**Pan-cancer molecular classes transcending tumor lineage across 32 cancer types, multiple data platforms, and over 10,000 cases**

Supplementary Methods

TCGA patient cohort

The results here are based upon data generated by TCGA Research Network (http://cancergenome.nih.gov/). Molecular data from 11232 human cancers were aggregated from public repositories (Supplementary Data 1). Tumors spanned 32 different TCGA projects, each project representing a specific cancer type, listed as follows: LAML, Acute Myeloid Leukemia; ACC, Adrenocortical carcinoma; BLCA, Bladder Urothelial Carcinoma; LGG, Brain Lower Grade Glioma; BRCA, Breast invasive carcinoma; CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, Cholangiocarcinoma; CRC, Colorectal adenocarcinoma (combining COAD and READ projects); ESCA, Esophageal carcinoma; GBM, Glioblastoma multiforme; HNSC, Head and Neck squamous cell carcinoma; KICH, Kidney Chromophobe; KIRC, Kidney renal clear cell carcinoma; KIRP, Kidney renal papillary cell carcinoma; LIHC, Liver hepatocellular carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; DLBC, Lymphoid Neoplasm Diffuse Large B-cell Lymphoma; MESO, Mesothelioma; OV, Ovarian serous cystadenocarcinoma; PAAD, Pancreatic adenocarcinoma; PCPG, Pheochromocytoma and Paraganglioma; PRAD, Prostate adenocarcinoma; SARC, Sarcoma; SKCM, Skin Cutaneous Melanoma; STAD, Stomach adenocarcinoma; TGCT, Testicular Germ Cell Tumors; THYM, Thymoma; THCA, Thyroid carcinoma; UCS, Uterine Carcinosarcoma; UCEC, Uterine Corpus Endometrial Carcinoma; UVM, Uveal Melanoma. Cancer molecular profiling data were generated through informed consent as part of previously published studies and analyzed in accordance with each original study’s data use guidelines and restrictions.

Datasets

RNA-seq and miRNA-seq data were obtained from The Broad Institute Firehose pipeline (<http://gdac.broadinstitute.org/>). All RNA-seq samples (representing n=10224 cases) were aligned using the by UNC RNA-seq V2 pipeline(1). Expression of coding genes was quantified for 20531 features based on the gene models defined in the TCGA Gene Annotation File. For miRNA-seq, data were generated using either the Illumina GAIIx or HiSeq 2000 platforms (representing n=10128 cases); to help correct for batch effects between data platforms (GAIIx versus HiSeq), we used the combat software(2) (with platform as the “batch” and cancer type as the “experimental group,” using R code provided by ref(2)), where 423 miRNA features (features with average expression>1 across all samples) were included in the final compiled dataset and downstream analyses. Proteomic data generated by RPPA across 7663 patient tumors (“Level 4” data) were obtained from The Cancer Proteome Atlas (<http://tcpaportal.org/tcpa/>). DNA methylation profiles had been generated by TCGA using either the Illumina Infinium HumanMethylation450 (HM450) or HumanMethylation27 (HM27) BeadChips (Illumina, San Diego, CA) (n=10959 cancer cases having methylation data). To correct for batch effects between data platforms (HM450 versus HM27), we used the combat software (with platform as the “batch” and cancer type as the “experimental group”)(2). GISTIC 2.0 was applied to the transformed copy number data obtained from Affymetrix SNP 6.0 arrays (n=10845 tumor profiles in all), with a noise threshold used to determine copy gain or loss; low-level gene gain, high-level gene amplification, low-level copy loss, or high-level copy loss were inferred using the “thresholded” calls as made by Broad Firehose pipeline (using +1, +2, -1, or -2, respectively).

Somatic mutation calls were obtained from the publicly-available “MC3” TCGA MAF file (covering n=10224 patients); variants called by two or more algorithms were used in this study. The MC3 dataset was produced using six different variant calling algorithms from four centers on the over 10,000 tumor/normal pairs in TCGA. The MC3 MAF provides uniform calling of mutations across all TCGA projects and samples. The MAF is described in more detail and provided at <https://www.synapse.org/#!Synapse:syn7214402>.

