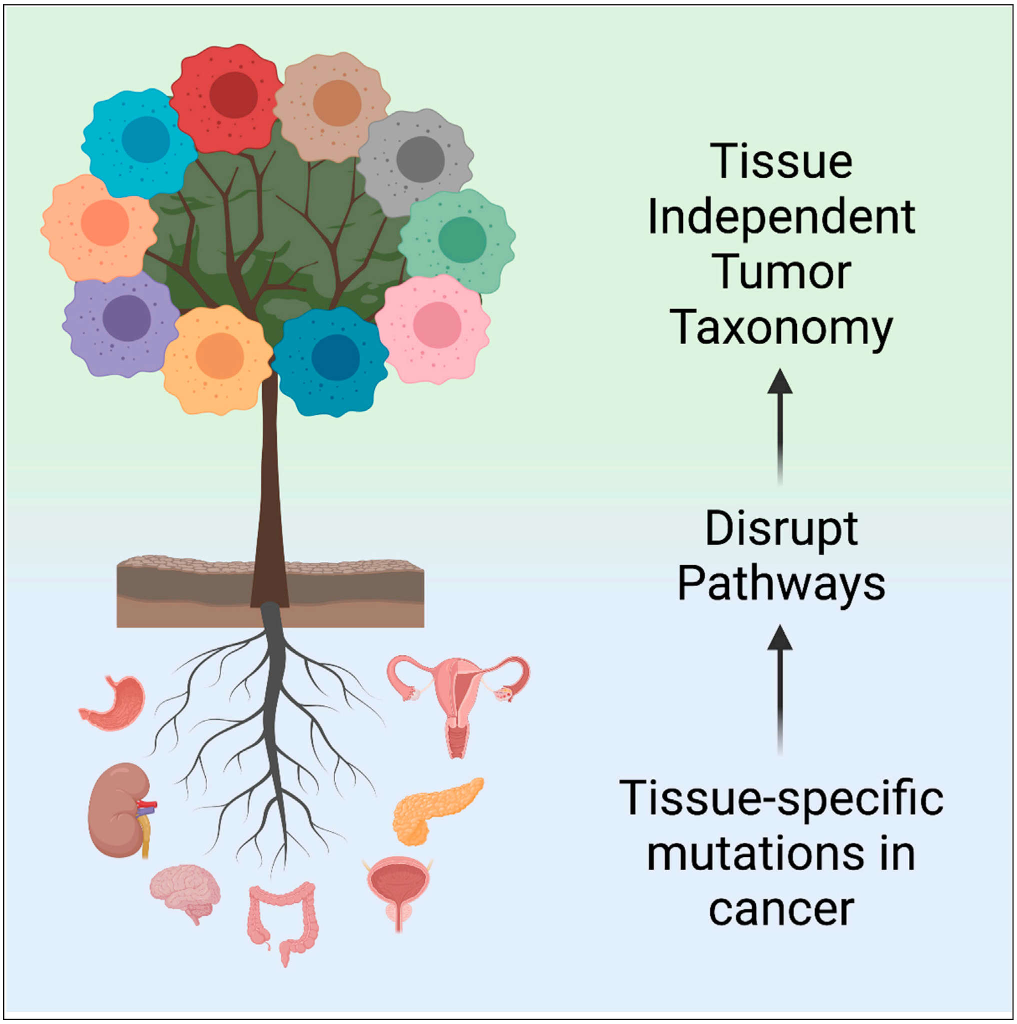
iScience

CellPress OPEN ACCESS

Article

A molecular taxonomy of tumors independent of tissue-of-origin


Tissue Independent Tumor Taxonomy
Disrupt Pathways
Tissue-specific mutations in cancer

Peter T. Nguyen, Simon G. Coetzee, Daniel L. Lakeland, Dennis J. Hazelett

dennis.hazelett@cshs.org

Highlights

Pathway-based clustering is applied to reclassify tumors using somatic mutations

Clusters represent distinct molecular features that cut across tissue-of-origin

Risk of death is modelled using tissue origin and cluster membership

Nguyen et al., iScience *24*, 103084
October 22, 2021 © 2021 The Authors.
https://doi.org/10.1016/j.isci.2021.103084

iScience

CellPress
OPEN ACCESS

Article

A molecular taxonomy of tumors independent of tissue-of-origin

Peter T. Nguyen,[1,4] Simon G. Coetzee,[1,4] Daniel L. Lakeland,[2] and Dennis J. Hazelett[1,3,5,*]

SUMMARY

Cancer is an organism-level disease, impacting processes from cellular metabolism and the microenvironment to systemic immune response. Nevertheless, efforts to distinguish overarching mutational processes from interactions with the cell of origin for a tumor have seen limited success, presenting a barrier to individualized medicine. Here we present a pathway-centric approach, extracting somatic mutational profiles within and between tissues, largely orthogonal to cell of origin, mutational burden, or stage. Known predisposition variants are equally distributed among clusters, and largely independent of molecular subtype. Prognosis and risk of death vary jointly by cancer type and cluster. Analysis of metastatic tumors reveals that differences are largely cluster-specific and complementary, implicating convergent mechanisms that combine familiar driver genes with diverse low-frequency lesions in tumor-promoting pathways, ultimately producing distinct molecular phenotypes. The results shed new light on the interplay between organism-level dysfunction and tissue-specific lesions.

INTRODUCTION

Advances in technology have greatly expanded our view into the mechanisms of cancer at the systems biology level. Next-generation sequencing technologies have made it possible to study germline and somatic mutations, expression profiles, DNA methylation, and copy number variations from the same tissue. To take advantage of these tools, several large consortia, including the cancer genome atlas (TCGA) and pan-cancer analysis of whole genomes (PCAWG) sequenced large numbers of tumors and collected data from multiple assays to analyze together, with the goal of integration and increased understanding of the mechanisms of cancer.

Considerable progress has been made analyzing these data. Statistical analyses identify hundreds of global and tissue-specific cancer driver genes (Dees et al., 2012; Jiang et al., 2019; Kumar et al., 2015; Lawrence et al., 2014; Tamborero et al., 2013; Tokheim et al., 2016; Zhao et al., 2019) using approaches aimed at detecting genes mutated at a greater rate than expected due to chance. Dees et al. (Dees et al., 2012) developed a pipeline to separate driver mutations from passenger mutations. Tamborero et al. (Tamborero et al., 2013) expanded on this idea to identify tumor mutation biases towards specific regions of a protein sequence. Lawrence et al. (Lawrence et al., 2014) focused on identifying cancer driver genes with intermediate mutation frequencies using tumor-normal pairs. It has been estimated that fewer than five mutations in key oncogenes and/or tumor suppressors would be sufficient to transform a normal cell to a cancerous state (Iranzo et al., 2018; Vogelstein and Kinzler, 2015). Mutations in cancer driver genes are commonplace in healthy tissue and correlate with age and environmental exposures (Martincorena, 2019).

Other studies provide a comprehensive view of mutations, gene expression, and genomic signatures, with the goal of understanding common themes of all cancers independent of tissue of origin. Understanding cancer as a disease of the cell has long been a goal of the field as characterized in essays by Hanahan and Weinberg (Hanahan and Weinberg, 2000, 2011). The first of these studies in genomics identified 11 subtypes from 12 cancer types, using integrative analysis with co-equal weighting of gene expression, methylation, copy number, and proteomics data (Hoadley et al., 2014). The principal finding was that tissue-of-origin is the predominant driving factor, though ~10% could be reclassified independent of tissue-of-origin. In a second study involving 33 cancer types and a much greater number of tumors, the authors identified 28 clusters that could be further subdivided into organ specific groups, including pan-gastrointestinal, pan-gynecological, pan-squamous, pan-gynecological/squamous, and pan-kidney (Hoadley et al., 2018).

[1]The Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA

[2]Lakeland Applied Sciences LLC, Los Angeles, CA 91001, USA

[3]Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA

[4]These authors contributed equally

[5]Lead contact

*Correspondence: dennis.hazelett@cshs.org

https://doi.org/10.1016/j.isci.2021.103084

More recently, pan-cancer mutations and other data have been looked at as disruptions of normal pathway activity. Pathway-centric analysis of tumors provides additional benefits over gene-centric analysis: (1) It is significantly less noisy because it aggregates molecular events across multiple genes in the same pathway and (2) the identification of a potential causal mechanism is easier to interpret as genetic aberrations are linked to molecular pathways. In a perspective, Creixell et al. (Creixell et al., 2015) described approaches of pathway and network analysis applicable to next-generation sequencing data, the most common approach being fixed-gene enrichment analysis to identify over-represented pathways. Sanchez-Vega et al. (Sanchez-Vega et al., 2018) carried out an exhaustive analysis of pathway enrichment using whole genome data. Focusing on ten frequently altered canonical signaling pathways in cancer, they explored the mutation frequencies of coding regions of genes in these pathways for each cancer type. Network approaches address the sparseness of mutations, allowing genes to be influenced by mutations in nearest network neighbors (Horn et al., 2018; Iranzo et al., 2018; Leiserson et al., 2014). These methods are powerful especially when applied to smaller numbers of tumors. One of the challenges in the field is that there may be thousands of moderate effect genes that occur at such low frequency that they are impossible to detect using positive selection theory. Some researchers have attempted to address this challenge by applying machine learning to cancer data to discover groups of functionally related genes as they interact with larger pathways and networks (Colaprico et al., 2020; Kim and Kim, 2018; Mourikis et al., 2019). These studies have proven very effective at highlighting fundamental disease phenotypes at the pathway level across cancers with different origins at the cellular and tissue level.

In this study, we attempt to understand how tissue-specific gene disruptions create common cancer phenotypes by focusing on discrete molecular pathways as the unit of disruption. Our approach strips all cell-type-specific information from the mutation data and equates gene-level mutations to cell-biological pathway disruptions. We use this heuristic to evaluate all cancers and show that, surprisingly, tumors that exhibit tissue-specific gene mutation patterns nonetheless fall into common categories of pathway disruption having unique prognoses in each cancer type.

RESULTS

Taxonomy of tumors based on disrupted molecular pathways

To study cancer pathways we obtained a set of 7,607 solid tumors from The Cancer Genome Atlas (TCGA) through the Genomic Data Commons (GDC) (gdc.cancer.gov) portal. TCGA data are most appropriate given the relative completeness of the patient metadata for survival and staging. We chose to analyze exome sequencing data because the affected target gene is known unambiguously. Therefore, we selected all missense, nonsense, frameshift, stop-loss, untranslated regions, and splicing mutations. To minimize bias from well-studied diseases and processes, we selected 377 Reactome pathways (reactome.org) (see Table S1) of interest corresponding to basic cellular processes and biochemical pathways, excluding those corresponding to catalytic categories (e.g., ''transcription factors'') or disease associations (e.g., ''mutated in colon cancer'') and filtered our gene list on membership in these pathways (total of 8,940 genes).

