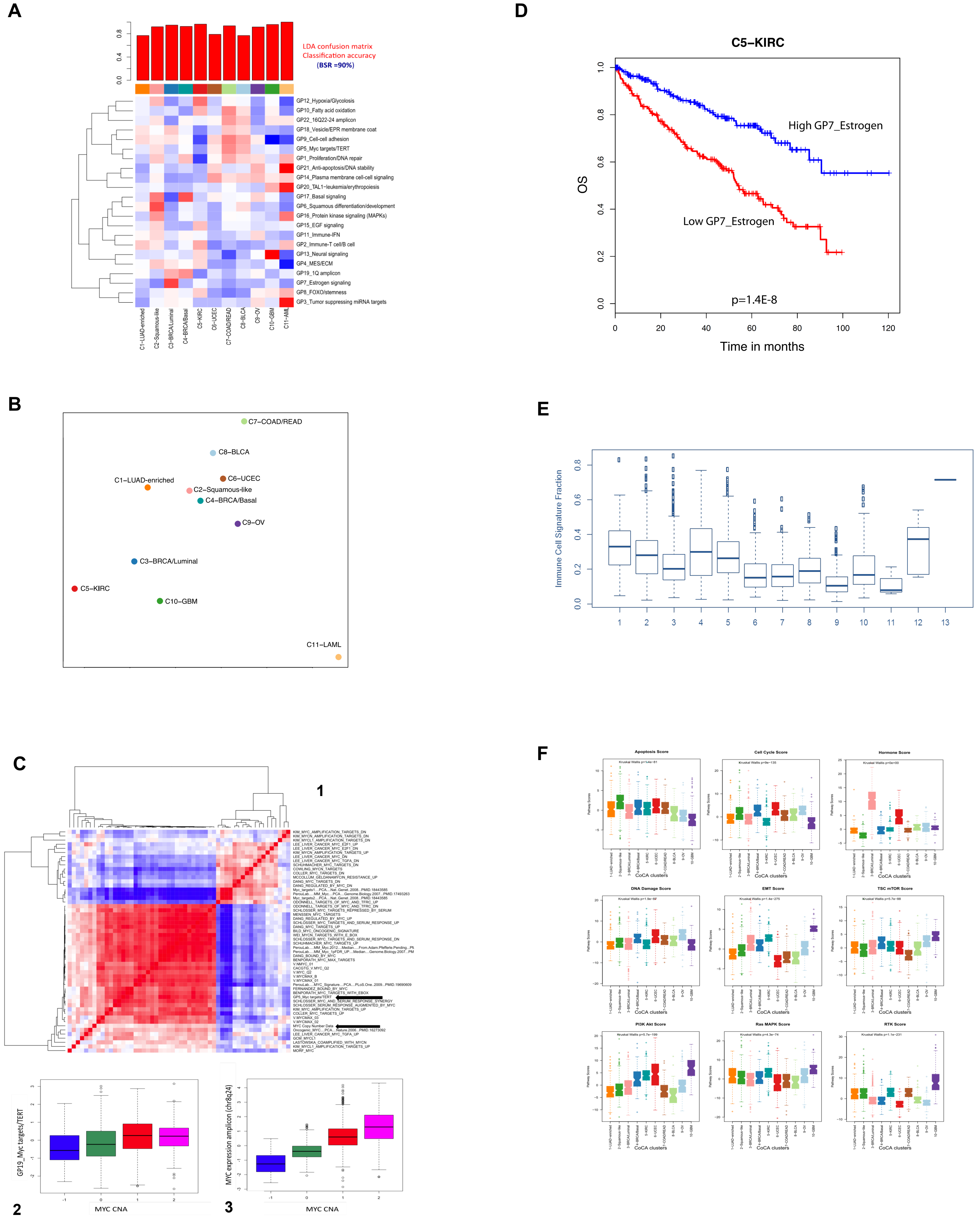
Supplemental Figure

Figure S4. Genomic Determinants of the Integrated Subtypes, Related to Figure 2



