Title

iSubGen generates integrative disease subtypes by pairwise similarity assessment.

Journal

Cell Reports: Methods, 4(11)

Authors

Fox, Natalie
Tian, Mao
Markowitz, Alexander
et al.

Publication Date

2024-11-18

DOI

10.1016/j.crmeth.2024.100884

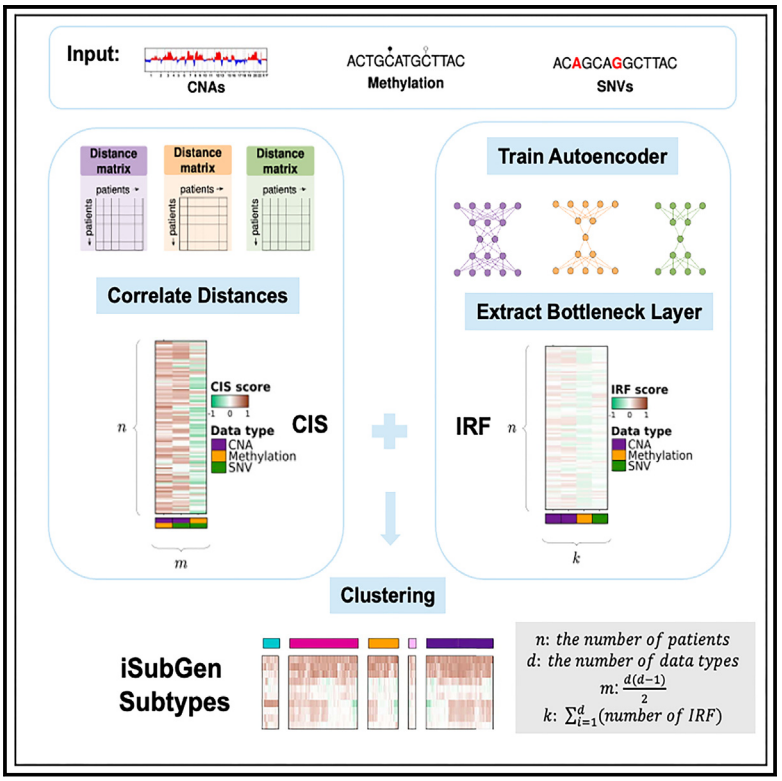
Peer reviewed

Cell Reports
Methods

Report

iSubGen generates integrative disease subtypes by pairwise similarity assessment

Graphical abstract



Highlights

- Creating disease subtypes informed by diverse data types is challenging
- We present iSubGen (integrative subtype generation) to address this problem
- Integrative similarity (CIS) score measures the connection between various data types
- iSubGen recapitulates known subtypes in cancers and handles missing data robustly

Authors

Natalie S. Fox, Mao Tian, Alexander L. Markowitz, Syed Haider, Constance H. Li, Paul C. Boutros

Correspondence

maotian@mednet.ucla.edu (M.T.), pboutros@mednet.ucla.edu (P.C.B.)

In brief

Fox et al. develop a deep-learning-based disease subtyping method, iSubGen (integrative subtype generation), which can seamlessly handle multi-omics data. iSubGen accounts for both individual data features and the inter-relationships between different data types, with the capability of handling missing data.

Fox et al., 2024, Cell Reports Methods *4*, 100884
November 18, 2024
https://doi.org/10.1016/j.crmeth.2024.100884

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Report

iSubGen generates integrative disease subtypes by pairwise similarity assessment

Natalie S. Fox,[1,2,3,4,5] Mao Tian,[2,3,4,*] Alexander L. Markowitz,[2,3,4] Syed Haider,[6] Constance H. Li,[1,2,3,4,5] and Paul C. Boutros[1,2,3,4,7,8,9,10,*]

[1]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada
[2]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA
[3]Institute for Precision Health, University of California, Los Angeles, Los Angeles, CA, USA
[4]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA, USA
[5]Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada
[6]The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK
[7]Department of Pharmacology and Toxicology, University of Toronto, Toronto, ON M5S 1A8, Canada
[8]Department of Urology, University of California, Los Angeles, Los Angeles, CA, USA
[9]Broad Stem Cell Research Center, University of California, Los Angeles, Los Angeles, CA, USA
[10]Lead contact
*Correspondence: maotian@mednet.ucla.edu (M.T.), pboutros@mednet.ucla.edu (P.C.B.)
https://doi.org/10.1016/j.crmeth.2024.100884

MOTIVATION Identifying disease subtypes is a strategy to address heterogeneity by identifying patient subgroups with more homogeneous presentation, progression, and response. Many studies have identified subtypes using single data types or by clustering groups of single data-type clusters; however, the best way to create subtypes by integration of diverse data types remains unclear. To address this, we created a multi-dimensional subtyping framework that incorporates two key innovations: a consensus integrative similarity score, which quantifies inter-relationships between different data types, and independent reduced features generated through deep-learning-based autoencoders, which standardize dimensionality across data types.