To avoid bias toward pathways with more genes, we counted disruptions if one or more member genes were mutated. We do not attempt to calculate enrichment for mutations within a pathway. Binarized pathways are likely noisy for a couple reasons. First, point mutations can be deleterious (attenuating, hypomorphic, or antimorphic) or activating (neomorphic or hypermorphic) in genes, and these can in turn be oncogenes or tumor suppressors. For this study we assume a significant fraction of these mutations are generically disruptive to pathway activity because it is impossible to know the tumor promoting effects of all mutations, including rarely studied genes. Second, low- and non-expressed genes accumulate mutations at a higher rate because of transcription coupled repair (Aitken et al., 2020; Cummings et al., 2020; Kandoth et al., 2013; Kim and Kim, 2018). To address this issue, we identified low expressed genes in each type of cancer and eliminated them for that cancer type only. Highly expressed genes could also have high mutation rates owing to transcription induced mutagenesis (Park et al., 2012). This phenomenon could result in cell-type-specific biases that might result in predisposition to different classes of cancer; therefore, we did not exclude these genes from our analysis. After selecting pathways and genes, we compiled a matrix of pathways, assigning a Boolean value of 1 to each pathway with one or more genes mutated and 0 for all others (Figure 1A).

We investigated this dataset using multiple correspondence analysis (MCA) (Lê et al., 2008), and visually summarized the analysis with UMAP (Figure 1B and see interactive media from Data S1 (junkdnalab.

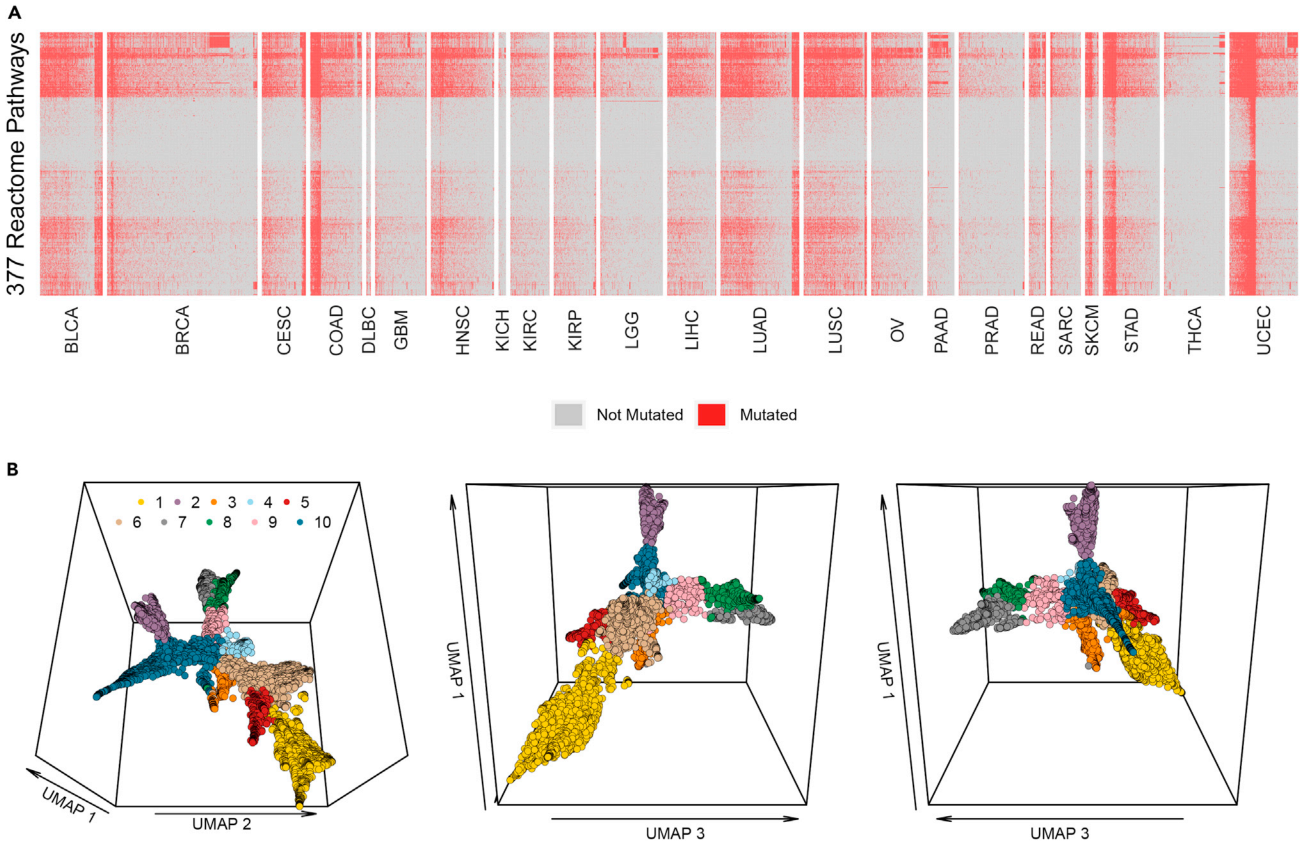


Figure 1. Clustering pathways of tumor samples

(A) Each of 377 selected Reactome pathways (rows) is classified as disrupted if one or more genes is mutated in the tumor sample (columns). Red cells denote pathway disruption. Tumor types reflect standardized abbreviations from the TCGA project.

(B) Different rotational perspectives of the same MCA-based UMAP projection in three-dimensional space. Each dot corresponds to a tumor sample. The same colors indicate the tumor's cluster identity throughout this manuscript.

shinyapps.io/PANCAN_supplemental/)) (McInnes et al., 2020). We used the resulting graph coordinates to perform density based clustering with HDBSCAN (McInnes et al., 2017), which resulted in identification of 10 well-defined clusters capturing about 80% of the tumor samples. To capture the remaining samples into one of these 10 clusters we used kNN (see STAR Methods for details on clustering methods).

Independence from tissue-of-origin

Having defined tumors in terms of pathway disruption profile, we sought to understand whether different cancer types segregate into one or more predominant clusters. To our surprise, most cancer types were *not* heavily biased in one cluster, and all well-represented cancer types had tumors in every cluster (see Figure 2A and full tumor profiles in Figure S1, interactive media Data S1), suggesting that these clusters indicate molecular pathology largely independent of tissue-of-origin. As an example of one cancer that does have a biased pathway profile, pancreatic adenocarcinoma (PAAD) was predominantly found in cluster 8 (Figures 2A and S1 and Data S1), but even PAAD comprises tumors from the nine remaining classes. Thus, patients with these tumors have potentially different underlying molecular pathologies.

Independence from molecular and histological subtype

Many cancers have molecular or histological subtypes defined based on gene expression, pathology or other-omics profiles. These subtypes often have different standards of care owing to different drug sensitivity (or other clinical trial data). If histological subtypes represent true molecular phenotypes, one predicts they should segregate with pathway-based clusters, supporting the clusters as proxies for molecular pathology sub-typing. To our surprise, we found a similar result to the previous analysis of cancer types. To illustrate this, we projected annotations for each of the breast cancer subtypes, composed of Triple-negative/Basal-like, Her2 positive,

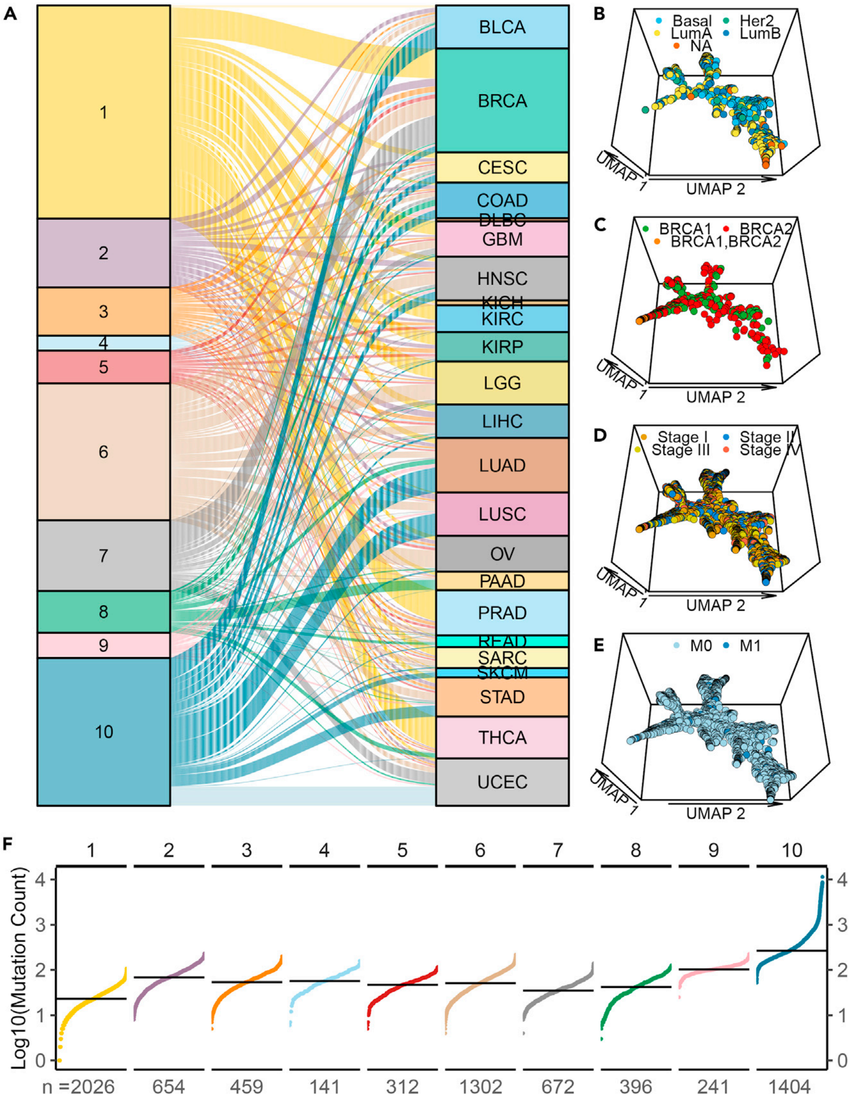


Figure 2. Pathway-based clustering independent of tissue-of-origin

(A) Sankey plot of correspondence between cancer type and cluster identity.

(B) Projection of breast cancer subtypes onto the UMAP. See also Figure S2.

(C) Projection of BRCA1/2 somatic mutation onto the UMAP.

(D) Projection of tumor stage onto the UMAP, regardless of cancer type.

(E) Projection of metastatic status onto the UMAP. Abbrevs: M0 = non-metastatic tumors, M1 = metastatic tumors.