Multiplatform-based subtype discovery across cancers

By k-means clustering method (using the “kmeans” R function, with 30 clusters and 1000 maximum iterations and 1000 random sets), cancer cases were subtyped according to each of the individual data platforms for DNA methylation, DNA copy alteration, mRNA expression, miRNA expression, and protein expression. For mRNA expression, the top 2000 genes with the highest average standard deviation by cancer type were used in the clustering, with log base 2-transformed values centered to standard deviations from the median across all tumors. For DNA methylation, the top 2000 genes with the highest average standard deviation by cancer type were used in the clustering, with no centering of values. For DNA copy, log base 2 (tumor/normal) copy values were collapsed, or averaged, into cytoband regions. For miRNA expression, all available features in the batch corrected dataset were used, with log base 2-transformed values centered to standard deviations from the median across all tumors. For protein expression (by RPPA), all feature were used, with log base 2-transformed values centered to standard deviations from the median across all tumors. Subtypes defined from each platform were coded into a series of indicator variables for each subtype, with the matrix of 1 and 0s then clustered by a Cluster of Cluster Analysis (COCA) (3-5) to define integrated subtypes. The clustering of the COCA matrix was carried out using both hierarchical clustering (Supplementary Figure 4A, average linkage) and k-means clustering (Supplementary Figure 4B).

Pan-cancer molecular class discovery

For the mRNA expression platform, log base 2-transformed expression values within each cancer type (as defined by TCGA project) were normalized to standard deviations from the median within each cancer type. By k-means clustering method (using the “kmeans” R function, with 10 clusters and 1000 maximum iterations and 25 random sets), cancer cases were subtyped, based on the top 2000 features with the most variable expression on average by cancer type (using standard deviation of the log2-transformed expression values as a measure of variability). No genes located on the X or Y chromosome were used in the clustering to define the pan-cancer classes (i.e. these genes were purposely excluded from the top 2000 genes used to cluster the expression profiles). Using top standard deviation as a selection criteria effectively worked against these top 2000 genes including any significant number having artificially reduced standard deviations by virtue of their being expressed in only a fraction of cases.

As a guide for determining an optimal number of subtypes to use as the final solution, a delta area plot generated using ConsensusClusteringPlus(6) was considered, using 3000 randomly selected TCGA cancer cases. For the k=10 subtype solution, we defined the top differential genes associated with each subtype, or pan-cancer “class”. Taking the top 2000 mRNA features, we first computed the two-sided t-test for each gene and each class, comparing expression levels of each class with that of the rest of the tumors. We then selected the top 100 genes with the lowest p-value for each subtype; however, for the c2 class only three genes were associated, and for the c7 class only 51 genes were associated, resulting in 854 top class-specific genes in all (Supplementary Data 2). In a similar manner as carried out for mRNA platform, top features associated with pan-cancer class for RPPA, miRNA, and DNA methylation platforms were identified (Figure 1B and Supplementary Figure 5).

Somatic alteration categories

For the pathway-centric view of nonsilent gene mutations, copy alterations, and epigenetic silencing events in TCGA cohort (Supplementary Figure 8A), key pathways and genes considered included: HIPPO pathway (*NF2*, *SAV1*, *WWC1*), NRF2 pathway (*NFE2L2*, *KEAP1*, *CUL3*, *SIRT1*, *FH*), chromatin modification (*CREBBP*, *EHMT1*, *EHMT2*, *EP300*, *EZH1*, *EZH2*, *KAT2A*, *KAT2B*, *KDM1A*, *KDM1B*, *KDM4A*, *KDM4B*, *KDM5A*, *KDM5B*, *KDM5C*, *KDM6A*, *KDM6B*, *KMT2A*, *KMT2B*, *KMT2C*, *KMT2D*, *KMT2E*, *NSD1*, *SETD2*, *SMYD4*, *SRCAP*), SWI/SNF complex (*ACTB*, *ACTL6A*, *ACTL6B*, *ARID1A*, *ARID1B*, *ARID2*, *BCL11A*, *BCL11B*, *BCL6*, *BCL6B*, *BRD7*, *BRD9*, *DPF1*, *DPF2*, *DPF3*, *PBRM1*, *PHF10*, *SMARCA2*, *SMARCA4*, *SMARCB1*, *SMARCC1*, *SMARCC2*, *SMARCD1*, *SMARCD2*, *SMARCD3*, *SMARCE1*), mTOR pathway (*AKT1*, *AKT2*, *AKT3*, *MTOR*, *PIK3CA*, *PIK3R1*, *PTEN*, *RHEB*, *STK11*, *TSC1*, *TSC2*, *IDH1*, *IDH2*, *VHL*), MYC (*MYC*, *MYCN*), Wnt/beta-catenin (*APC*, *AXIN1*, *CTNNB1*, *FGF19*, *NCOR1*), and p53-related (*ATM*, *CCND1*, *CCNE1*, *CDK4*, *CDKN1A*, *CDKN2A*, *E2F2*, *E2F3*, *FBXW7*, *MDM2*, *RB1*, *TP53*). The DNA methylation level as interrogated by cg13601799 and cg13672843 was used for CDKN2A and VHL, respectively; as carried out previously (4), a beta value of 0.2 or above was considered evidence for epigenetic silencing. For a number of oncogenes (*AKT1*, *MTOR*, *PIK3CA*, *RHEB*, *NFE2L2*, *BRAF*, *EGFR*, *ERBB2*, *ERBB3*, *HRAS*, *KRAS*, *NRAS*) mutations in oncogenes were considered in the analyses, if the mutations occurred in “hotspot” residues as reported by Chang *et al*.(7); all nonsilent mutations in putative tumor suppressor genes (e.g. *TP53*) were considered in the analyses.