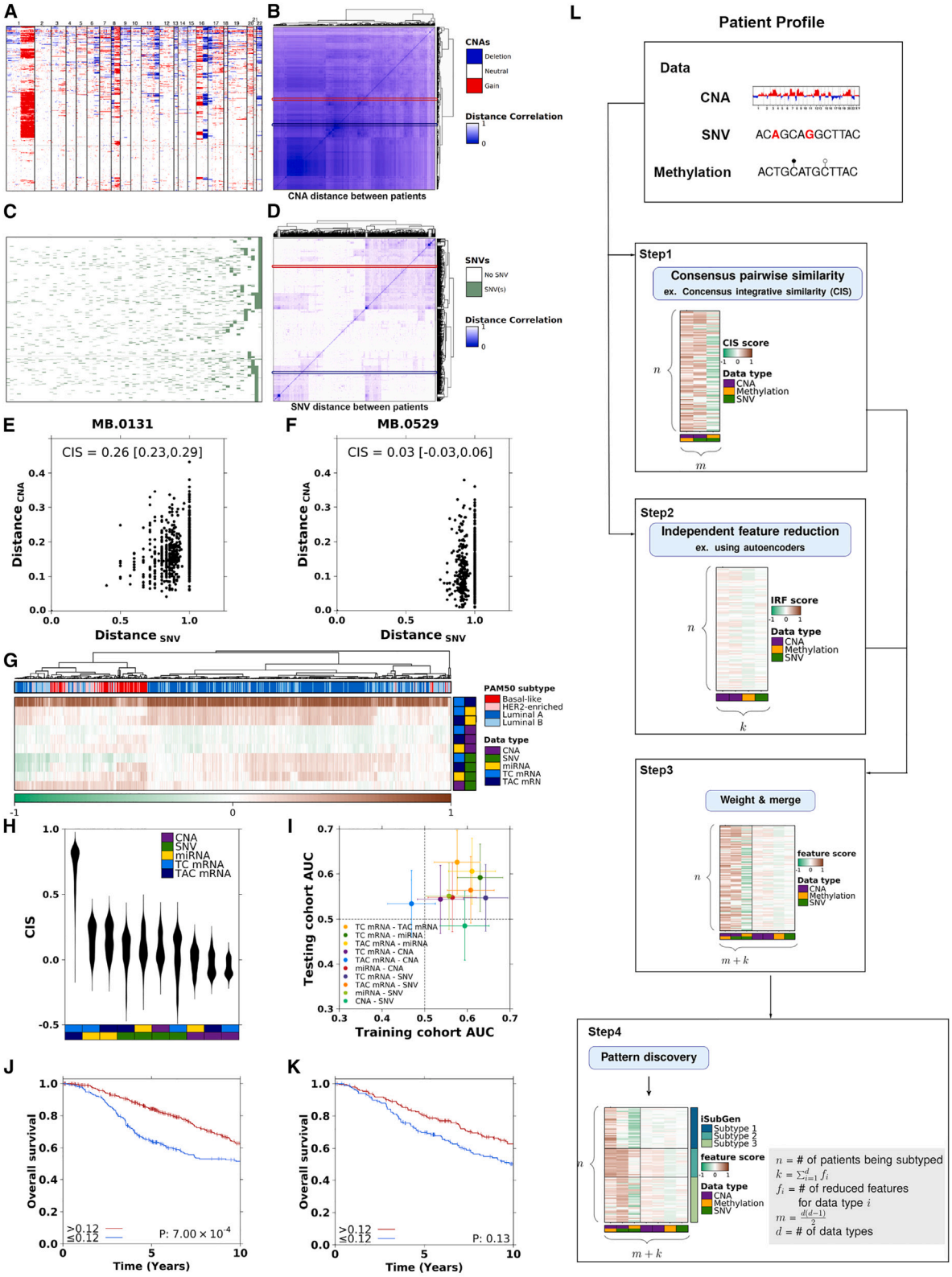
SUMMARY

There are myriad types of biomedical data—molecular, clinical images, and others. When a group of patients with the same underlying disease exhibits similarities across multiple types of data, this is called a subtype. Existing subtyping approaches struggle to handle diverse data types with missing information. To improve subtype discovery, we exploited changes in the correlation-structure between different data types to create iSubGen, an algorithm for integrative subtype generation. iSubGen can accommodate any feature that can be compared with a similarity metric to create subtypes versatilely. It can combine arbitrary data types for subtype discovery, such as merging genetic, transcriptomic, proteomic, and pathway data. iSubGen recapitulates known subtypes across multiple cancers even with substantial missing data and identifies subtypes with distinct clinical behaviors. It performs equally with or superior to other subtyping methods, offering greater stability and robustness to missing data and flexibility to new data types. It is available at https://cran.r-project.org/web/packages/iSubGen.

INTRODUCTION

Most diseases show substantial interpatient variability in presentation, progression, and response to treatment; this heterogeneity is a hallmark of cancer, autoimmune disorders, and neurological disorders, among many others.[1–6] Clinical heterogeneity in behavior often reflects common patterns of disease features, called subtypes, which can be important for clinical management by reducing the heterogeneity in presentation and progression.[2,7,8]

Disease subtypes play a particularly important role in cancer, where almost all tumors arise from a single cell, and features of that cell shape tumor initiation, progression, and evolution.[9,10] The location of the primary cancer lesion influences the types of interventions possible and their efficacies, leading cancers to be grouped clinically based on their tissue of origin. Individual tissues contain cells of different types and distinct gene-expression landscapes, and these evolve into cancers with distinct characteristics.[11] Further, cells of a single cell type can lead to different types of cancer based on the identity and timing of



(legend on next page)

driver mutations and on the microenvironmental pressures they experience during tumorigenesis.[10,12,13]

The variable, but repeatedly observed, evolutionary courses of cancers originating in a single anatomical location are termed "cancer subtypes." Historically, cancer subtypes have been defined histopathologically.[3–6] More recently, high-throughput molecular assays have discovered and defined subtypes.[7,14–17] Both approaches can identify groups of cancers with less heterogeneous prognoses and responses to treatment.[7,14,18,19] Subtypes can sometimes be discovered from a single data type,[7] but often cannot be precisely defined without considering multiple layers of biological information.[14]

The classical approach to subtype discovery is to apply unsupervised learning methods to a subset of input data that varies substantially between individuals. These input data can be binary (e.g., single-nucleotide variants [SNVs]), categorical (e.g., copy-number alterations [CNAs]), continuous (e.g., mRNA abundance), bounded continuous (e.g., methylation β values, ranging from 0 to 1), or have other distributional features. Many molecular data are gene based, but some represent processes such as pathway activity or trinucleotide mutational signatures.[20]

This classic approach has several limitations when applied to multiple data types simultaneously. First, standard unsupervised learning methods can produce artifactual results when applied to datasets with highly variable distributional features, often implicitly assigning heavier weights to data types with many features or larger numerical ranges. To address this, some integrative subtyping algorithms transform input into a latent variable space or use summary features from each individual data type.[21–23] Second, clinical practice routinely produces partial information, and most unsupervised learning methods struggle to accommodate large amounts of missing data.[24,25] Third, most methods do not exploit differential covariance or correlation across data types nor provide clear understanding of how each data type contributes to the final subtyping.

We created iSubGen (integrative subtype generation) to create subtypes by directly quantifying inter-relationships between different data types. iSubGen recapitulates known molecular and histological subtypes, robustly handles missing data, supports high subtype and feature numbers, and seamlessly integrates gene-based and non-gene-based features.

RESULTS

Development dataset

To develop iSubGen, we used the 1,991-patient METABRIC breast cancer dataset (European Genome-Phenome Archive: EGAS00000000083), which has clinical, CNA, SNV, microRNA (miRNA) abundance, and mRNA abundance data, with the latter computationally deconvolved into tumor cell (TC) and tumor-adjacent cell (TAC) components.[14,26–28] We initially focused on the 1,071 patients with complete data and split these into the 684-patient training cohort and 367-patient testing cohort as in the original publication.[14] Initial method development used the 684 patients in the training cohort with complete data.

Consensus integrative similarities

Typical approaches to subtype identification quantify the relationship between each pair of patients using a similarity metric. For an n-patient cohort, this information is encoded in an $n \times n$ similarity matrix, which can be clustered using unsupervised machine learning.[29] Thus, clustering of CNA profiles (Figure 1A) generates CNA subtypes (Figure 1B), and clustering of SNV profiles (Figure 1C) generates SNV subtypes (Figure 1D) in the METABRIC training dataset.

To integrate multiple data types into subtyping, there are two basic strategies. First, all data can be standardized to a common scale and a single metric applied to the appended matrix. Thus, for an n-patient dataset with m data types each having p_m features, this results in performing similarity calculations on an $n \times \Sigma p_m$ feature matrix, producing a final $n \times n$ similarity matrix. This approach intrinsically preferences data types with more features or larger values because they hold more weight in similarity calculations.[30] An alternative strategy instead analyzes each data type separately and relates the m separate $n \times n$ similarity matrices. For example, each data type can be clustered separately, after which the patient classifications from each data type can themselves be clustered.[11] This discretizes patient classifications and intrinsically weights each data type either equivalently if cluster number is held constant or as a function of cluster number if it is not.

To create a more flexible method of merging multiple data types, we directly reduced the pair of $n \times n$ similarity matrices

Figure 1. Integrative similarities

Pairwise integrative similarities in the training cohort of METABRIC breast cancer patients.

(A) CNAs with genes ordered by genomic position on the x axis and patients on the y axis. Gains are red and deletions are blue.

(B) CNA patient-by-patient similarity matrix using Jaccard distance as the similarity metric.

(C) SNVs for genes mutated in more than 13 patients. Genes (x axis) are ordered by mutation frequency. Patients are on the y axis.

(D) SNV patient-by-patient similarity matrix calculated using SNVs without patient recurrence filtering and Jaccard distances.

(E and F) Comparison of CNA and SNV similarities relative to patient MB.0131 (E) and MB.0529 (F). The CNA and SNV profiles for MB.0131 and MB.0529 are indicated in (A) and (B) by the boxes and arrows in red and blue, respectively. Jaccard distances are used for measuring similarity in both CNA and SNV. MB.0131 was randomly selected as an example of a patient with positive CIS. MB.0529 was randomly selected as an example of a patient with a CIS near zero.