(F) Somatic mutation frequencies for each cluster. Vertical axis shows log mutation count, horizontal axis is cluster identity. Each dot represents an individual tumor sample, ranked lowest to highest by mutation count. The median mutation count in each cluster is indicated by the horizontal line.

normal-like, and luminal A and B subtypes onto the UMAP. These are among the most heavily studied molecular subtypes in cancer, each with different prognoses and standards of care. We did not observe exclusive segregation by cluster for these subtypes (Figure 2B). We also projected histological subtype data for the remaining cancers (see Figure S2 and see interactive media from Data S1); we find that the subtypes, though often biased to one or more clusters, are almost never exclusive. We interpret these data to mean that the clusters do not correspond to previously identified molecular subtypes within the parent cancer types.

Independence from drivers of genome instability

There are several familial cancer-causing mutations studied for differences in basic biology, survival, and treatment outcomes. The functions of these genes are related to risk factors such as genome stability, proof-reading & DNA damage repair, and telomere length. *BRCA1/2* genes are key for DNA double-stranded break repair (Davies et al., 2001; Moynihan et al., 1999) and germline mutations in these genes confer elevated risk for breast, prostate, and ovarian cancers. The mechanism of risk is thought to involve loss of heterozygosity, resulting in loss of the wildtype, functional allele (Gudmundsson et al., 1995), so we projected somatic mutations for *BRCA1* and *BRCA2* genes onto the UMAP. We did not observe segregation of these mutations into specific clusters (Figure 2C and see interactive media from Data S1). We also projected mismatch repair (MMR) genes *MSH2*, *MSH6*, *MLH1*, *MLH3*, *PMS1* and *PMS2*, *BRIP1*, *RAD51*, *CHEK2* and *APC*. None of these genes except for *APC* exhibited remarkable specificity with respect to cluster assignment (Figure S3). To look at other risk factors such as maintenance of DNA methylation levels and telomere length, we projected somatic mutations of the *TET2* and *TET3* genes, plus *TERT*, *TEP1*, and *DKC1*, and observed similar lack of cluster bias (Figure S3).

Independence of stage, mutation count and mutation profile

Tumor staging is based on pathological criteria, including tumor diameter, which can vary greatly in importance between different tissues. Stage is used clinically as a proxy for advancement toward a more deadly state and metastasis. Thus, it is possible that more advanced tumors have common pathway disruption profiles. The UMAP, which features a series of lobe-like structures on a common backbone of tumor samples, could reflect progression through a series of stages. The backbone starts with a cluster of tumors (class 1) that has the fewest point mutations and culminates in a cluster (class 10) which has nearly every pathway disrupted (Figure 2F). However, we don't observe a trend in the overall mutation burden across the backbone of the UMAP. Nonetheless, to test the hypothesis that the molecular-pathway disruption clusters represent advancement through stages, we projected staging data onto the UMAP. Similar to tissue of origin and other categories of tumor, we do not observe any bias among the stages to specific clusters (Figure 2D), suggesting that stage does not contribute to cluster identity.

Finally, as a measure of tumor advancement, metastasis is the condition in which certain phenotypic criteria are met: loss of differentiation, cell-cell contacts, epithelial to mesenchymal transition, immune system evasion, and tissue invasiveness (Hanahan and Weinberg, 2011). To determine whether any clusters correspond to an especially advanced stage of cancer across tissue types, we projected the metastases onto our UMAP, and surprisingly we observed even distribution of the samples across clusters (Figure 2E). This final observation suggests that our pathway-disruption clustering is dependent on particular combinations of gene mutations affecting different pathways that can each give rise to advanced stages of disease and metastasis, regardless and independent of overall mutational burden.

Tissue specific genes define cluster membership

To identify pathway enrichment across all cancers, we created a list of pathway disruptions with percent mutated samples and top genes (Table S1). As expected, these analyses reveal the broad importance of many well known pathways that are disrupted in cancer, including ''PIP3 activates Akt signaling'' (77% of samples), ''MAP1K/MAP3K signaling'' (70% of samples), ''Mitotic G2-G2/M phases'' (67% of samples), ''Cellular senescence'' (64% of samples), ''G2/M Checkpoints'' (62% of samples), *etc.*

To discover what pathways are most important for clustering, we calculated percent enrichment *within a cluster* relative to all other clusters and ranked pathways from highest to lowest enrichment. We visualized enrichment as a heatmap (Figure 3A). Using this approach, we identified about fourteen pathways per cluster (enrichment score $\geq$0.3, 95% confidence; see STAR Methods) (Table S2). Clusters 7, 8, and 9 had several pathways in common. To explore the specific pathways marking each cluster, we projected disruptions for each of the 377 pathways onto the UMAP (Table S2 and see interactive media from Data S1). Clusters 3 and 5 were distinguished by metabolic pathways including RNA and protein biosynthesis (Table S2). Similarly, cluster 4 was distinguished by mutations affecting regulation of DNA and histone methylation (''DNA methylation,'' ''PRC2 methylates histones and DNA,'' and ''Nucleosome assembly''). Clusters 7–9 have in common mutations in extracellular, intracellular, and immune-related signaling pathways (see Figure 3B and Table S2). Cluster 2 had the highest pathway enrichment levels of the three, having mutations in hedgehog signaling, ''β-catenin degradation,'' ''cellular response to hypoxia,'' ''regulation of cell cycle'' and ''apoptosis'' among others.

Prior efforts to extract signatures from pan-cancer datasets met with difficulty in distinguishing tumor samples from tissue-specific -omics data signatures. Given our pathway-disruption based clustering, this raises

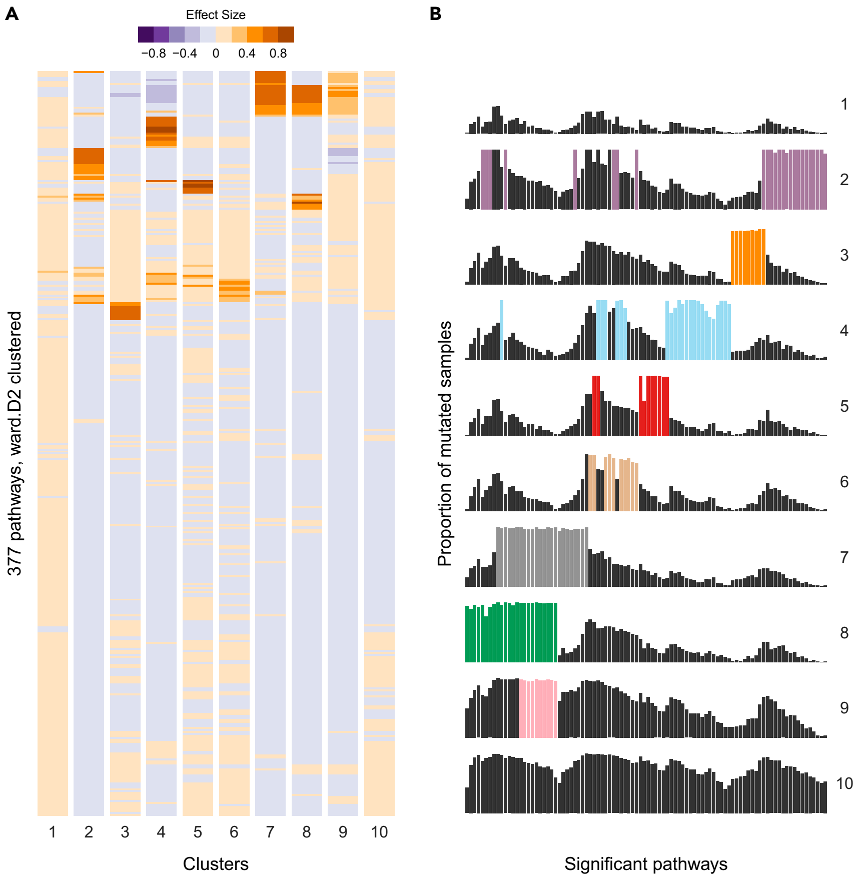


Figure 3. Pan-cancer enrichment of pathway disruptions

(A) Heatmap shows relative enrichment of each pathway (rows) within numbered clusters (columns). Color represents effect size as percent enrichment.

(B) Proportion of mutated samples in each significant pathway (columns; union set of pathways with effect size ≥0.30 in each cluster) within cluster (rows).

the question, are tumor phenotypes driven by common driver genes, ''silent'' tissue-specific effectors (i.e., too few samples to detect above statistical significance thresholds), or a combination of both? To answer this question, we compared top pathway genes for each cluster relative to TCGA background to find differentially mutated genes. We ranked odds ratios and selected the top ten enriched and depleted genes (p value <0.01) for each cluster (Figure 4; odds ratios plot). Clusters 7 and 8, which share multiple enrichment in signaling pathways, are largely driven by mutations in *PI3K* and its orthologs and *Ras* genes, respectively (compare *PIK3CA* and *KRAS* panels of Figure S4 and see interactive media from Data S1). Interestingly, cluster 9, which also shared multiple enrichment in signaling pathways with clusters 7 and 8, is enriched for both *PIK3CA* and *KRAS*. Clusters 3 and 5, defined by enrichment in metabolic pathways, had mutations in ribosomal proteins and nuclear pore complexes, respectively. Cluster 4 had mutations in nucleosome structural or subunit genes. Cluster 2 had mutations in proteasomal subunits involved in protein degradation. We also observed that genes enriched for one cluster are depleted from others (i.e., *TP53* is enriched in cluster 6, but depleted in cluster 7; *PIK3CA* is enriched in cluster 7, but depleted in clusters 3 and 8). Next, we investigated the proportion of samples per cancer type for the significant genes within a cluster (Figure 4; heatmap). Surprisingly, clusters were not predominated by one or more highly mutated genes across all cancers. Instead, when observing the mutation rate for these genes within samples that belong to a cluster, the mutation rate is heterogeneous across tumors by tissue origin (e.g., in cluster 4, CESC was enriched for *H2AFX*, OV was enriched for *HIST1H2BD*, and UCEC was enriched for *HIST1H2AC*). Even among the top enriched genes within clusters there is no global pattern, indicating that our clusters are not driven by individual genes, but rather networks as a whole. Taken together, our data identify a framework of cancer type-specific mutations associated with specific clusters.