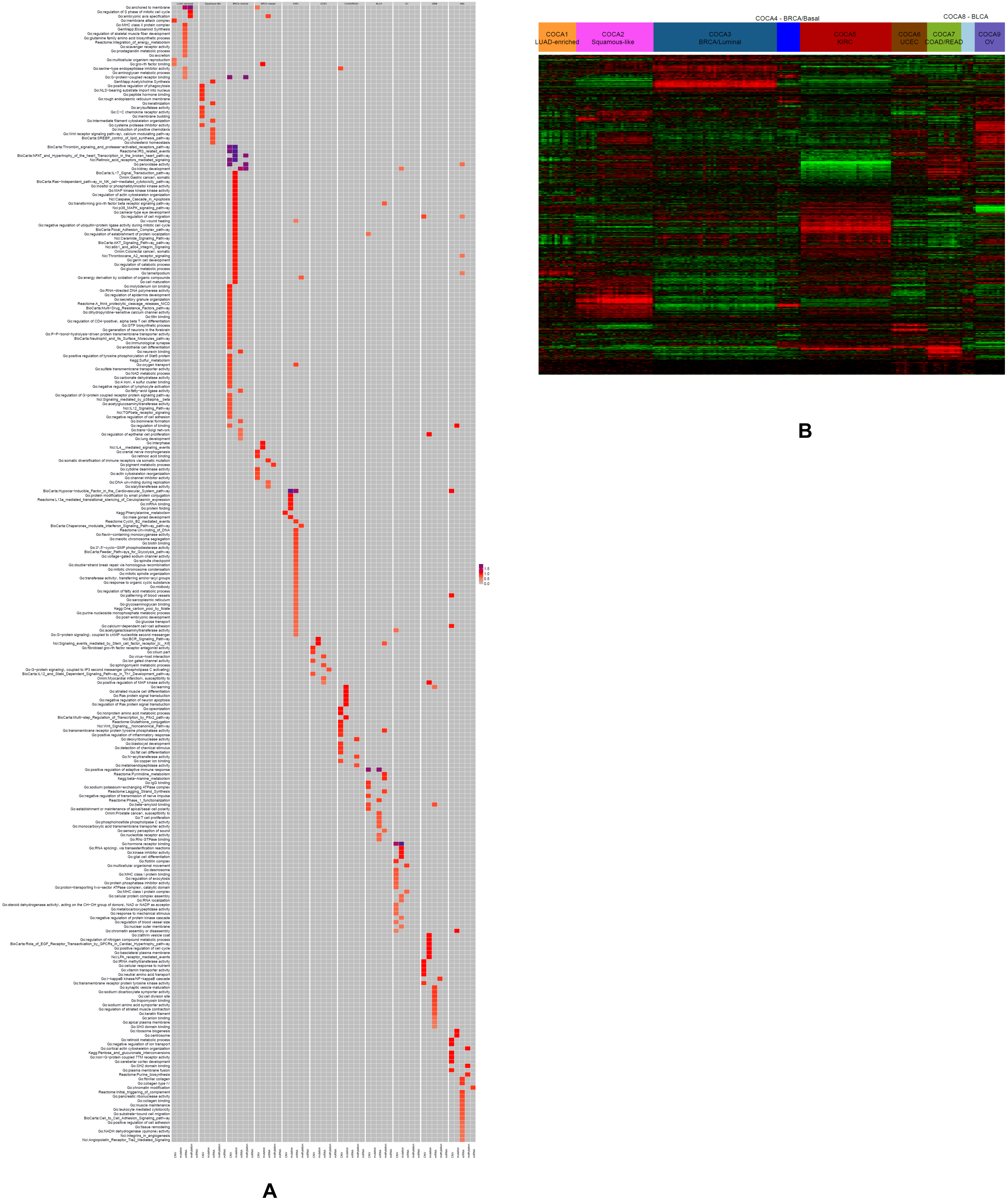
Supplemental Figure

Figure S5. Expression-based Determinants of the Integrated Subtypes, Related to Figure 3