Gene and protein signatures

Gene transcription signature scores associated with pathway (e.g. scores for p53, EMT, NRF2/KEAP1, hypoxia, KEGG: Glycolysis/Gluconeogenesis, KEGG: Pentose Phosphate pathway, KEGG: Fatty Acid metabolism, KEGG: TCA Cycle, and KEGG: Oxidative Phosphorylation or OX-PHOS, k-ras, MYC, YAP1, WNT, and NOTCH) were computed as follows. Log base 2-transformed Values for genes in the TCGA pan-cancer mRNA dataset were either normalized across cancers (standard deviations from the median) or normalized within each cancer type (standard deviations from the median of the given cancer type), as indicated in the Results. For p53, NRF2/KEAP1, hypoxia, WNT, NOTCH, and KEGG signatures, the average expression of the set of genes within a given signature were computed. For k-ras, MYC, and YAP1 signatures, normalized expression profiles were scored for the above signatures using our previously described “t-score” metric(1). EMT signature score was computed as done previously(8), with the sum of the normalized epithelial marker gene values subtracted from the sum of the normalized mesenchymal marker gene values. Gene targets of p53 were from ref(9). Gene transcription signature scores of NRF2/KEAP1 pathway where generated as described(10), on the basis of four different signatures as described previously(4). Hypoxia signature was based on the set of canonical HIF1A targets from Harris(11). Gene transcription signature scores of YAP1 pathway where generated as described previously, on the basis of four different signatures(4). MYC signature (from data by Coller et al.(12)) was from ref(13), and the Settleman k-ras sensitivity signature was from ref(14). WNT signature was defined previously(15) (summing up values for WNT antagonist, agonist, and target genes). NOTCH signature was defined previously(16). Normalized pan-cancer RPPA profiles were scored for an mTOR pathway signature, defined as the sum of phosphoprotein levels of mTOR, 4EBP1 (S65, T37/T46, and T70 RPPA features), P70S6K, and S6 (S235/S236 and S240/S244 features).

Analysis of External Normal Tissue and Cell Datasets

The Fantom datasets of gene expression by cell type(17) were analyzed using a previously utilized approach(5). Briefly, the top 2000 most variable mRNAs (used above for the clustering analyses) were examined in both fantom human and fantom mouse expression datasets. Logged expression values (base 2) for each gene in the fantom dataset were centered on the median of sample profiles. For each fantom differential expression profile (genes centered within the fantom dataset), the inter-profile correlation (Spearman’s) was taken with that of each TCGA pan-cancer differential expression profile (with genes normalized within each TCGA project to standard deviations from the median). Fantom profiles represented in the inter-profile correlations table were clustered by hierarchical clustering.

Analysis of External Multi-Cancer Datasets

We examined an external gene expression profiling dataset of multiple cancer types from the Expression Project for Oncology (expO) (GSE2109), classifying each external tumor profile by pan-cancer class as defined by TCGA data. Within each cancer type, log base 2-transformed genes in the expO dataset were normalized to standard deviations from the median. As a classifier, the top set of 854 mRNAs distinguishing between the ten pan-cancer classes in TCGA (from Figure 1B and Supplementary Data 2) was used. For each pan-cancer class, the average value for each gene was computed, based on the centered TCGA expression data matrix (represented in Figure 1B). The Pearson’s correlation between each expO profile and each TCGA pan-cancer class averaged profile was computed. Each expO case was assigned to TCGA pan-cancer class, based on which class profile showed the highest correlation with the given expO profile. Lymph node, Axillary Lymph Node, Parotid gland, and Retroperitoneum expO profiles were not included in the analysis. In a similar manner, cell line profiles from the Cancer Cell Line Encyclopedia (CCLE) expression dataset (18) were each assigned to a pan-cancer class. Correlations between expO expression profiles and fantom profiles, as well as between CCLE profiles and fantom profiles, were computed as described above, with the correlations averaged by specific categories of fantom profiles.