(G) Patients grouped by clustering CISs.

(H) The distributions of CIS for each data-type pair.

(I) Area under the receiver operating characteristic curve for predicting overall survival at 5 years using CISs. Error bars represent the 95% confidence intervals.

(J and K) Overall survival differences for patients dichotomized using $CIS_{TAC\ mRNA\text{-}miRNA}$ at the maximum geometric mean of the true positive rate and the false positive rate in the training cohort (J) and using the training cohort threshold in the testing cohort (K). p values are from log-rank tests.

(L) Schematic overview of iSubGen with three data types as an example for n patients. Each data type was separately run through feature reduction and combined in pairs for comparison of the patient profiles using similarity measures. Output from feature reduction (n rows by k columns) and similarity comparison (n rows by m columns) were rescaled and reweighted, if necessary, and merged into a single matrix for unsupervised machine learning to create the final classifications. We used the autoencoder bottleneck layer for independent feature reduction and CISs as our pairwise similarity measures.

for two data types into a continuous value representing the similarity between any two patients' similarity profiles (Figures 1E and 1F). Thus, the two n-length similarity vectors for a single patient, one per data type, are collapsed into a single value. Here, we used Spearman's correlation to measure similarity. We used resampling to robustify this value, leading to a consensus integrative similarity (CIS) for each patient (Figures 1E and 1F). A vector of n CISs is created for each pair of data types, yielding an $n \times [m \times (m-1)/2]$ matrix encompassing the inter-relationships between data types for each patient. Figure 1G shows this matrix for 684 patients from the METABRIC dataset with simple unsupervised clustering applied to it. Luminal breast cancers cluster together and basal-like breast cancers cluster together, suggesting that CIS values can reflect disease biology.

CIS values are near zero for data types with independent (orthogonal) information, positive for data types with shared information, and negative when patients similar to one another in one data type are dissimilar in the other. In METABRIC, the median CIS across all data types was near zero (Figure 1H; median 0.06, range −0.38 to 0.88). The two data types that shared the most information were TC mRNA abundance and TAC (stromal) mRNA abundance (median $CIS_{TC\ mRNA\text{-}TAC\ mRNA}$ 0.77, range −0.07 to 0.88). The relationships between different types of information encapsulated in CISs were predictive of clinical features. In the training cohort, four of ten CISs predicted 5-year survival (area under the receiver operating characteristic [AUROC] > 0.6) without applying any statistical learning. This was validated in the 367-patient testing cohort (Figure 1I). For example, stronger associations between TAC mRNA and miRNAs were associated with improved overall patient survival (Figures 1J and 1K).

To further test the validity of CISs, we evaluated whether CIS constituting mRNA abundance retained key information such as mRNA-based subtypes of breast cancer (PAM50) and whether other CISs were also predictive of breast cancer molecular subtypes. Using the training and testing cohorts, we compared CIS distributions between PAM50 subtypes (Figure S1A). Almost all CISs differed among PAM50 subtypes (19/20, ANOVA $q < 0.01$). A random forest trained using CIS values predicted subtypes with AUROCs in the testing cohort ranging from 0.58 to 0.95 (Figure S1B). $CIS_{TC\ mRNA\text{-}miRNA}$ was the most important feature for the random forest luminal A classifier followed closely by $CIS_{TC\ mRNA\text{-}TAC\ mRNA}$ and $CIS_{TAC\ mRNA\text{-}miRNA}$ (Figure S1C). We also trained a random forest in The Cancer Genome Atlas (TCGA) breast cancer dataset to predict subtypes from CIS values. This achieved AUROCs ranging from 0.71 to 0.88 (Figure S1D). Different features were important for classification of each subtype (Figure S1E).

To determine whether CISs were associated with known subtypes in other cancers, we exploited pan-cancer TCGA data (Broad GDAC Firehose 2016-01-28 Release: https://gdac.broadinstitute.org) of 12 cancer types with six data types per patient (mRNA abundance, miRNA abundance, methylation, CNAs, SNVs, and SNV trinucleotide signatures). We created pan-cancer training and testing cohorts each comprising 1,709 patients. All CIS combinations in the training and testing cohorts distinguished cancer types (30/30 in both cohorts, ANOVA $q < 0.01$; Figure S1F). CIS distributions for some cancer types were bimodal, such as thyroid cancer (THCA) $CIS_{mRNA\text{-}SNV}$. Histopathological subtypes may cause this bimodality: in THCA, patients with tall cell thyroid cancer had higher $CIS_{mRNA\text{-}SNV}$ than those with follicular thyroid cancer ($p = 3.09 \times 10^{-3}$; Figure S1G). Bimodality and high variance in CIS across many cancers increases the chance of finding subpopulations/subtypes. Random forest classifiers trained on CISs predicted all cancer types with $AUROC_{testing\ cohort} > 0.9$ (Figure S1H). CISs vary in importance for predicting cancer types, with different CISs being important in distinguishing each cancer type (Figure S1I). For example, $CIS_{methylation\text{-}mRNA}$ was most important in identifying liver cancers (LIHC), $CIS_{mRNA\text{-}miRNA}$ for predicting kidney clear cell cancers (KIRC), and $CIS_{SNV\text{-}mRNA}$ for predicting kidney papillary cancers. Thus, CISs can distinguish histological cancer types and subtypes.

iSubGen framework and integrative subtyping

CISs capture the changing relationships between different types of data. To integrate them with information present in patterns of a single data type, we created a second set of engineered features. This feature set was generated by training an autoencoder for each data type, using its bottleneck layer as the set of independent reduced features (IRFs). iSubGen is thus a four-step subtype generation framework: consensus pairwise similarity construction (CIS generation), data-type independent feature reduction (IRF generation), weighting of features, and unsupervised machine learning (Figure 1L). The CIS values represent how different data types inter-relate, while the IRF values identify general patterns within each data type. A detailed schematic overview of the algorithm is online at the iSubGen GitHub repository (https://github.com/uclahs-cds/package-iSubGen) and in the package vignette (https://cran.r-project.org/web/packages/iSubGen/vignettes). This strategy helps to balance groups of engineered features so that their relative weights are not primarily a function of the total feature number. In step three of the subtyping framework, the user sets the weightings of CIS vs. IRF and merges the two feature sets to create the combined engineered feature matrix. This provides a parameterizable decision for users that can optimize based on internal features (e.g., cluster silhouette profiles) or external ones (e.g., separation of metadata). Finally, applying pattern discovery to the combined sets of engineered features generates the final iSubGen subtypes. Here, we performed pattern discovery using consensus clustering,[29] but iSubGen supports multiple algorithms at each step. For example, CISs can use different correlation metrics or mutual information, with or without subsampling.