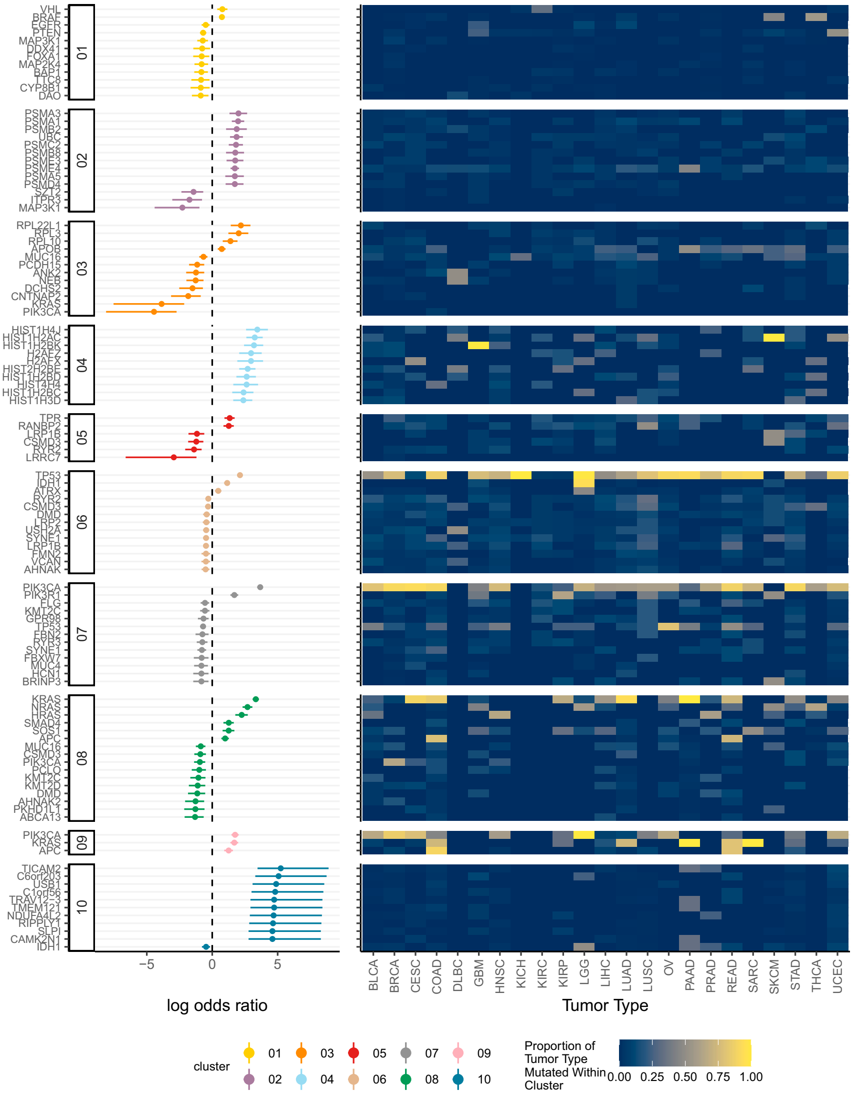


Figure 4. Gene level analysis reveals tissue-specific class signatures

Odds ratio plot; Column uses a logarithmic axis to represent odds ratio with a 95% confidence interval. Rows: significant genes from each cluster. Each cluster was compared against the background (all other clusters) to find differentially mutated genes. Significant genes (p value <0.01) were selected and limited to the top ten results for each cluster. Depleted significan genes (left). Enriched significant genes (right). Heatmap; Columns: cancer type. Rows: significant genes. The heatmap shows proportion of samples per cluster and cancer type mutated for each gene.

Enrichment of pathways in metastasis is cluster-specific

Because metastatic tumors are distributed across all ten clusters, we first compared them with non-metastatic tumors following the logic we used to investigate cluster specific enrichment. Using non-metastatic tumors as background, we found very low levels of enrichment (<10%) in a handful of pathways. We reasoned that the individual clusters might be too different to detect global metastasis enrichment signals given the small sample size (n = 215 metastatic tumor samples).

Therefore, we calculated cluster-specific enrichment in metastatic tumors and found a total of 31 enriched pathways (significant with enrichment score $\geq$0.3) across all clusters (Table 1). A number of enrichments represented pathways that were already shown to be enriched in non-metastatic samples of other clusters. For example, ''Signaling by *PTK6*'' is enriched in non-metastatic samples of cluster 8 (Table S2), but not in 7 and 9. This pathway is enriched in metastatic tumors of clusters 7 and 9 ($p<10^{-3}$, Table 1). This is also true of

Table 1. Cluster-specific enriched pathways (effect size ≥0.30) in metastasis

Pathway	Cluster	Effect size	−log(p)
Signaling by NOTCH1	2	0.34	2.11
Signaling by NOTCH2	2	0.40	2.82
Plasma lipoprotein clearance	2	0.31	1.94
Hedgehog 'off' state	3	0.31	2.07
Non-homologous end-joining (NHEJ)	3	0.33	2.53
Effects of PIP2 hydrolysis	4	0.49	2.04
Signaling by SCF-KIT	4	0.54	2.55
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.	4	0.60	2.41
Fc epsilon receptor (FCERI) signaling	4	0.55	2.12
Platelet calcium homeostasis	4	0.44	1.67
Energy dependent regulation of mTOR by LKB1-AMPK	5	0.31	1.80
GPCR ligand binding	5	0.35	1.93
Non-homologous end-joining (NHEJ)	5	0.59	3.70
DNA double-strand break response	5	0.34	1.63
RNA polymerase II transcription pre-initiation and promoter opening	5	0.35	2.36
Response to elevated platelet cytosolic Ca^{2+}	5	0.34	1.61
RNA polymerase II transcription initiation and promoter clearance	5	0.35	2.38
Protein ubiquitination	5	0.35	1.87
Glycogen metabolism	5	0.36	2.52
TCF dependent signaling in response to WNT	7	0.34	2.90
Neurexins and neuroligins	7	0.30	2.71
Signaling by PTK6	7	0.37	3.21
Erythropoietin activates RAS	7	0.31	3.28
Pre-NOTCH expression and processing	9	0.30	5.00
Transcriptional activation of mitochondrial biogenesis	9	0.35	1.80
Beta-catenin independent WNT signaling	9	0.33	5.00
Hedgehog 'off' state	9	0.35	1.83
G1/S DNA damage checkpoints	9	0.40	5.00
Signaling by PTK6	9	0.30	5.00
Erythropoietin activates RAS	9	0.38	2.37
Butyrate response factor 1 (BRF1) binds and destabilizes mRNA	10	0.30	4.00

"Erythropoietin activates *RAS*," which is enriched in non-metastatic tumors of cluster 8 (Table S2) and also in metastatic tumors of clusters 7 and 9. Cluster 4 metastases were enriched for "Fc epsilon receptor (FCERI) signaling," a key neutrophil pathway, which is also specific to clusters 2, 7 and 8 non-metastatic tumors. Thus, metastases pathways from one cluster are often enriched in non-metastases of other clusters.

Pathway disruption clusters vary in short-term prognosis of survival

If our clusters represent biological states distinct from tissue of origin, they may have different prognoses within cancer types or across all cancers. These analyses are limited by confounding factors of age, stage at diagnosis, sex, ethnicity, and tissue-specific disease progression. To explore these ideas, we used Bayesian inference to test models of survival using public longevity data from the CDC. The model estimates the effect of cancer type and cluster-specific cancer effects independently, resulting in a cancer and cluster-specific estimate of an effective age function and the aging rate multiplier, k.

To illustrate the model, we simulate survival for six related groups (Figure 5A). We start with a cohort of women aged 30 randomly selected from the US population, and show expected survival (light blue). We then use typical values from the posterior samples to set the effective age of that cohort to the effective age of a BRCA patient (yellow). Finally to illustrate how the model varies by effective age we add an additional artificial 10 years to the BRCA effective age (dark blue). Compare these results to a uniformly selected group of women ranging from 30 to 70 years old at random (green). Immediately, survival changes due to the mixture of ages, without malignancy. By computing the effective ages for this mixture of patients, we see that malignancy further reduces expected lifespan (pink). Adding an additional 10 years to effective age would reduce the lifespan even further (dark orange). Hence, simply due to the change in distribution of age-at-diagnosis, we expect two equally deadly cancers to have *different* survival curves. The model takes this into account and cancer aggressiveness is estimated accurately without the confounding effect of age through the effective age and aging rate parameters.

We found that cancer types, as expected, have a range of prognoses relative to the general population. In Figures 5B–5D we show how effective age baseline, and effective age rate (which modifies the baseline based on your actual age) as well as the k_{tis} which accelerates or decelerates further aging, varies considerably from cancer to cancer. For example a 20 year old diagnosed with THCA or PRAD would be treated similarly to a 40 year old based on the Effective Age parameter. It's noticeable that for every year beyond 20 in the PRAD diagnosis we add only an extra ~0.2 years (Effective Age Rate). The k_{tis} parameter represents an acceleration (greater than 1) or deceleration (less) of the further aging beyond diagnosis. For example, for BRCA patients each year past diagnosis adds about 1 year of risk, whereas for LGG patients it adds around 2, and for SARC patients it adds only 0.25 or so.

In Figures 5E–5H we see three particularly deadly cancers (Glioblastoma:GBM, Pancreatic:PAAD, Ovarian:OV), and one cancer where diagnosis hardly changes risk relative to background (Thyroid: THCA). Cancers with posterior probability for relative risk of less than 1 should be interpreted carefully. This Bayesian model is for a state of information. The information that a person is diagnosed with cancer may lead us to expect shorter survival than the general population of matched age, or longer survival than the general population of matched age. Shorter prognosis could result from cancer aggressiveness, injuring the body and causing death. By contrast, longer prognosis could result from the cancer being relatively mild, and therefore diagnosis could be an informational signal that the patient is health conscious, with the comparison group having more people whose cancers and other health issues go undiagnosed. It is important to note therefore that the diagnosis can increase our expectation of life relative to the comparison group, even if it decreases the expectation of life of the individual relative to the counterfactual where they did not have cancer.

Our estimates of tissue-specific cluster effects were for the Effective Age parameter. These cluster effects were multiplied by the tissue specific coefficient to form the full effective age. Priors for this parameter were relatively peaked around 1 reflecting the purpose of this coefficient as a multiplicative perturbation (Figure 6). A cluster-specific rate of 1 represents the typical rate for this tissue type. For several cancers (e.g., PRAD, kidney chromophobe (KICH), diffuse large B-cell lymphoma (DLBC), thyroid cancer (THCA)) the posterior estimates are largely indistinguishable from the prior, reflecting that either there were too few mortalities in the data to make an estimate (as expected for PRAD and THCA) or too few samples, period. We did not observe cluster-specific trends that held true across cancer types, which could result from different cancers having different standards of care for example. For some cancers, however, there is evidence that certain forms are more or less aggressive. PAAD class 8 seems to rise your effective age somewhat relative to other forms, as does LGG group 7, whereas BRCA group 5, COAD group 8, and LGG group 5 all seem somewhat lower risk than the other classes within those tissues.