Statistical Analysis

All p values were two-sided unless otherwise specified. Where multiple testing was not an issue, statistical significance was defined at the 0.05 threshold. All available TCGA data in the public domain at the time of this study was utilized, and no patients were deliberately excluded.

Differential expression between comparison groups was assessed using t-test on log-transformed values (base 2). False Discovery Rates (FDRs) were estimated using the method of Storey and Tibshirini(19). The top global differential features presented in main Figure 1 and elsewhere were all well within acceptable FDR considerations (Supplementary Data 2). For example, for RNA-seq platform, 2000 features were originally tested in order to derive the top 854 features (Figure 1B) (aiming for the top 100 genes from each of the ten classes, though for classes 2 and 7 in particular 100 genes that met the cutoff threshold were not found). Of those 854 genes, the largest nominal p-value associated with a gene was 1e-5. Of the 854 genes, 851 had a p-value of 2e-15 or less. Taking the Storey and Tibshirani approach for FDR correction—where we first multiplied the nominal p-values by 10 before the FDR calculation across genes (to conservatively correct for our testing of each of the ten classes)—the estimated FDR for the entire set of 854 was 0.0001. Pearson’s correlations and t-tests using expression data were assessed using log2-transformed values, as log transformation would bring the values closer to a normal distribution(20). In addition to t-tests, Mann-Whitney U-tests were also performed, which do not assume a specific distribution of the data.

Visualization using heat maps was performed using both JavaTreeview (version 1.1.6r4)(21) and matrix2png (version 1.2.1)(22). Heirarchical clustering was carried out using Cluster 3.0 (C Clustering Library version 1.52)(23). R software (version 3.1.0) was used for k-means clustering, survival analysis, and generation of box plots. A stratified Log-rank test was used to evaluate survival differences among pan-cancer classes, when correcting for cancer type. Patient survival data from TCGA were current as of March 31, 2016. The log-rank test assumes that censoring is unrelated to prognosis, that the survival probabilities are the same for subjects recruited early and late in the study, and that the events happened at the times specified. One-sided Fisher’s exact tests and chi-squared tests assume that the binary data are independent.

Evaluation of Batch Effects

Having defined our pan-cancer classes, we examined whether these would be strongly associated with TCGA sample batches. We examined batch effects in two ways: 1) by asking which batches may strongly associate with pan-cancer class, and 2) by asking whether the differential expression patterns that define the pan-cancer classes remain significant after factoring in TCGA batch information. Batch information was obtained for 9346 cases from [http://bioinformatics.mdanderson.org/main/TCGABatchEffects:Overview]. The significance of enrichments between our pan-cancer classes (c1 through c10) and TCGA batches—as defined by batch ID, Tissue Source Site (TSS), and sample shipping date—were calculated (Supplementary Figure 6). As another approach to examining batch effects, we took the 854 mRNAs that best distinguished between the pan-cancer classes (the genes represented in Figure 1A), and asked whether most all of these would remain significant, if batch information was incorporated along with class in assessing the significance of differential expression. For each gene, we used a linear regression model, with dependent variable (continous variable) of expression and with independent variables (all categorical variables): class + BatchId + PlateId + ShipDate + TSS. For the given class, originally associated with the gene (per Figure 1A and Supplementary Data 2), we assessed the significance of differential expression for the given class (as compared with c1 as the reference, or with c2 as the reference in the case of c1 genes). Of the 854 genes originally significantly associated with one of the pan-cancer classes, 849 (99.4%) were also significant (p<0.05) in the linear model (Supplementary Data 2).

Data availability

All data used in this study are publicly available. TCGA data are available through the Genome Data Commons (https://gdc.cancer.gov/) and the Broad Instititute’s Firehose data portal (https://gdac.broadinstitute.org). Somatic mutation calls made from TCGA whole exome sequencing data are available from synapse (https://www.synapse.org/#!Synapse:syn7214402). FANTOM consortium datasets are available from <http://fantom.gsc.riken.jp/data/>. Cancer Cell Line Encyclopedia (CCLE) datasets available from <http://www.broadinstitute.org/ccle>.ExpO expression dataset is available via the Gene Expression Omnibus.