Pan-cancer grouping discovery with iSubGen

To demonstrate how iSubGen combines CISs and IRFs to generate robust subtypes, we applied it to the pan-cancer cohort evaluated in Figures S1F–S1I. Using six data types (miRNA abundance, mRNA abundance, methylation, CNA, SNV, and trinucleotide signatures), we subtyped the two 1,709-patient subsets of 12 cancer types separately using iSubGen (Figures S2A–S2F and Table S1). Each subset was independently analyzed to evaluate iSubGen subtype consistency. Comparing the adjusted Rand index of the iSubGen clusters with TCGA cancer types, we identified 14 iSubGen groupings

CellPress
OPEN ACCESS

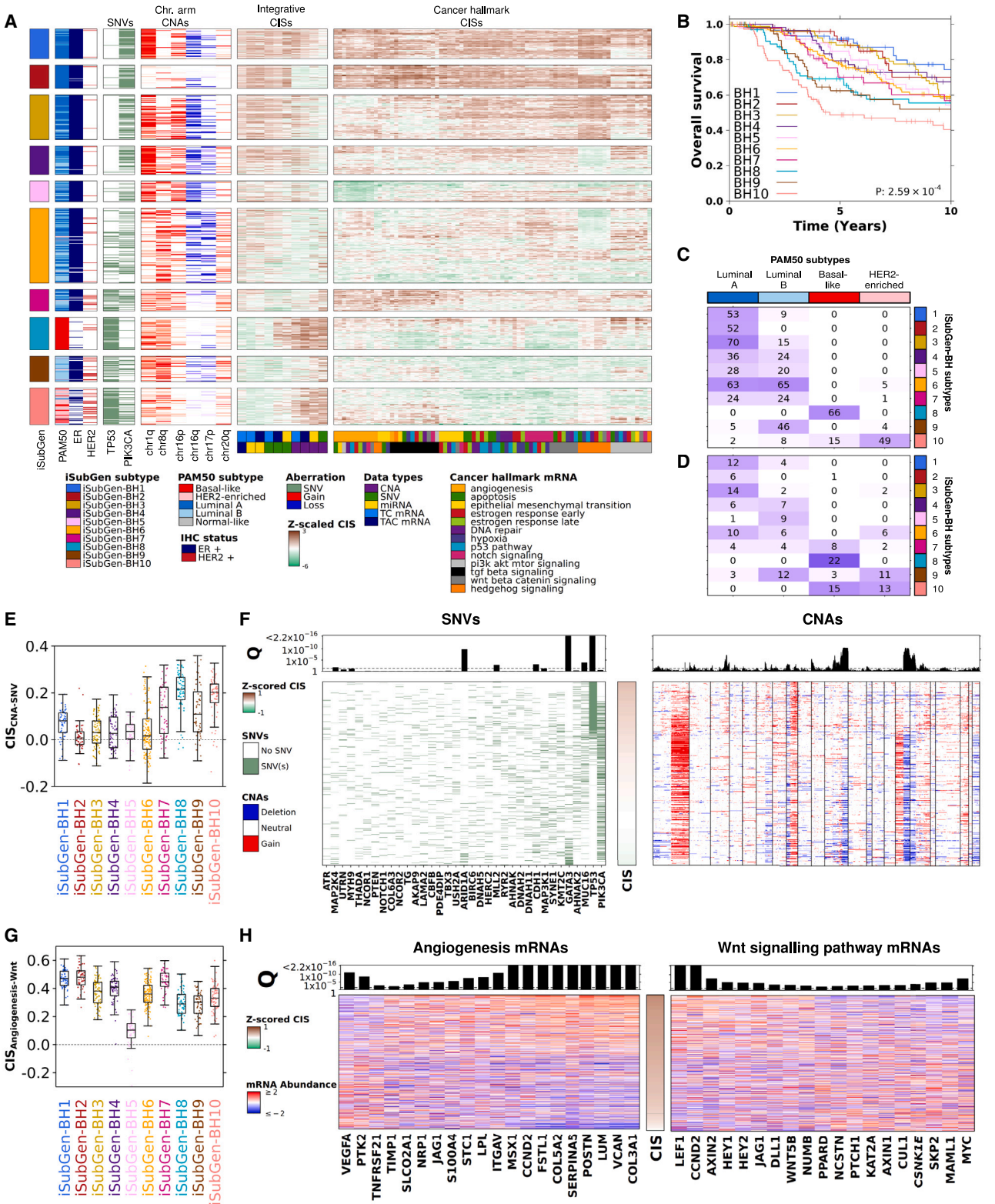


Figure 2. Breast cancer iSubGen combining integrative omics features and cancer hallmark mRNA features

(A) Using iSubGen, the breast cancer patients in the training cohort were classified into ten subtypes using integrative omics features and mRNA cancer hallmark features.

(legend continued on next page)

in both subsets: iSubGen-P1 through iSubGen-P14 and iSubGen-Q1 through iSubGen-Q14 in the discovery and validation cohorts, respectively.

iSubGen-P1, which is composed almost entirely of skin cutaneous melanoma (SKCM), had the highest $CIS_{SNV\text{-}signature}$ (Figures S2A and S2B). Lung adenocarcinomas (LUAD), stomach and esophageal carcinoma (STES), breast cancers (BRCA), bladder cancers (BLCA), and head and neck squamous cell cancers (HNSC) were classified together in multiple groups. THCA were separated from the other cancers into two thyroid cancer groups: iSubGen-P10/iSubGen-Q10 and iSubGen-P11/iSubGenQ11 (Figures S2B and S2D). iSubGen-P10/iSubGen-P11 contained 90% (135/150) and iSubGen-Q10/iSubGen-Q11 97% (146/150) of THCA patients in their respective cohorts. The obvious differences between these two thyroid cancer groups were elevations of $CIS_{SNV\text{-}methylation}$ ($p < 2.2 \times 10^{-16}$), $CIS_{SNV\text{-}mRNA}$ ($p < 2.2 \times 10^{-16}$), and $CIS_{SNV\text{-}miRNA}$ ($p < 2.2 \times 10^{-16}$) in iSubGen-P10/iSubGen-Q10 relative to iSubGen-P11/iSubGen-Q11, representing subtypes of thyroid cancer (Figures S2C and S2E). The CIS values for iSubGen-P and iSubGen-Q groupings had high concordance (Figure S2F). iSubGen generates CIS and IRF values that are both useful for supervised learning and that allow unsupervised learning to independently create concordant classifications in two pan-cancer datasets.

Integrative molecular-based and pathway-based breast cancer subtyping

To demonstrate the utility of iSubGen for integrative multi-modal subtype discovery, we next applied it to the METABRIC breast cancer cohort, integrating 19,877 mRNA features for both TC and TAC mRNA, 18,852 CNA features, 823 miRNA features, and SNV mutation status of 173 driver genes. When applied to these five data types, iSubGen identified five subtypes (Figure S2G and Table S2), which differed in patient survival (Figure S2H). We named subtypes such that patients in iSubGen-B1 had the best outcome and iSubGen-B5 the worst. iSubGen-B1 and iSubGen-B2 had lower tumor grade (Figure S2I) and size (Figure S2J) than iSubGen-B4 and iSubGen-B5. The five iSubGen-B subtypes were tightly associated with the PAM50 subtypes.[7,31] Notably, iSubGen-B5 contained most HER2-enriched and basal-like breast cancers in both training and testing cohorts (Figure S2K), linked to its lower $CIS_{TC\ mRNA\text{-}TAC\ mRNA}$ ($p < 2.2 \times 10^{-16}$), $CIS_{TC\ mRNA\text{-}miRNA}$ ($p < 2.2 \times 10^{-16}$), and $CIS_{TAC\ mRNA\text{-}miRNA}$ ($p < 2.2 \times 10^{-16}$) (Figure S2L). This reflects a higher transcriptome similarity among luminal breast cancers than among HER2-enriched or basal-like cancers. Even among the luminal cancers, the good-outcome iSubGen-B1 and iSubGen-B2 subtypes had higher $CIS_{SNV\text{-}mRNA}$ and $CIS_{SNV\text{-}miRNA}$ ($p < 2.2 \times 10^{-16}$; Figure S2L). We similarly identified strong associations between iSubGen-B and METABRIC IntClust subtypes[14] (Figure S2M). We calculated the Akaike information criterion (AIC) scores of accelerated failure time models using iSubGen, PAM50, and IntClust subtypes (Figures S2H and S2N). The PAM50 (AIC: 2,165) and iSubGen (AIC: 2,168) subtypes had similar scores, while IntClust performed slightly worse (AIC: 2,181). We conducted a comprehensive comparison between iSubGen ($K = 10$) and IntClust subtypes ($K = 10$). The contingency table showed a high overlap for certain subtypes, such as IntClust-7 and -9 within iSubGen-2, while other subtypes did not correspond as neatly but were spread over several subtypes (Figure S2O). To understand the distinction between iSubGen and IntClust, we performed survival and clinical association analyses for each iSubGen subtype within IntClust categories and vice versa (Table S4). Despite all being categorized in IntClust-4, iSubGen-4 and iSubGen-9 exhibited significant differences in patient outcome and in molecular features such as ER status and PAM50 subtype (Figures S2P–S2T). Since the normal-like PAM50 subtype in breast cancer might result in part from contamination of tumor-adjacent normal tissue,[32–34] we excluded these samples from our initial analyses. When normal-like samples were included, iSubGen again produced subtypes (Figures S3A–S3D and S3E; Table S2) strongly associated with PAM50 subtypes (Figures S3B and S3C) and associated with patient survival (Figure S3F).