To compare our pathway-based clustering to other published methods, we evaluated the performance in predicting time to death (tdeath) with other published classification models. We downloaded cluster membership from two studies conducted within the TCGA consortium (Hoadley et al, 2014, 2018). We generated tdeath predictions from uniform random p, by inverting the cumulative distribution function of our death

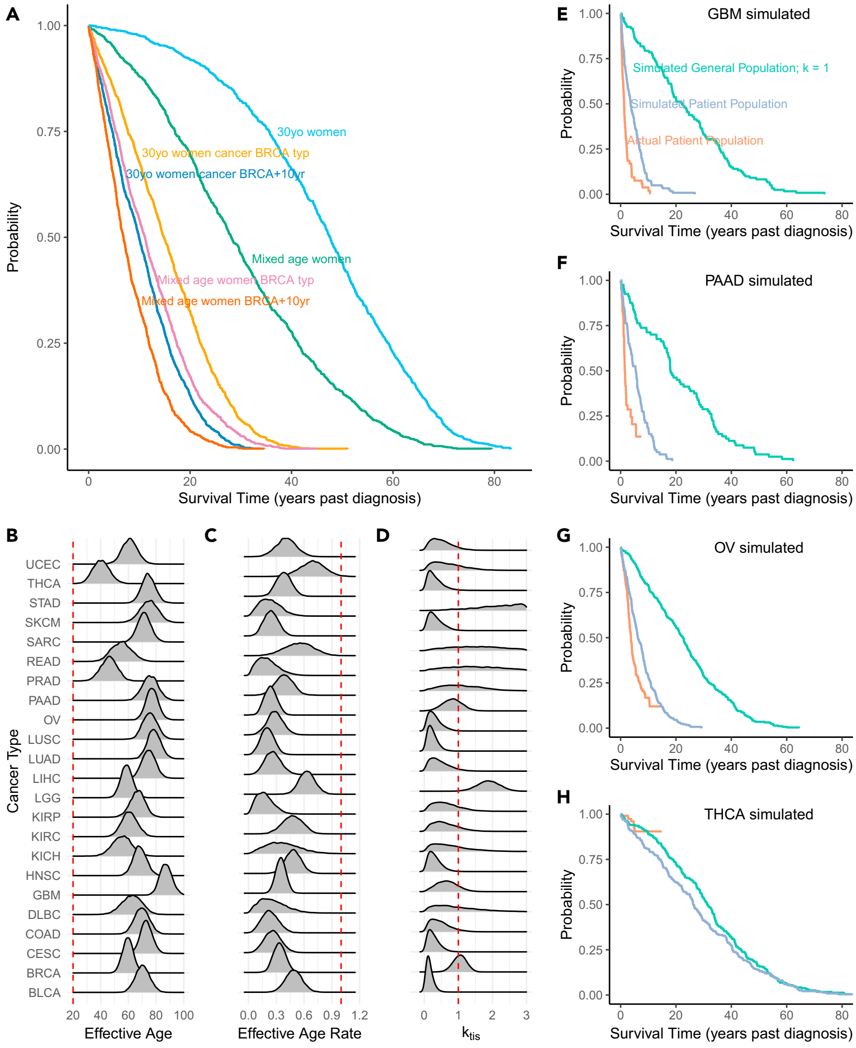


Figure 5. Comparison of different cancers, simulated and actual results

Comparison of survival in six simulated cohorts.

(A) A group consisting of randomly selected 30 year old women, unaffected (light blue) *vs.* cancer using the typical values from the posterior samples for a BRCA patient (yellow) and an artificial 10 years added to the effective age for a BRCA patient (dark blue), another group consisting of uniformly selected 30–70 year-old women at random, unaffected (green) and cancer without (pink) and with (dark orange) the artificial 10 years.

(B) Posterior probability density of the effective age baseline (A_0) by cancer type. Red dashed line represents age for unaffected individuals.

(C) Posterior probability density of effective age rate (A_r) by cancer type. Red dashed line represents the effective age rate for unaffected individuals.

(D) Posterior probability density of the k_{tis} multiplier. Values greater than 1 correspond to decreased age-independent life expectancy. Red dashed line represents k_{tis} for unaffected individuals.

(E–H) Comparison of our model to the actual survival for the patients in our dataset. We compare a random group with correct age distribution for the given cancer (turquoise) to the model predicted survival for the given average effective age (blue) and the actual survival for the patients in our dataset (orange). Differences with actual data are likely due to a mixture of cluster identities which we exclude from these simulations for simplicity.

probability function. Using the actual tdeath from TCGA as the reference, we show that our pathway-based clustering (referred to as kNN) had slightly better performance compared to COCA (Hoadley et al., 2014), iCluster (Hoadley et al., 2018) and tissue-of-origin (tis). (Figure S5A). In addition, we computed the entropy of the error distribution in predictions (Figure S5B). All models had similar performance with kNN performing slightly better as indicated by a slightly smaller error entropy.

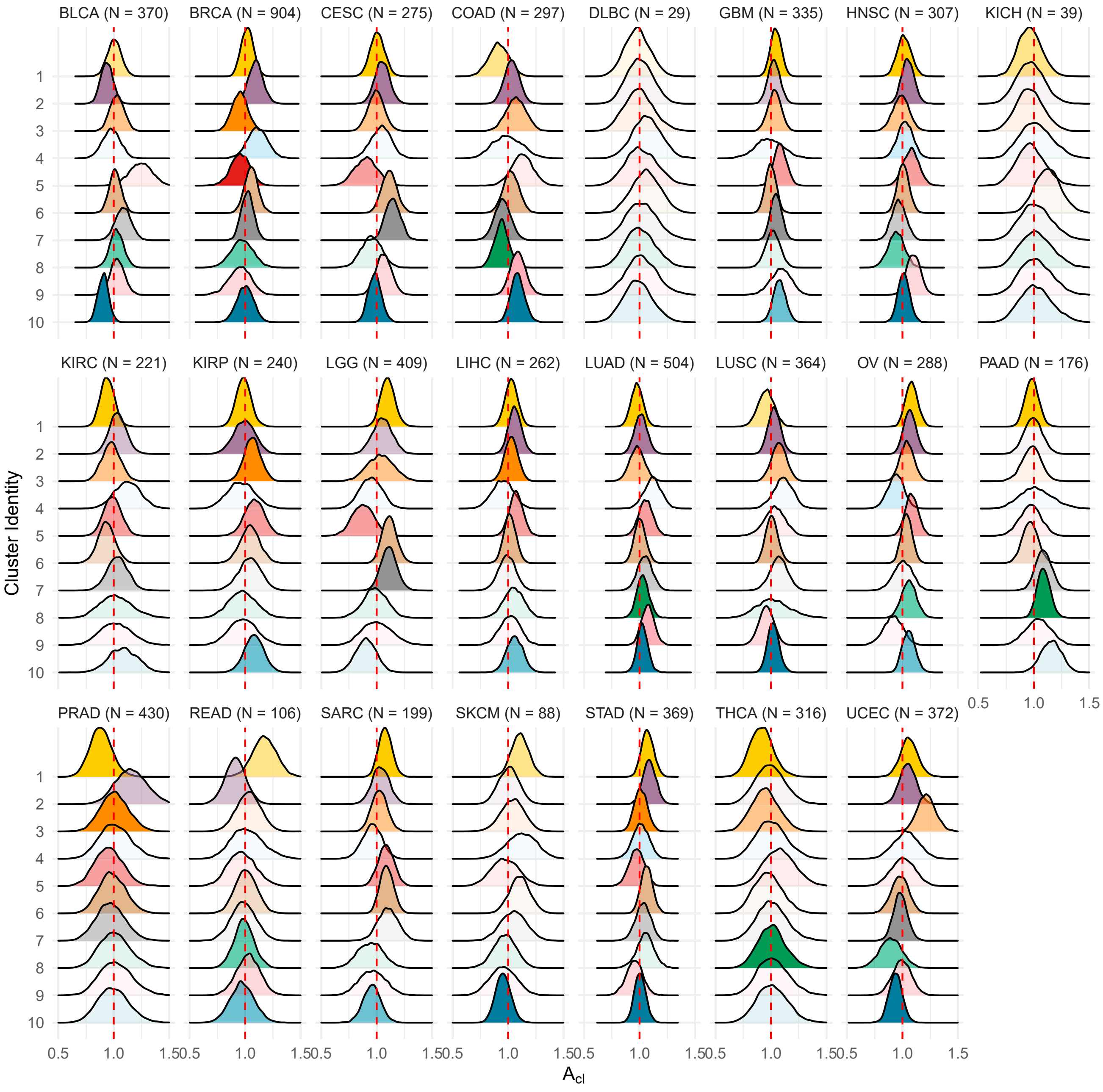


Figure 6. Cluster specific A_{cl} values for each cancer type

The overall effective age value is the product of $A_0 x A_{cl} + A_r x(A^1 – 20)$. The cluster-specific A represents relative aggressiveness of each cluster within cancer type. Alpha transparency (α) is set to reflect varying confidence in the posterior distribution when the number of observations is N < 10, α = 0.1; N < 25, α = 0.5; N ≥ 25 α = 1.0.

DISCUSSION

Classification of tumors independent of tissue-of-origin

One of the biggest hurdles in cancer research is the sparsity of data; ~20,000 protein-coding genes is comparable with the number of tumor samples, even with multiple mutations per sample. We sought to simplify the problem by employing a "knowledge-base driven analysis" (Khatri et al., 2012), investigating cancer as a disease of basic cellular and biochemical pathways. We accomplished this by translating gene-level mutations into pathway level disruptions. Our approach differs from previously described methods (Creixell et al., 2015) in that we chose to focus on the pathway (defined in the methods) as the unit of disruption instead of the gene, where individual mutations may be sufficient to alter pathway activity.

To our knowledge, this approach has not previously been attempted despite its relative simplicity. We limited our analysis to mutations with likely deleterious effects in genes that are actively expressed in each cancer type, thus avoiding bias from transcription coupled repair. Our method of filtration differs from the "rank-and-cut" method of (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020) but represents a reasonable attempt to account for the same biases. We also restricted our analysis to biochemical pathways, excluding curated gene sets related to diseases, syndromes, or classes of proteins with shared catalytic activity or conserved domains which are potentially problematic (Khatri et al., 2012). We chose this approach to limit redundancy and exclude biologically unrelated collections of genes.

The hypothesis that cancer results from dysfunction in basic cellular processes common to eukaryotic cells was introduced and later expanded on in a pair of essays by Hanahan and Weinberg (Hanahan and Weinberg, 2000, 2011). An alternative hypothesis is that every tumor belongs to one of a large number of syndromes which are unique to each tissue-of-origin, that share some mechanisms and treatment strategies. Recent publication of TCGA consortium papers present a view largely, and surprisingly, consistent with this latter hypothesis (Hoadley et al, 2014, 2018). Perhaps owing to the intractable complexity of genomics, proteomics, and patient metadata in all its forms, the inescapable conclusion thus far is that tissue-of-origin remains the most important driver of tumor characteristics at every scale and by every measure. Our observations contrast with this view, and instead support an interpretation of publicly available data in which all tumors manifest one of a limited number of phenotypes resulting from disruptions of basic pathways.