Author Contributions

Conceptualization: C.J.C.; Methodology: C.J.C., F.C., Y.Z.; Investigation: C.J.C., F.C., Y.Z., D.L.G., B.D., D.J.K., M.I.; Formal Analysis: C.J.C., F.C., Y.Z.; Data Curation: C.J.C.; Visualization; C.J.C.; Writing – Original Draft: C.J.C.; Writing – Review & Editing: C.J.C.,D.L.G., B.D., D.J.K., M.I.; Supervision: C.J.C.

References

1. The\_Cancer\_Genome\_Atlas\_Research\_Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature **2013**;499(7456):43-9 doi 10.1038/nature12222.

2. Johnson W, Rabinovic A, Li C. Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics **2007**;8(1):118-27.

3. Hoadley K, Yau C, Wolf D, Cherniack A, Tamborero D, Ng S*, et al.* Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. Cell **2014**;158(4):929-44.

4. Chen F, Zhang Y, Şenbabaoğlu Y, Ciriello G, Yang L, Reznik E*, et al.* Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma. Cell Rep **2016**;14(10):2476-89.

5. Chen F, Zhang Y, Parra E, Rodriguez J, Behrens C, Akbani R*, et al.* Multiplatform-based Molecular Subtypes of Non-Small Cell Lung Cancer. Oncogene **2016**;E-pub Oct 24.

6. Wilkerson M, Hayes D. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics **2010**;26(12):1572-3.

7. Chang M, Asthana S, Gao S, Lee B, Chapman J, Kandoth C*, et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol **2016**;34(2):155-63.

8. Gibbons D, Creighton C. Pan-cancer survey of epithelial-mesenchymal transition markers across The Cancer Genome Atlas. Dev Dyn **2017**;E-pub 2017 Jan10.

9. The\_Cancer\_Genome\_Atlas\_Network. Comprehensive molecular portraits of human breast tumours. Nature **2012**;490(7418):61-70 doi 10.1038/nature11412.

10. Abazeed M, Adams D, Hurov K, Tamayo P, Creighton C, Sonkin D*, et al.* Integrative radiogenomic profiling of squamous cell lung cancer. Cancer Res **2013**;73(20):6289-98.

11. Harris A. Hypoxia--a key regulatory factor in tumour growth. Nat Rev Cancer **2002**;2(1):38-47.

12. Coller HA, Grandori C, Tamayo P, Colbert T, Lander ES, Eisenman RN*, et al.* Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. Proc Natl Acad Sci USA **2000**;97:3260-5.

13. Creighton C. Multiple oncogenic pathway signatures show coordinate expression patterns in human prostate tumors. PloS one **2008**;3(3):e1816.

14. Singh A, Greninger P, Rhodes D, Koopman L, Violette S, Bardeesy N*, et al.* A gene expression signature associated with "K-Ras addiction" reveals regulators of EMT and tumor cell survival. Cancer Cell **2009**;15(6):489-500.

15. Gingras M, Covington K, Chang D, Donehower L, Gill A, Ittmann M*, et al.* Ampullary Cancers Harbor ELF3 Tumor Suppressor Gene Mutations and Exhibit Frequent WNT Dysregulation. Cell Rep **2016**;14(4):907-19.

16. Kwon O, Zhang L, Wang J, Su Q, Feng Q, Zhang X*, et al.* Notch promotes tumor metastasis in a prostate-specific Pten-null mouse model. J Clin Invest **2016**;126(7):2626-41.

17. FANTOM\_Consortium\_and\_the\_RIKEN\_PMI\_and\_CLST\_(DGT), Forrest A, Kawaji H, Rehli M, Baillie J, de Hoon M*, et al.* A promoter-level mammalian expression atlas. Nature **2014**;507(7493):462-70.

18. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin A, Kim S*, et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature **2012**;483(7391):603-7.

19. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA **2003**;100:9440-5.

20. Zwiener I, Frisch B, Binder H. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. PloS one **2014**;9(1):e85150.

21. Saldanha AJ. Java Treeview--extensible visualization of microarray data. Bioinformatics **2004**;20:3246-8.

22. Pavlidis P, Noble W. Matrix2png: A Utility for Visualizing Matrix Data. Bioinformatics **2003**;19(2):295-6.

23. de Hoon M, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics **2004**;20(9):1453-4.