To demonstrate that iSubGen can be useful with only a single molecular data type, we next focused on the mRNA abundance data of METABRIC, evaluated as a set of 13 cancer hallmark pathways[35,36] (Figure S3G and Table S2). In the training cohort, iSubGen identified seven hallmark subtypes associated with overall survival (Figure S3I). The CISs between cancer hallmarks were generally higher (median CIS 0.5) than CISs between different data types (median CIS 0.06, $p < 2.2 \times 10^{-16}$). iSubGen hallmark-based breast cancer subtypes (iSubGen-H1 through iSubGen-H7) were associated with PAM50 (Figure S3H), iSubGen-B subtypes (Figure S3J), and IntClust (Figure S3K). In general, higher CISs between hallmarks in the iSubGen-H subtypes were associated with better overall patient survival (Figures S3G and S3I). iSubGen subtypes are concordant with the idea that tumors with more dysregulation across data types and signaling pathways have poorer outcomes.

The iSubGen-B and iSubGen-H subtypes assess breast cancer by two different paradigms: iSubGen-B is a genome-wide approach, and iSubGen-H is a pathway approach. We combined the engineered features from these two approaches within a single model to demonstrate the flexibility of iSubGen. Together,

(B) Overall survival for the iSubGen-BH breast cancer subtypes. p value is from a log-rank test.
(C) Comparison of the iSubGen breast cancer subtypes and PAM50 subtypes in the training cohort. Heatmap coloring represents the number of the patients in each overlap.
(D) Comparison of the iSubGen breast cancer subtypes and PAM50 subtypes in the testing cohort.
(E) Comparison of $CIS_{CNA\text{-}SNV}$ distributions between iSubGen-BH subtypes in the training cohort.
(F) Association of SNV and CNA features with the $CIS_{CNA\text{-}SNV}$. The top barplot shows significance between each feature and $CIS_{CNA\text{-}SNV}$ using Wilcoxon rank-sum tests and false discovery rate (FDR) adjustment.
(G) Comparison of CIS between angiogenesis mRNA set and Wnt/β-catenin signaling mRNA set for iSubGen-BH subtypes in the training cohort.
(H) Association of $CIS_{angiogenesis\text{-}Wnt/\beta\text{-}catenin\ signaling}$ with Z-scaled mRNA abundance from each gene set ($q < 0.01$). The top barplot shows significance between each mRNA and the CIS using Spearman's correlation and FDR adjustment. Genes are ordered by Spearman's correlation with correlations decreasing out from middle and the CIS panel.

these six molecular data types comprise 39,725 molecular features and 13 pathway activities. We identified ten subtypes in our training cohort (Figure 2A and Table S2) and again named them by their association with overall survival: iSubGen-BH1 through iSubGen-BH10 (Figure 2B). The iSubGen-BH subtypes associated with PAM50 subtypes and improved the separation of basal-like breast cancers and HER2-enriched cancers relative to iSubGen-B and iSubGen-H (Figure 2C). These associations were validated in the testing cohort via centroid classification (Figure 2D), highlighting the reproducibility of iSubGen subtypes. iSubGen has the potential to capture features that are not present in traditional histological or PAM50 subtypes. For example, analysis of the basal-like cancer subtypes in iSubGen-BH8 and iSubGen-BH10 revealed distinct patterns of SNV-paired CISs, suggesting two subtypes of basal-like cancer differentiated by their somatic mutation profiles.

To characterize the iSubGen-BH CISs, we examined their associations with the individual input features. Higher $CIS_{CNA\text{-}SNV}$ was associated with iSubGen-BH7 ($p = 1.4 \times 10^{-7}$), iSubGen-BH8 ($p < 2.2 \times 10^{-16}$), iSubGen-BH9 ($p = 1.7 \times 10^{-7}$), and iSubGen-BH10 ($p < 2.2 \times 10^{-16}$) compared to iSubGen-BH1 through iSubGen-BH6 (Figure 2E). We identified six SNV associations (out of 156) and 1,283 CNA associations (out of 10,662; $q < 0.01$) where mutation of a specific gene was associated with higher or lower $CIS_{CNA\text{-}SNV}$ (Figure 2F). Patients with *TP53* SNVs had high $CIS_{CNA\text{-}SNV}$, while *GATA3* SNVs and *ARID1A* SNVs were associated with low $CIS_{CNA\text{-}SNV}$. Among the associated CNAs, deletion of the q arms of chromosome 11 and chromosome 16 were associated with lower $CIS_{CNA\text{-}SNV}$. We also examined individual input feature association with the hallmark CISs. Lower $CIS_{angiogenesis\text{-}Wnt/\beta\text{-}catenin}$ differentiated iSubGen-BH5 from the other iSubGen-BH subtypes ($p < 2.2 \times 10^{-16}$; Figure 2G). There were 36 angiogenesis mRNAs and 42 Wnt/β-catenin signaling mRNAs in the individual hallmark gene sets that were used to calculate $CIS_{angiogenesis\text{-}Wnt/\beta\text{-}catenin}$. We found three genes, including *VEGFA*, from the angiogenesis gene set and ten genes, including *MYC*, from the Wnt/β-catenin signaling gene set where higher mRNA abundance was associated with lower $CIS_{angiogenesis\text{-}Wnt/\beta\text{-}catenin\ signaling}$ ($p < 0.01$; Figure 2H). There were 17 genes from the angiogenesis gene set and nine genes from the Wnt/β-catenin signaling gene set where lower mRNA abundance was associated with lower $CIS_{angiogenesis\text{-}Wnt/\beta\text{-}catenin\ signaling}$ ($q < 0.01$). Thus, iSubGen enhances subtype development by integrating individual features (IRFs) with feature-feature interactions.