We attempted to account for our clusters in terms of more trivial explanations. For example, it could be that the clusters are consistent with disease progression. We were unable to identify any such trend in the number of mutations, the relative staging or metastasis; and each cluster instead was associated with unique combinations of pathways. Some cancer types are unevenly distributed among the clusters, though we could not identify any cancers that were exclusive to a single cluster. Only cancer types with the fewest samples were found to be absent from one or more clusters at all. Additionally, we annotated the clusters with patient attribute values (e.g., age of diagnosis, gender, and race) and were unable to identify any trends.

Somewhat surprisingly to us, this finding extends to histological subtypes of breast, head and neck cancers, leukemias, *etc.* This result implies that histological subtypes could reflect differences in cell-of-origin, rather than fundamental differences in cancer phenotype. The four major subtypes of breast cancer correspond to histological and molecular expression profiles that define them and how they respond to experimental stimuli (Prat and Perou, 2010). It has been hypothesized that differences in molecular regulators of development in precursor cell types present in breast epithelium drive histological phenotypes (Skibinski and Kuperwasser, 2015; Zhang et al., 2017). Consistent with this view we found that breast tumor samples of the Luminal A subtype were heavily biased toward membership in clusters 1 and 7, and basal tumor samples were biased toward cluster 6. However, both subtypes also contained samples in every other cluster (without exception), and Luminal B and Her2 positive samples are distributed across clusters. Our interpretation of these data is that inherited cell-of-origin signatures could predispose certain precursor cells within the breast epithelium to forming tumors of one cluster or another but are not determinative. This view is compatible with the previously stated hypothesis but opens the way for a more granular view of individual tumors.

It would be surprising if we did not observe bias for some cancers and subtypes amongst our classes, because some treatment regimens have greater efficacy for patients of a given cancer or histological subtype (Prat and Perou, 2010). Nonetheless, the basis for some tumors being treatment-refractory in spite of receiving the standard of clinical care for diagnostic markers remains elusive. Doubtless some of this is due to chance events, as tumors can metastasize and remain dormant years before they are detected at distal sites, or resistant clones may have already arisen at undetectable levels (Hanahan and Weinberg, 2011), but our analysis suggests the possibility of identifying more informative molecular, histological or cellular subtypes that could form a basis for future stratification of patients into different precision treatment regimens.

Tissue specific manifestation of pathway-centric disruptions

Our results illustrate how unique combinations of mutations in pan-cancer driver genes with tissue-specific pathway disruptions result in common categories when viewed at the level of the pathway knowledge-base.

Top cancer driver genes (e.g., *PIK3CA* and *TP53*) are found in most of the clusters, in spite of the fact that they contribute to many cluster-specific pathways. This can only be explained by less commonly mutated genes complementing the unique combination of driver genes in each tumor, and we speculate that many of these less commonly mutated genes are sensitive to increased mutation frequency in different tissues. For example, the most frequently mutated genes (*TP53*, *PIK3CA*, and *MUC16*) in cluster 2 were also frequently mutated in several other clusters. Yet, pathway enrichment in cluster 2 was defined by less commonly mutated genes in the proteasomal degradation pathway (e.g., *PSMA3*, *PSMA1*, or *PSME4*). The ubiquitin proteasome system maintains cellular protein homeostasis by regulating protein turnover. Pathway mutations found in cluster 2 (e.g., "B-catenin degradation," "cellular response to hypoxia," "regulation of cell cycle," and *etc*) covered several aspects of the cancer hallmarks described by Hanahan and Weinberg. These hallmarks included "sustaining proliferative signaling," "evading tumor suppressors," and "resisting cell death" (Bhattacharjee et al., 2014). Cluster 3 was defined by less commonly mutated genes related to ribosomal subunit. Pathway enrichment for cluster 3 included dysregulation in "rRNA processing in the nucleus and cytosol" and consequently, pathways involved in eukaryotic translation and nonsense mediated decay. Mutations in the ribosomal subunit have been associated with selective translation of mRNA to upregulate proliferation (Kampen et al., 2019). Pathway enrichment in cluster 3 consists of hallmarks relating to "sustaining proliferative signaling," "evading tumor suppressors," "resisting cell death," "deregulating cellular energetics," and "genome instability and mutations." Cluster 4 was defined by less commonly mutated genes related to nucleosome structure that resulted in pathway enrichment in DNA methylation and DNA double-strand break repair pathway. Cluster 4 was also enriched for "telomere maintenance," a mechanism contributing to cancer progression by counteracting telomere shortening. Mutations in the nucleosome as found in cluster 4 have been proposed to hit all attributes of hallmarks of cancer. Cluster 7 and 8 shared multiple enrichment in signaling pathways related to the hallmarks, showing that some clusters share common driver mechanisms: "sustaining proliferative signaling," "evading tumor suppressors," "resisting cell death," and "inducing angiogenesis." These clusters were strongly defined by mutations in *PI3K* or *RAS*, respectively, suggesting a shared etiology in the EGFR pathway for which these genes represent alternative intracellular signaling mechanisms. Taken together, our clusters highlight the similarities and differences at a gene and pathway level between and within different cancer types – uncovering the molecular mechanism in cancer. In the future, we hope to explore the defining features of the clusters by connecting the pathway mutations to cancer hallmarks.

Estimates of survival reveal pathway-dependent differences

By modeling CDC longevity data as a baseline risk function we showed that each clusterexhibits cancer-type specific effects on survival expectancy. However, considering that within each cancer type there are different clinical standards of care, and even within classes of drugs the preferred treatment can vary between cancers, it makes sense that we observe tissue-specific cluster effects. Contrast ovarian *vs.* breast cancer, which are both hormonally driven cancers, for example. Ovarian cancer has but one main treatment axis, platinum, whereas breast cancer patients have a variety of treatment regimens based on molecular subtype and other factors. Unfortunately, given the diversity of drug classes and treatments, we lack sufficient power to explore these variables in the TCGA data. It is our hope that future studies will help to distinguish between treatment-specific effects on survival given different pathway disruption clusters.

Implications for the evolution of cancer

Our findings imply separate processes in the etiology of cancer that can be broadly thought of as general cancer promoting, cluster-specific mutations and metastasis. General cancer promoting processes include genome stability and immortality, as ''enabling characteristics'' of the cancer phenotype (Hanahan and Weinberg, 2011). Such pathways are disrupted in most clusters and are frequently the result of aberrations involving common driver genes such as BRCA1/2, MMR genes, mitotic checkpoints, cohesion complexes, *etc.* Cluster-specific evolution must involve the acquisition of disruptions to pathways that may individually be harmful (e.g., highly proliferative cells are more likely to senesce) but together produce more specialized cancer phenotype and increased fitness. Importantly, our observations do not imply the order in which these mutations should accumulate. This could be addressed in a future study by evolutionary analysis of clonality, drawing inference from variant allele frequencies as in Gerstung et al. (2020). However, since many of the genes in the non-cluster-specific pathways involve the known driver genes, it is reasonable to surmise that these mutations promote or enable acquisition of cluster-specific defects via random mutation and natural selection, thus producing the clusters we observed. In support of this, the pan-cancer analysis of whole genomes consortium (PCAWG) found that oncogenic driver mutations are highly enriched in early arising clones, whereas later arising clones have much greater diversity in driver mutations (Gerstung et al. 2020). Moreover, driver genes that are known to be responsible for discrete mutation signatures such as *APOBEC*, *BRCA1* and *BRCA2* produce mutational hotspots reflecting varying selective pressures in different tissues (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020).

One of the drawbacks of bulk tumor whole genome sequencing data is the problem of tumor heterogeneity. Consortium samples are likely to contain contamination from support tissue, stroma, inflammatory cells, immune cells of the innate and adaptive immune systems, and all potentially harboring cancer supporting mutations (Tripathi et al., 2012). We think it will be instructive to explore these ideas in the context of single cell experiments.

On metastasis as a convergence of phenotypes

We report that enrichment in metastatic tumors across all clusters yielded generally lower effect sizes and larger p values than the cluster specific analysis, suggesting that signal is diluted when clusters are pooled, and supporting the view that metastasis has cluster-specific requirements. Since the number of metastatic samples is relatively low, this part of our analysis is likely underpowered and subject to expanded analysis with larger cohorts. The fact that most metastatic enrichment is cluster-specific and has a tendency to overlap with cluster-specific pathways from non-metastatic tumors of neighboring clusters suggests that newly acquired mutations result in similar clusters converging on one or more deadly phenotypes with critical features of end-stage cancer. Thus, even at relatively low power, our analysis of metastasis uncovered differences between tumors against the noisy backdrop of tissue specific profiles. As a major caveat to our pan-cancer analysis, we acknowledge that many therapies *do* target highly-specific driver genes, markers, and signaling pathways (e.g., *TP53*, *EGFR*, or *HER2*), but understanding the broader context of the genetic background and pathway vulnerability of tumors containing such markers may aid in creating smarter combination therapies. We submit that when we discover the requirements of each cluster with respect to pathway disruptions and metastasis we may be able to target them therapeutically and prevent further adaptation.

Limitations of the study

We must remark on the limitations of our work exemplified in clusters 1 and 10, for which we did not find many distinctive associations with pathways. Cluster 1 had a relatively low proportion of mutated pathways, although it is broadly enriched in many of the same tumor-promoting pathways common to other clusters. In addition, our data show clearly that this cluster is as likely to contain stage IV metastatic tumors as it is to contain those of stage I. This cluster likely represents a group of tumors with aberrations in methylation, copy number, or other structural variants. Consistent with this, kidney chromophobe and thyroid cancers have high proportions of structural variations *vs.* other variant types (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020) and are heavily skewed to cluster 1 membership. Likewise, cluster 10 represents a group of hyper-mutated tumors that harbor so many mutations that virtually no pathway is unaffected. It seems likely that a significant fraction of the "mutant" samples for each pathway are burdened with excess passenger mutations. This could be addressed with more sophisticated filtering of likely passenger mutations (e.g., Jaganathan et al., 2019; Sundaram et al., 2018). In the future, we hope to incorporate these other data into a comprehensive pathway-centric analysis as we have done here for point mutations and indels.