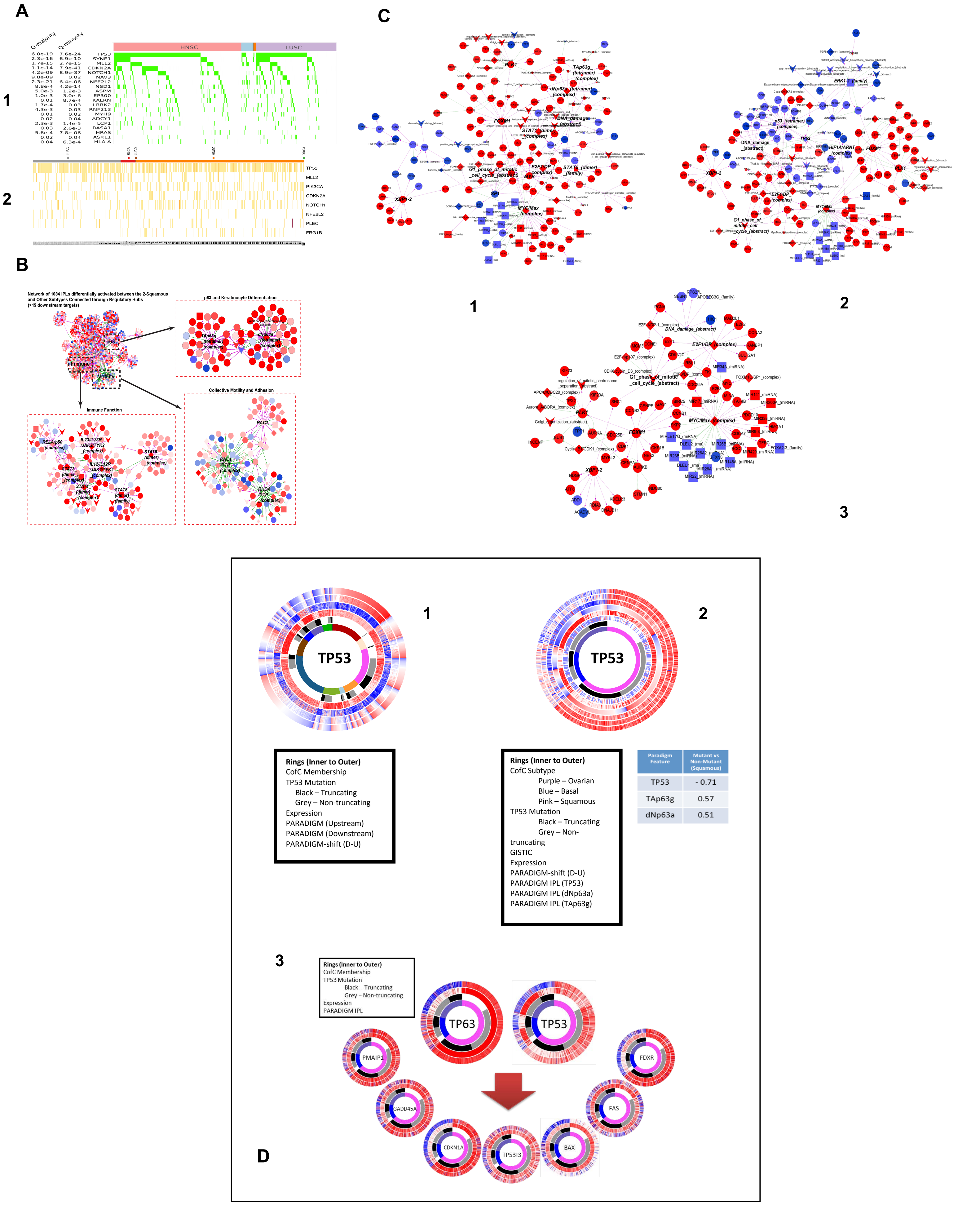
Supplemental Figure

Figure S6. Multi-platform Determinants of the Integrated Subtypes, Related to Figure 3



Supplemental Figure

Figure S7. Convergence of Squamous-like Subtype and Features Common to Squamous, Breast Basal, and Ovarian Subtypes, Related to Figure 4 and Figure 5



Rings (Inner to Outer)
CofC Membership
TP53 Mutation
Black – Truncating
Grey – Non-truncating
Expression
PARADIGM (Upstream)
PARADIGM (Downstream)
PARADIGM-shift (D-U)