Subtyping using genic and non-genic molecular data

To evaluate whether iSubGen could be used to generate subtypes from more diverse molecular data, we sought to subtype using a combination of gene-based and mutational-process information. We used trinucleotide signatures[20] for 557 patients from three kidney cancer types from the pan-cancer TCGA datasets,[37–39] which to our knowledge have not been previously integrated into multi-modal subtyping strategies. Each cancer type had six available data types: CNA, SNV, trinucleotide signature exposures, methylation, mRNA abundance, and miRNA abundance. We randomly divided patients with all six data types into equal-sized training and testing cohorts. In the training cohort, we used iSubGen to identify eight subtypes: iSubGen-K1 through iSubGen-K8 (Figure 3A and Table S3). iSubGen-K1 contained almost all kidney chromophobe (KICH) cancers, while iSubGen-K2 through iSubGen-K5 comprised predominantly kidney papillary (KIRP) cancers and iSubGen-K6 through iSubGen-K8 predominantly clear cell (KIRC) cancers. Centroid classification in the testing cohort validated the presence and relative frequencies of these subtypes (Figure 3B). Interestingly, KIRC patients in iSubGen-K5 had poorer overall survival than other KIRC patients in both training and testing cohorts (Figure 3C), while KIRP patient survival was not associated with iSubGen-K subtypes (Figure 3D). Trinucleotide signatures were associated with both histological classifications[20] and iSubGen-K subtypes (Figure 3E), which provided a possible etiology for each subtype. KIRC patients classified in iSubGen-K6 and iSubGen-K8 had fewer point mutations attributed to SBS5 than iSubGen-K7. KIRP patients had many mutations attributed to SBS2 and SBS13 relative to KICH and KIRC, suggesting stronger overall AID/APOBEC activity. We also compared iSubGen subtypes to histological classifications (Figure S4A), whereby three subtypes were strongly associated with chromophobe tumors (78.9%–97.6%), three with clear cell tumors (84.4%–100%), and one with papillary tumors (75%). Both chromophobe and clear cell tumors can be subdivided into novel subtypes. The iSubGen kidney subtypes revealed significant disparities in cancer stages, though not in age or sex, underscoring their utility as distinct entities with unique tumor biology (Figures S4C–S4E). iSubGen provides a framework for integrating both mutational and mutational-process information into subtype discovery.

iSubGen subtyping is robust to missing data

Because human cancers vary in size and are often profiled from biopsy rather than surgical specimens, it is common for only a subset of possible molecular assays to be performed. For this and many other reasons, missing data are common in genomic studies. To evaluate iSubGen's performance in the face of missing data, we randomly separated TCGA lung cancer data into a 512-patient training cohort and a 509-patient testing cohort.[40,41] Overall, 446 of 1,021 patients (47%) lacked one or more of the six data types used in classification, split evenly between training and testing cohorts (Figure 4A and Table S3). To select the number of iSubGen subtypes, we assessed the association between histological subtypes and different numbers of iSubGen clusters using the adjusted Rand index in the training cohort. Lung cancers formed two subtypes, iSubGen-L1 and iSubGen-L2. Subtype structure was robust to missing data (Figures 4B and Table S3). iSubGen-L1 largely comprised lung adenocarcinomas, and iSubGen-L2 largely comprised lung squamous cell carcinomas in both training (Figure 4E) and testing (Figure 4F) cohorts. Overall, 89% (230/257) of training and 87% (227/260) of testing cohort lung adenocarcinomas were in iSubGen-L1. Similarly 87% (221/255) of training and 80% (198/249) of testing cohort lung squamous cell carcinomas were in iSubGen-L2 (Figure S4B). iSubGen-L2 had higher median $CIS_{mRNA\text{-}SNV}$ than iSubGen-L1 ($median_{training\ L1} = -0.04$, $median_{training\ L2} = 0.18$, $p_{training} < 2.2 \times 10^{-16}$; $median_{testing\ L1} = -0.04$, $median_{testing\ L2} = 0.16$, $p_{testing} < 2.2 \times 10^{-16}$; Figures 4C and 4D).

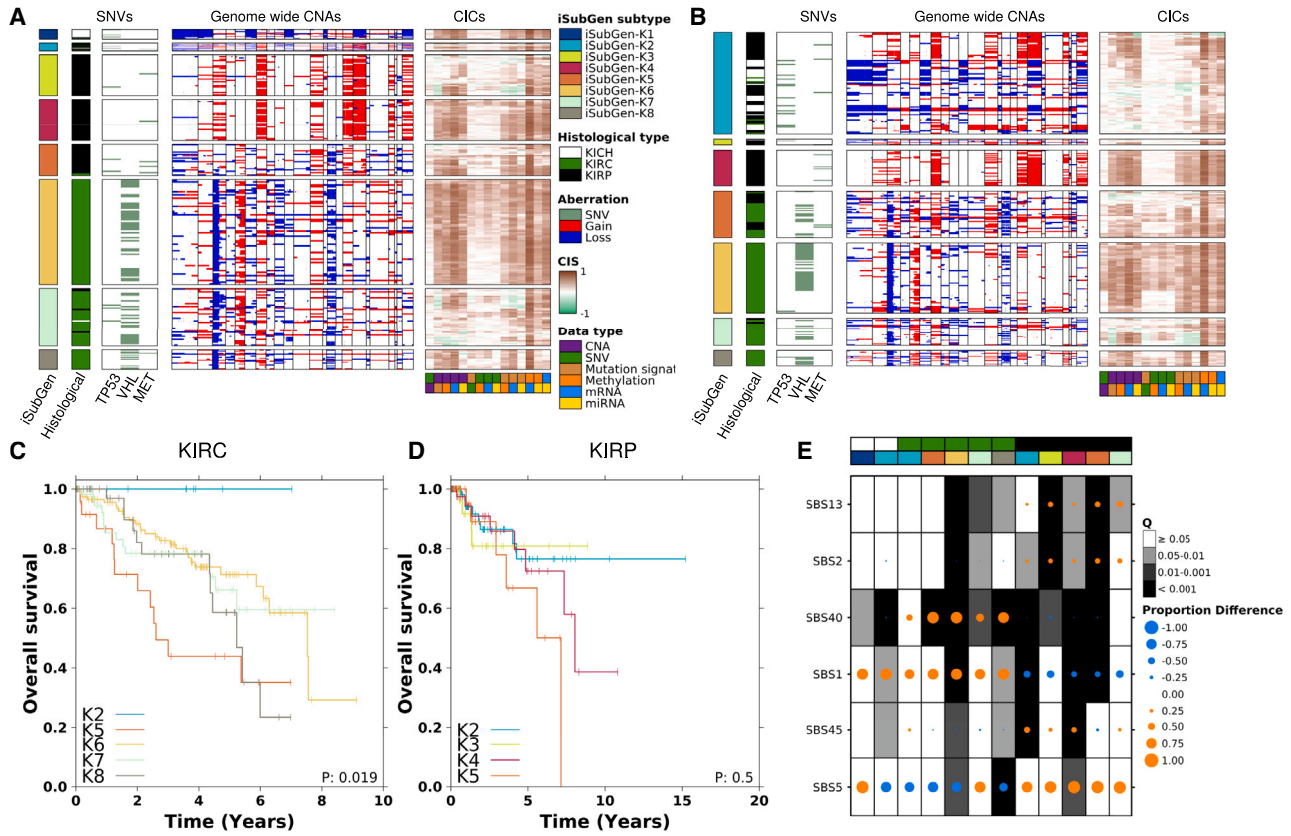


Figure 3. Kidney cancer iSubGen using non-gene-based features

(A) Using iSubGen, patients in the training cohort were classified into six subtypes.

(B) Centroid classification of the iSubGen-K subtypes in the testing cohort.

(C) Overall survival between iSubGen classifications for KIRC patients. Groups with fewer than ten patients are not included. *p* values are from log-rank tests.

(D) Overall survival between iSubGen classifications for KIRP patients. Groups with fewer than ten patients are not included. *p* values are from log-rank tests.

(E) Association of patients in the training and testing cohorts with trinucleotide mutation signatures. Each column is a group of patients. The top covariate shows the TCGA histological cancer type, and the second covariate is the iSubGen classification of the patients. Patient groups with fewer than ten patients are not shown. Each dot is sized to the proportion of patients with mutations from the trinucleotide signature. If the dot is orange, the proportion for the group is greater than the proportion of patients not in the group. Similarly, if the dot is blue, the proportion is less than in the other patients. The background shading is the *q* value from the proportion test comparing the proportion for patients in the group to the proportion for those not in the group.