The outcome of our model comparison with other published classification models in predicting tdeath showed kNN performing slightly better. However, for the purpose of predicting lifetime within this dataset, all models are practically equivalent. One possible explanation for this is the cluster identity does not provide much new information from tissue-of-origin. However another possibility is that because treatment is not specialized to cluster, the treatments are not taking advantage of potential differences between clusters which might dramatically change survival. Most of the cluster-specific rate (A_{cl}) peaked around 1 for all models – leaving the remaining variables (A_0, A_r, and k_{tis}), which are tissue-specific, to compute tdeath. Although our study focused on somatic mutations, the incorporation of other types of omics data may help differentiate the classes within cancer type. In addition, more work in solidifying the pathway model into cancer hallmarks could provide biologically meaningful distinction of the clusters. Finally, if treatments can be specialized to biological information encoded in clusters, perhaps this could drive dramatic improvements in survival among some clusters in the future.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ Selection of pathways
  - ○ Filtering genes
  - ○ Clustering
  - ○ Survival
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Class and stage-specific enrichment calculations

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103084.

ACKNOWLEDGMENTS

We wish to thank David Van Valen, Kate Lawrenson, Simon Knott, and Megan Hitchins, for critical comments on this manuscript, and Tim Triche Jr for suggested improvements to the preprint. We also thank Ivetth Corona for early discussions and feedback. The graphical abstract was created by BioRender.com. Research reported in this publication was supported in part by Graduate Research Education at Cedars-Sinai (P.N.), by a Cedars-Sinai Precision Health award (D.H.), by institutional commitments to D.H., and by the Office of the Director, National Institutes of Health under Award Number OT2OD026675 (D.H.).

AUTHOR CONTRIBUTIONS

P.N., S.G. and D.L. conducted the experiments. D.H. designed the experiments and wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 24, 2021
Revised: July 6, 2021
Accepted: August 30, 2021
Published: October 22, 2021

REFERENCES

Aitken, S.J., Anderson, F., Connor, P.O., Sundaram, V., Feig, C., Rayner, T.F., Lukk, M., Aitken, S., Luft, J., Kentepozidou, E., et al. (2020). Pervasive lesion segregation shapes cancer genome evolution. Nature *583*, 265–270. https://doi.org/10.1038/s41586-020-2435-1.

Bhattacharjee, P., Mazumdar, M., Guha, D., and Sa, G. (2014). Ubiquitin–proteasome system in the hallmarks of cancer. In Role of Proteases in Cellular Dysfunction (Springer). https://doi.org/10.1007/978-1-4614-9099-9_9.

Chang, W., Cheng, J., Allairem, J.J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2021). Shiny: Web Application Framework for R.

Colaprico, A., Olsen, C., Bailey, M.H., Odom, G.J., Terkelsen, T., Silva, T.C., Olsen, A.V., Cantini, L., Zinovyev, A., Barillot, E., et al. (2020). Interpreting pathways to discover cancer driver genes with moonlight. Nat. Commun. *11*, 69. https://doi.org/10.1038/s41467-019-13803-0.

Colaprico, A., Silva, T., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T., Malta, T., Pagnotta, S., Castiglioni, I., et al. (2015). TCGAbiolinks: an r/bioconductor package for integrative analysis of TCGA data. Nucl. Acids Res. *44*. https://doi.org/10.1093/nar/gkv1507.

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., et al. (2015). Pathway and network analysis of cancer genomes. Nat. Methods *12*, 615–621. https://doi.org/10.1038/nmeth.3440.

Cummings, B.B., Karczewski, K.J., Kosmicki, J.A., Seaby, E.G., Watts, N.A., Singer-Berk, M., Mudge, J.M., Karjalainen, J., Satterstrom, F.K., O'Donnell-Luria, A.H., et al. (2020). Transcript expression-aware annotation improves rare variant interpretation. Nature *581*, 452–458. https://doi.org/10.1038/s41586-020-2329-2.

Davies, A.A., Masson, J.-Y., McIlwraith, M.J., and ZStasiak, A. (2001). Role of BRCA2 in control of the RAD51 recombination and DNA repair protein. Mol. Cell *7*, 273–282. https://doi.org/10.1016/S1097-2765(01)00175-7.

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. Genome Res. *22*, 1589–1598. https://doi.org/10.1101/gr.134635.111.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Rosebrock, D., Mitchell, T.J., Rubanova, Y., Anur, P., Yu, K., et al. (2020). The evolutionary history of 2,658 cancers. Nature *578*, 122–128. https://doi.org/10.1038/s41586-019-1907-7.

Gudmundsson, J., Johannesdottir, G., Bergthorsson, J.T., Arason, A., Ingvarsson, S., Egilsson, V., and Barkardottir, R.B. (1995). Different tumor types from BRCA2 carriers show wild-type chromosome deletions on 13q12–q13. Cancer Res. 55, 4830–4832.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674. https://doi.org/10.1016/j.cell.2011.02.013.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *10*, 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9.

Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell *173*, 291–304.e6. https://doi.org/10.1016/j.cell.2018.03.022.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., Zhang, J., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929–944. https://doi.org/10.1016/j.cell.2014.06.049.

Horn, H., Lawrence, M.S., Chouinard, C.R., Shrestha, Y., Hu, J.X., Worstell, E., Shea, E., Ilic, N., Kim, E., Kamburov, A., et al. (2018). NetSig: network-based discovery from cancer genomes. Nat. Methods *15*, 61–66. https://doi.org/10.1038/nmeth.4514.

Iranzo, J., Martincorena, I., and Koonin, E.V. (2018). Cancer-mutation network and the number and specificity of driver mutations. Proc. Natl. Acad. Sci. U S A *115*, E6010–E6019. https://doi.org/10.1073/pnas.1803155115.

Jaganathan, K., Panagiotopoulou, S.K., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting splicing from primary sequence with deep learning. Cell *176*, 535–548. https://doi.org/10.1016/j.cell.2018.12.015.

Jiang, L., Zheng, J., Kwan, J.S.H., Dai, S., Li, C., Li, M.J., Yu, B., To, K.F., Sham, P.C., Zhu, Y., et al. (2019). WITER: a powerful method for estimation of cancer-driver genes using a weighted iterative regression modelling background mutation counts. Nucl. Acids Res. *47*, e96. https://doi.org/10.1093/nar/gkz566.

Kampen, K., Sulima, S., Vereecke, S., and De Keersmaecker, K. (2019). Hallmarks of ribosomopathies. Nucl. Acids Res. *48*, 1013–1029. https://doi.org/10.1093/nar/gkz637.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339. https://doi.org/10.1038/nature12634.

Khatri, P., Sirota, M., and Butte, A.J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. *8*, e1002375. https://doi.org/10.1371/journal.pcbi.1002375.

Kim, H., and Kim, Y.-M. (2018). Pan-cancer analysis of somatic mutations and transcriptomes reveals common functional gene clusters shared by multiple cancer types. Sci. Rep. *8*, 6041. https://doi.org/10.1038/s41598-018-24379-y.

Kumar, R.D., Searleman, A.C., Swamidass, S.J., Griffith, O.L., and Bose, R. (2015). Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. Bioinformatics *31*, 3561–3568. https://doi.org/10.1093/bioinformatics/btv430.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501. https://doi.org/10.1038/nature12912.

Leiserson, M.D., Vandin, F., Wu, H., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., et al. (2014). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat. Genet. *47*, 106–114. https://doi.org/10.1038/ng.3168.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an r package for multivariate analysis. J. Stat. Softw. *25*, 1–18. https://doi.org/10.18637/jss.v025.i01.

Martincorena, I. (2019). Somatic mutation and clonal expansions in human tissues. Genome Med. *11*. https://doi.org/10.1186/s13073-019-0648-4.

McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. J Open Source Softw *2*, 1–2. https://doi.org/10.21105/2Fjoss.00205.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2020). UMAP: Uniform Manifold Approximation and Projection*3*, 29th (J Open Source Softw), pp. 1–2. https://doi.org/10.21105/joss.00861.

Mourikis, T.P., Benedetti, L., Foxall, E., Temelkovski, D., Nulsen, J., Perner, J., Cereda, M., Lagergren, J., Howell, M., Yau, C., et al. (2019). Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. Nat. Commun. *10*, e3101. https://doi.org/10.1038/s41467-019-10898-3.

Moynihan, M.E., Chiu, J.W., Koller, B.H., and Jasin, M. (1999). Brca1 controls homology-directed DNA repair. Mol. Cell *4*, 511–518. https://doi.org/10.1016/S1097-2765(00)80202-6.

Park, C., Qian, W., and Zhang, J. (2012). Genomic evidence for elevated mutation rates in highly expressed genes. EMBO Rep. *13*, 1123–1129. https://doi.org/10.1038/embor.2012.165.

Prat, A., and Perou, C.M. (2010). Deconstructing the molecular portraits of breast cancer. Mol. Oncol. *5*, 5–23. https://doi.org/10.1016/j.molonc.2010.11.003.

R Core Team (2020). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

RStudio Team (2020). RStudio: Integrated Development Environment for r. RStudio (PBC).

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W.K., Luna, A., La, K.C., Dimitriadoy, S., Liu, D.L., Kantheti, H.S., Saghafinia, S., et al. (2018). Oncogenic signaling pathways in the cancer genome. Cell *173*, 321–337.e10. https://doi.org/10.1016/j.cell.2018.03.035.

Skibinski, A., and Kuperwasser, C. (2015). The origin of breast tumor heterogeneity. Oncogene *34*, 5309–5316. https://doi.org/10.1038/onc.2014.475.

Stan Development Team (2020). RStan: The R Interface to Stan.

Sundaram, L., Gao, H., Padigepati, S.R., McRae, J.F., Li, Y., Kosmicki, J.A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J., et al. (2018). Predicting the clinical impact of human mutation with deep neural networks. Nat. Genet. *50*, 1161–1170. https://doi.org/10.1038/s41588-018-0167-z.

Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013). OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics *29*, 2238–2244. https://doi.org/10.1093/bioinformatics/btt395.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93. https://doi.org/10.1038/s41586-020-1969-6.

Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. Proc. Natl. Acad. Sci. U S A *113*, 14330–14335. https://doi.org/10.1073/pnas.1616440113.

Tripathi, M., Billet, S., and Bhowmick, N.A. (2012). Understanding the role of stromal fibroblasts in cancer progression. Cell Adh. Migr. 6, 231–235.

Vogelstein, B., and Kinzler, K.W. (2015). The path to cancer –three strikes and you're out. N. Engl. J. Med. *373*, 1895–1898. https://doi.org/10.1056/NEJMp1508811.

Zhang, M., Lee, A.V., and Rosen, J.M. (2017). The cellular origin and evolution of breast cancer. Cold Spring Harb. Perspect. Med. *7*, a027128. https://doi.org/10.1101/cshperspect.a027128.