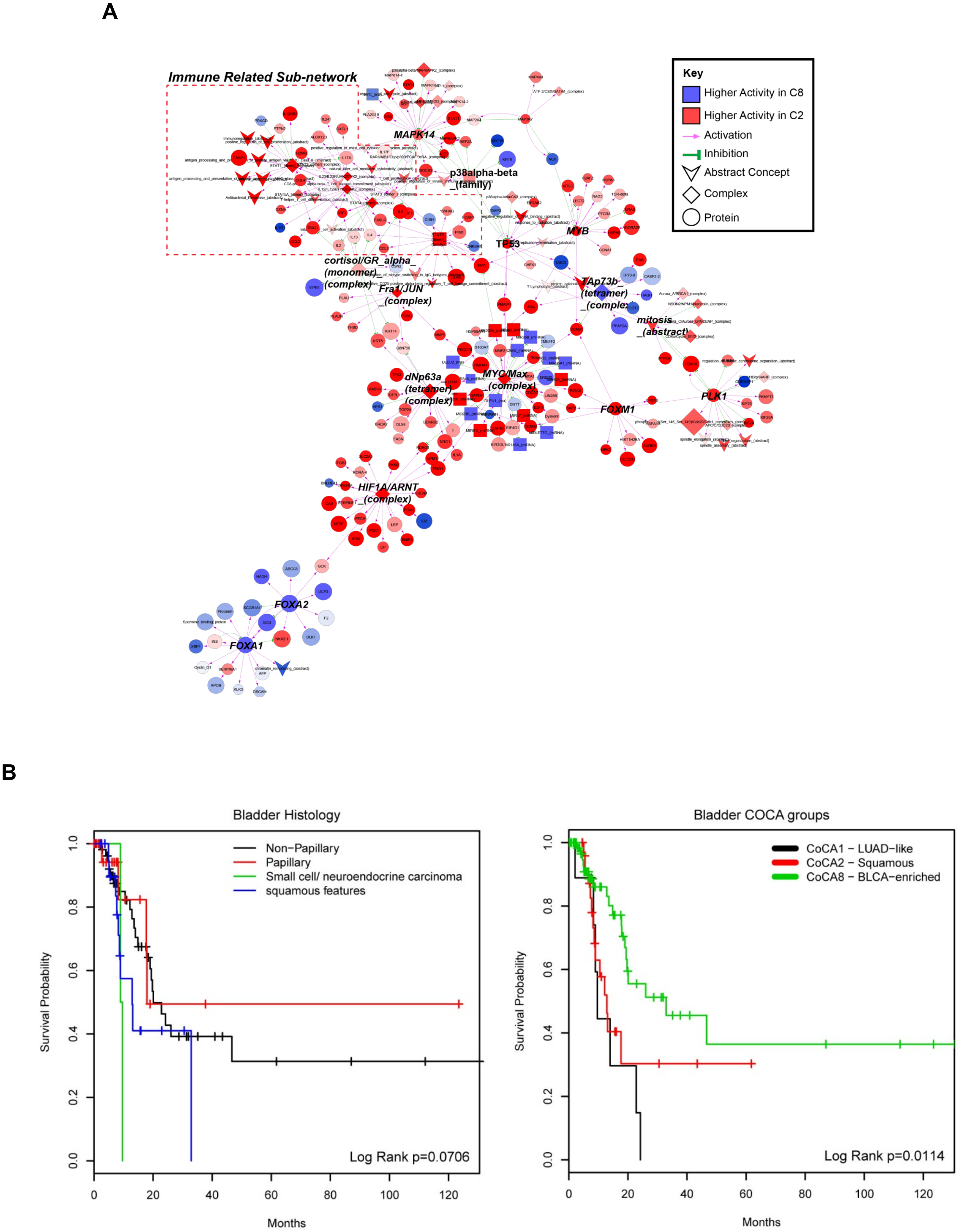
Rings (Inner to Outer)
CofC Subtype
Purple – Ovarian
Blue – Basal
Pink – Squamous
TP53 Mutation
Black – Truncating
Grey – Non-truncating
GISTIC
Expression
PARADIGM-shift (D-U)
PARADIGM IPL (TP53)
PARADIGM IPL (dNp63a)
PARADIGM IPL (TAp63g)

Paradigm Feature	Mutant vs Non-Mutant (Squamous)
TP53	- 0.71
TAp63g	0.57
dNp63a	0.51

Rings (Inner to Outer)
CofC Membership
TP53 Mutation
Black – Truncating
Grey – Non-truncating
Expression
PARADIGM IPL

Supplemental Figure

Figure S8. Divergence of the Bladder Cancer Subtype, Related to Figure 6


A
Immune Related Sub-network
MAPK14
p38alpha-beta _(family)
cortisol/GR_alpha_ (monomer) (complex)
Fra1/JUN _(complex)
TP53
MYB
TAp73b_ (tetramer) _(complex)
mitosis _(abstract)
MYC/Max _(complex)
dNp63a_ (tetramer) _(complex)
FOXM1
PLK1
HIF1A/ARNT _(complex)
FOXA2
FOXA1
Key
🟦 Higher Activity in C8
🟥 Higher Activity in C2
→ Activation
⊣ Inhibition
▽ Abstract Concept
◇ Complex
○ Protein
B
Bladder Histology
Non−Papillary
Papillary
Small cell/ neuroendocrine carcinoma
squamous features
Log Rank p=0.0706
Survival Probability
Months
Bladder COCA groups
CoCA1 − LUAD−like
CoCA2 − Squamous
CoCA8 − BLCA−enriched
Log Rank p=0.0114
Survival Probability
Months