To visualize these underlying associations between data types, we focused on CISs from exemplar patients and on mRNA and SNV features prior to feature engineering. TCGA.44.5645 had lower overall $CIS_{mRNA\text{-}SNV}$ and near-median values for iSubGen-L1 patients (Figure 4G). TCGA.78.7155 had the highest $CIS_{mRNA\text{-}SNV}$ of all patients in the training cohort and had high $CIS_{mRNA\text{-}SNV}$ for iSubGen-L2 (Figure 4H). TCGA.44.5645 and TCGA.78.7155 both clustered with their respective histological subtype using mRNA abundance (Figure 4I). By contrast, SNVs did not separate patients by histological subtype (Figure 4J): TCGA.44.5645 had more total somatic SNVs than TCGA.78.7155 and clustered with other highly mutated tumors. Other randomly selected patients, including lung squamous cell carcinoma (LUSC) samples TCGA.05.4422 (iSubGen-L2) and TCGA.66.2793 (iSubGen-L2), and LUAD samples TCGA.50.5936 (iSubGen-L2) and TCGA.55.6970 (iSubGen-L1), also showed that LUSC samples have higher $CIS_{mRNA\text{-}SNV}$ than LUAD samples (Figures S4F–S4I). The low $CIS_{mRNA\text{-}SNV}$ values show that SNVs and mRNA provide orthogonal information, leading iSubGen to create composite subtypes that merge them.

To assess the association of CIS values with epidemiological features, we considered sex differences, which have been widely reported in lung cancer.[42–44] We tested whether CISs differed between tumors arising in patients with XX and XY germline chromosome conformations, whereby 7 of 15 CISs were associated with sex (Table S4). For example, XY patients had higher $CIS_{mRNA\text{-}SNV}$ than XX patients: mRNA and SNV profiles were more concordant in lung tumors arising in men than those arising in women. Thus, CISs reflect underlying epidemiological features, independent of missing data.

Benchmarking relative to other subtyping strategies

We compared iSubGen to other integrative subtyping algorithms: concatenation of the data types, clusters of clusters (COCA), similarity network fusion (SNF), and iClusterBayes.[22,45,46] We

CellPress
OPEN ACCESS

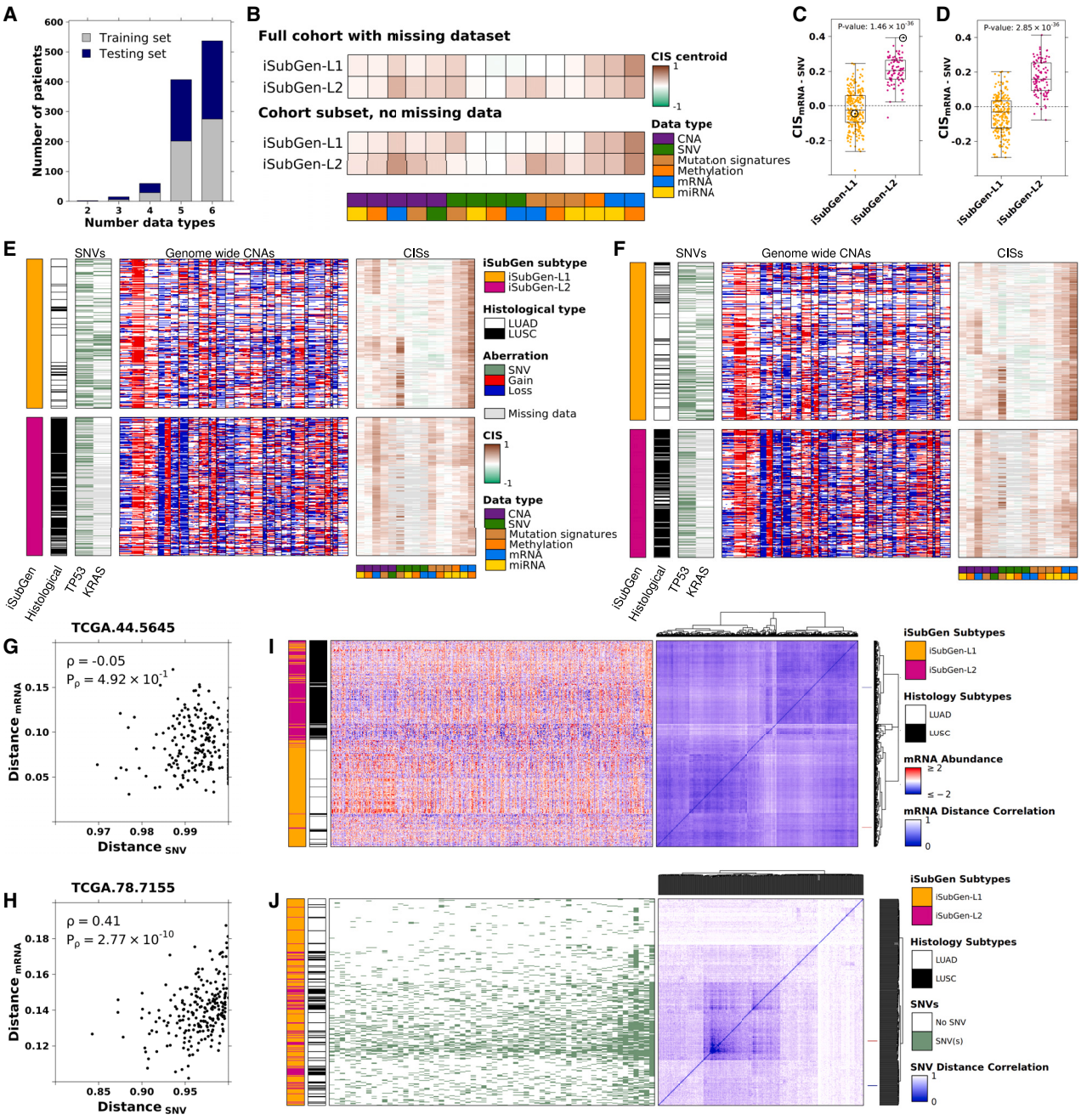


Figure 4. iSubGen is robust to missing data in lung cancer

(A) The number of data types held by each patient.

(B) Using iSubGen, patients from the training cohort were classified into two subtypes including patients with missing data. The top panel shows the centroids from subtyping with missing data. To assess the effect of missing data, a subset of patients that had all six data types were also independently clustered, and these centroids are shown in the bottom panel.

(C) $CIS_{mRNA\text{-}SNV}$ for the training cohort, with the CISs of TCGA.78.7155 and TCGA.44.5645 circled. *p* values are from Wilcoxon rank-sum tests.

(D) $CIS_{mRNA\text{-}SNV}$ for the testing cohort. *p* values are from Wilcoxon rank-sum tests.

(E) iSubGen classification of the training cohort.

(F) Centroid classification of the iSubGen-L subtypes in the testing cohort.

(G) The CIS for patient TCGA.44.5645, who is the patient with the median $CIS_{mRNA\text{-}SNV}$ in iSubGen-L1. Spearman rank correlation coefficient and associated *p* values are denoted in the plot.