Zhao, S., Liu, J., Nanga, P., Liu, Y., Cicek, A.E., Knoblauch, N., He, C., Stephens, M., and He, X. (2019). Detailed modeling of positive selection improves detection of cancer driver genes. Nat. Commun. *10*, 3399. https://doi.org/10.1038/s41467-019-11284-9.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA whole exome sequencing mutation data	Genomic Data Commons	https://dcc.icgc.org/repositories
ENSEMBL to All pathways	Reactome	https://reactome.org/download/current/Ensembl2Reactome_PE_All_Levels.txt
Complex to pathway	Reactome	https://reactome.org/download/current/Complex_2_Pathway_human.txt
Pathways hierarchy relationship	Reactome	https://reactome.org/download/current/ReactomePathwaysRelation.txt
Batch effects normalized mRNA	Pan-Cancer Atlas Hub	https://pancanatlas.xenahubs.net
United States Life Tables, 2017 – Volume 68, Number 7	National Center for Health Sciences	https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/NVSR/68_07/
Data generated for pathway-based clustering and analysis of TCGA mutation data	This paper	https://doi.org/10.5281/zenodo.5117696
Code for pathway-based clustering and analysis of TCGA mutation data	This paper	https://github.com/dennishazelett/TTmanu
Software and algorithms		
R 4.0.5	R Core Team, 2020	https://www.R-project.org
RStudio Server 1.4.1103	RStudio Team, 2020	https://www.rstudio.com
Bioconductor 3.11	Bioconductor	https://www.bioconductor.org/
TCGAbiolinks 2.18	Colaprico et al., 2015	https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html
rstan 2.21.2	Stan Development Team, 2020	https://cran.r-project.org/package=rstan
Shiny 1.6.0	Chang et al., 2021	https://cran.r-project.org/package=shiny
Other		
Data S1	This paper	https://junkdnalab.shinyapps.io/PANCAN_supplemental/

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to the Lead Contact, Dennis J. Hazelett (Dennis.Hazelett@cshs.org).

Materials availability

This study did not generate any new material.

Data and code availability

- This paper analyzes existing, publicly available data. These datasets are listed in the key resources table. Data generated in this study have been deposited at Zenodo and are publicly available as of the date of publication. DOI is listed in the key resources table.
- All code for producing the analyses and figures herein are included in this fully reproducible manuscript in R markdown format. R markdown files and all other models are available from our repository on the distributed version control site, Github, listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Selection of pathways

To understand the molecular mechanism of cancer at a pathway level, we used Reactome (https://reactome.org/), a knowledge-based pathway database. The mapping files of ENSEMBL genes to pathways, pathway hierarchy relationships, and complexes to top pathways were downloaded from https://reactome.org/download-data. Using these data, we imposed pathway criteria to define basic cellular processes and biochemical pathways: (1) human-derived pathways (*"HSA"*) (2) limited to grandchild node for each parent pathway (e.g. 'Beta-catenin independent WNT signaling' in 'Signal Transduction') (3) exclusion of pathways in the parent pathway: "Disease," "Muscle contraction," and "Reproduction" or pathway names that include any of the following keywords: "disease," "defect," "cancer," "oncogenic," "disorder," "mutant," "loss," "infection," "bacteria," or "listeria." While some of the excluded pathways have been shown to play an important role in cancer, they are highly specialized (e.g. "PI3K/AKT Signaling in Cancer"). Additionally, for most of the excluded pathways, a neutral version pathway of the pathway exists (e.g. "PIP3 activates Akt signaling"). Finally, we mapped the 18,577 Ensembl IDs from the TCGA dataset to the highly selected Reactome pathways. This operation produced a lookup table that consisted of 377 pathways mapped to 8,940 genes.

Filtering genes

We filtered likely erroneous mutations due to transcription coupled repair. Our approach was to determine the status of each gene (i.e. expressed or not expressed) in each tissue type in order to exclude lowly expressed genes. To do this, we obtained the TCGA RNA sequencing data adjusted for batch effect dataset (https://pancanatlas.xenahubs.net). Using the data, we removed the genes and tumor samples that were not included in our analysis, grouped the tumor samples by tissue type and computed the mean expression value for each gene. A minimum threshold of 10 transcripts per million was set for expressed genes based on an inflection point observed when plotting the mean expression values of genes ranked by expression in each tissue type. Genes that did not meet this threshold were considered not expressed. This operation produced a lookup table for gene expression status in each tumor sample for 18,127 genes.

Clustering

In order to classify tumors using this dataset, we used multiple correspondence analysis (MCA) (Lê et al., 2008). First, we determined the number of dimensions containing useful information by selecting the eigenvalue with the most explanatory power, using the average of 100 permutations of the data as baseline (Figure S6A). We then chose the maximum eigenvalue for which the p-value remained ≤ 0.05 (see cutoff in Figure S6B). Then we performed a UMAP analysis (McInnes et al., 2020), both in order to summarize the MCA graphically, and as a preprocessing step to boost the performance of density based clustering. The resulting map was notable for its lobed structure, with several reproducible projections regardless of random seed setting. A representative version of this 3D UMAP is shown in Figure S6C, rotated to enhance the visibility of the major features. Following this spatial mapping we attempted to define groupings of similar tumors within the spatial map using HDBSCAN, which performs hierarchical clustering and provides metrics of cluster stability and probabilities of cluster membership for each node (McInnes et al., 2017). However, HDBSCAN is sensitive to several parameters; key for our analysis are the minimum number of tumor samples in a cluster that capture the maximum number of tumor samples, measured by probability of membership of $\ge 5\%$ in at least one cluster. Thus, we created a score metric as the fraction of classified tumors with max probability $< 5\%$ in one cluster and chose a cluster size of 92 to minimize the score function (Figure S6D). HDBSCAN with these settings resulted in ten distinct high-density clusters which we then projected onto the UMAP (Figure S6E). This classified 6,038 out of 7,607 tumors but still resulted in a significant fraction of unclassified tumor samples. Since we ultimately wish to be able to classify any tumor using this scheme, we performed k Nearest Neighbors (kNN) analysis, which computes a similarity metric to every tumor in the set and then lets the *k* most similar tumors "vote" as to the identity of the query tumor sample based on their cluster labels. We set *k* to be the square root of the number of tumor samples (87). Using this method, we assigned cluster membership to the remaining tumors (Figure S6F and see interactive media from Data S1 (junkdnalab.shinyapps.io/PANCAN_supplemental/)).

Survival

We used Bayesian inference to explore the relative impact of cancer diagnosis on survival at the time of diagnosis. Our model assumes that at birth each person has a small initial probability per unit time to die, and that this probability grows exponentially in time at some baseline rate that matches the observed CDC data. We assume that at diagnosis this information will allow us to treat patients as if they were a different age, and with further time passing at a different rate, a self-similarity based model. The effective age at diagnosis is based on a baseline age for that tissue/class combination, plus a perturbation for each year over 20 years old at diagnosis. To fit the model, we constrain the baseline risks using the available CDC life tables (Figure S7). We also constrain the cancer specific parameters using the longevity data within the cancer tumor dataset. Our model estimates the effect of cancer type and class-specific cancer effects independently, resulting in a cancer and class specific estimate of the effective age.

Most TCGA cases are diagnosed late in life, so we fit a model that emphasizes accuracy in the right tail of the distribution. We modeled baseline longevity using a risk rate function. The risk rate of death at a time t days after the A birthday is given as:

$$R(t) = R_{20} exp\left(a(A-20) + D(t)ka\left(\frac{t}{365}\right)\right)$$

Time is split into two components. A represents the "effective age" of the patient at diagnosis, whereas t starts at 0 upon diagnosis and represents the days since diagnosis, a field recorded in the dataset. $D(t)$ is an indicator for whether diagnosis has occurred or not (before diagnosis $D(t) = 0$, after $D(t) = 1$). The model assumes that the risk of death is R_{20} at age 20 years and increases exponentially with a constant rate a up to the age at diagnosis. Risk of death assumes a new rate ka thereafter.

Given this risk per unit time, the probability of death at time $T+dT$ is the probability to survive to time T which is $(1-P(T))$ and then die in the remaining interval which is $R(T)dT$

$$\mathrm{d}P = (1-P(T))R(T)\mathrm{d}T$$

leading to the differential equation for the cumulative probability of death $P(T)$

$$\frac{\mathrm{d}P}{\mathrm{d}T} = (1-P(T)))R(T)$$

The solution of this ODE is:

$$P(T) = 1 - exp\left(-\int_0^T R(t)\mathrm{d}t\right)$$

And the density of deaths per unit time is the derivative with respect to T:

$$p(T) = R(T)exp\left(-\int_0^T R(t)\mathrm{d}t\right)$$

However, to model the risk of cancer effectively we found that it was necessary not to use the actual age A' but instead the "effective age" for that tissue/class type. Our effective age is calculated as:

$$A = A_0 A_{cl} + A_r(A' - 20)$$

where A_0 is the tissue specific effective age parameter (expressed in years), A_{cl} is a dimensionless multiplier for the given class, and A_r is the dimensionless age rate which determines how actual age A' affects effective age A using a linear perturbation. For a non-cancer patient, $A_0 = 20$, $A_{cl} = 1$, and $A_r = 1$ are the relevant benchmarks.

The $R(T)$ function has two important parameters, the a value which represents the background risk rates, and the k value, which is a function of cancer tissue type.

$$k = k_{tis}$$

The product $A_0 A_{cl}$ is obviously symmetric between the two factors. To disambiguate the meaning of the two, the prior for A_0 has high probability range from about 40 to 100 years, whereas the prior distribution of A_{cl} has peak probability at 1 and a relatively narrow width, due to the fact that it is a tissue specific multiplicative modifying factor. To understand the roles of these priors refer to (Figure S8).

Using this model, we split the data by Male/Female sex. Sex is known to be a risk factor for death, with males dying at slightly higher rates for all ages of interest in our dataset.

As an aside, we did not explicitly account for stage at diagnosis in our model for the following reasons. In the tumor sample data, stage at diagnosis is confounded with cancer type because some cancers are screened aggressively (e.g. colorectal and prostate cancers) while others are diagnosed typically after they become problematic for the patient's lifestyle (e.g. ovarian and pancreatic cancers). Secondly, such a model specification would suffer from added noise because the staging data are not well standardized across cancer types, have different criteria, and because it is unclear what the relationship between stage and advancement of disease is (for example some sub-stage 4 tumors are metastatic). To compound this latter issue, our tumor set is vastly under-powered given the uneven representation of stage across cancer types.

QUANTIFICATION AND STATISTICAL ANALYSIS

Class and stage-specific enrichment calculations

We calculated the enrichment of pathways in one set of tumors as the relative fraction of tumors (with estimated uncertainty) with a mutated gene in the pathway to all tumors, inclusive of the category of interest. To do this operation, we use the beta distribution to permute a posterior distribution on the fraction of tumors with a pathway mutated for each category based on the observed set, and compared this to the posterior obtained from the full set of tumors (all tumors) as the distribution of differences between all permuted samples. We considered a pathway enriched if the 95% range of credible differences thus obtained excludes 0 and the mean credible difference was greater than or equal to 30% enrichment, which excludes a large number of small differences that are not likely to be biologically relevant.