(legend continued on next page)

considered the other iCluster algorithms, but iCluster requires that all data types have the same features, and iClusterPlus is restricted to four or fewer data types.[21,47] Since we have more than four data types, iClusterBayes was used to represent the iCluster algorithms. iClusterBayes is limited to a maximum of six data types, which is the maximum here but could be a limitation for other studies. SNF and iClusterBayes also do not accept patients with missing data (particularly missing all data for a single data type), so lung cancer results were compared to results from the patient subset without missing data. We conducted robustness testing by selecting subsets of varying sample sizes (n = 100, 200, 400, 800, and 1,000) from the METABRIC dataset. We assessed subtypes by analyzing survival outcomes with log-rank tests and examining their associations with established classifications (PAM50, IntClust) with the adjusted Rand index (Figures S4J–S4L). Overall, iSubGen performed better in 159 of the 324 pairwise comparisons, as determined by the $-\log_{10}$ (log-rank p value) between iSubGen and the other algorithms. iSubGen outperformed COCA and SNF but performed similarly to iClusterBayes. The adjusted Rand index, which measures the concordance of subtypes with clinical ground truth, indicated that iSubGen was more effective than iClusterBayes but not COCA and SNF. In the pairwise comparison of the adjusted Rand index, iSubGen performed better than or equal to the others in 109 of 324 tests for PAM50 and in 130 of 324 tests for IntClust. Moreover, iSubGen achieved robustness when sample sizes were 400 or more and returned substantially more stable and coherent clusters with larger sample sizes. Overall, these findings confirm that iSubGen performs similarly to or better than the best methods across a range of metrics.

We have also conducted survival and association analyses within the TCGA kidney and lung datasets. Across 32 comparisons of datasets and k values, iSubGen ranked first 14 times and second 7 times (Figures S4M and S4N), thereby outperforming all other algorithms. However, silhouette analysis demonstrates a preference for specific data types across various cancer types (Figures S4O–S4Q). Overall, iSubGen had performance equivalent to the best performance of these alternative methods but was more stable across datasets, being able to handle arbitrary numbers of data types and tolerant to significant missing data.

DISCUSSION

Many factors influence the status and progression of cancer: somatic mutations (e.g., CNAs, SNVs), epigenomic alterations, chromatin reorganization, and external cellular signals all occur on a background of the individuals' health, stress, and exposures over a lifetime. Capturing how all these data types interrelate is an open problem subject to intensive investigation. We introduce iSubGen to capture the associations between different types of data. For iSubGen, we developed a metric called CIS to capture how different data types inter-relate. If feature patterns from two data types define the same patient associations (high CIS), then the two data types may reflect a regulatory relationship of some type. For example, a higher $CIS_{SNV\text{-}mRNA}$ could be explained by a set of mutated genes, such as transcription factors, that drive broad mRNA abundance patterns. CISs, along with a reduced form of the single-data-type information termed IRFs, both serve as intermediaries for integrative subtype discovery and can be used for direct supervised biomarker development.

iSubGen facilitates maximum data-type inclusion and modular replacement of the framework steps to personalize for different situations. However, with increased options comes increased parameterization and the need to check that the underlying engineered features are reproducible. Indeed, iSubGen does not directly incorporate prior information (although its robustness to missing data provides a natural pathway for doing so). Almost all subtype-development approaches face this challenge: there is no one metric to quantitatively optimize clustering results when selecting weightings, the number of clusters, and the ultimate subtypes. We considered inter-subtype differences, association with CISs, prognostic associations, and, since we were assessing well-characterized cancer types, association with known histopathological subtypes. In any subtyping, it is up to the user to decide what is most important in choosing a subtype number when multiple different values can bring statistically reasonable results using domain knowledge.

Subtypes provide fundamental understanding about polygenic disease—they identify groups of patients whose current diseases share similar appearance and thus might share both similar histories and future responses to treatment. Potential applications of iSubGen extend to almost any complex biological system. For example, to understand the effect of human microbiomes on health, we will need to recognize patterns across underlying human genetics, epigenetics, metabolomics, and the microorganisms present in the gut. Data arising from mobile devices provide a completely different setting with a plethora of data types to combine and interpret. iSubGen provides a flexible framework within which to capture multi-modal interactions in diverse science-data applications.

Limitations of the study

A caveat of iSubGen is its long execution time relative to other subtype discovery methods. The calculation of CIS values is computationally demanding because of the number of similarity calculations for, first, similarity matrices for each data type and second, CIS calculations between data types. CIS is a consensus metric and so requires iterations of all the similarity calculations. For large numbers of patients, as in our pan-cancer

(H) The CIS for patient TCGA.78.7155, who is the patient with the highest $CIS_{mRNA\text{-}SNV}$ in iSubGen-L2. Spearman rank correlation coefficient and associated p values are denoted in the plot.
(I) Z-scaled mRNA abundance for mRNAs with standard deviation greater than 2. mRNA is on the x axis and patients on the y axis. Far-right plot shows mRNA abundance patient-by-patient similarity matrix using 1 minus Pearson correlation as the similarity metric.
(J) SNVs for genes mutated in more than 50 patients. Genes (x axis) are ordered by mutation frequency. Patients are on the y axis. Far-right plot shows SNV patient-by-patient similarity matrix calculated using SNVs without patient recurrence filtering and Jaccard distance as the similarity metric.
In (I) and (J), TCGA.78.7155 is indicated in blue and TCGA.44.5645 in red.

analysis, creation of the CIS matrix can take several hours on a single CPU (one core), although this can be readily parallelized. Another limitation of iSubGen may be the lack of a feature-selection function within the package, and indeed the optimal strategy to feature select in the context of subtype discovery remains unclear.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Paul C. Boutros (pboutros@mednet.ucla.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data used in this research are all publicly available. Accession numbers are listed in the key resources table.
- All original code has been deposited at GitHub and is publicly available as of the date of publication. iSubGen is freely available as an R package from CRAN at https://cran.r-project.org/web/packages/iSubGen/index.html. Its code is available at https://github.com/uclahs-cds/public-R-iSubGen. An archival DOI is 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.

ACKNOWLEDGMENTS

P.C.B. was supported by a Terry Fox Research Institute New Investigator Award and a CIHR New Investigator Award, and by the NIH/NCI under awards P30CA016042, U24CA248265, U01CA214194, and U2CCA271894. N.S.F. was supported by a CIHR Canadian Graduate Scholarship, a CIHR Michael Smith Foreign Study Supplement, a Medical Biophysics Excellence University of Toronto Fund Scholarship, a University of Toronto Geoff Lockwood and Kevin Graham Medical Biophysics Graduate Scholarship, and a Prostate Cancer Canada Philip Feldberg Studentship.

AUTHOR CONTRIBUTIONS

Conceptualization, N.S.F. and P.C.B.; software, N.S.F.; data curation and resources, S.H., C.H.L., M.T., and A.L.M.; formal analysis, investigation, and visualization, N.S.F., M.T., and A.L.M.; data interpretation, N.S.F., M.T., and P.C.B.; writing – original draft, N.S.F.; writing – review & editing, M.T. and P.C.B.; supervision, project administration, and funding acquisition, P.C.B. All authors approved the manuscript.

DECLARATION OF INTERESTS

P.C.B. sits on the scientific advisory boards of Sage Bionetworks Inc., BioSymetrics Inc., and Intersect Diagnostics Inc. At the time of publication, N.S.F. was an employee of Hoffman-La Roche Limited (Roche Canada). All contributions by N.S.F. were completed prior to this employment.

STAR★METHODS

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

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2024.100884.

Received: December 6, 2021
Revised: July 6, 2023
Accepted: October 1, 2024
Published: October 23, 2024

REFERENCES

1. Thenganatt, M.A., and Jankovic, J. (2014). Parkinson Disease Subtypes. JAMA Neurol. *71*, 499–504. https://doi.org/10.1001/jamaneurol.2013.6233.
2. Isaacs, J.D., and Ferraccioli, G. (2011). The need for personalised medicine for rheumatoid arthritis. Ann. Rheum. Dis. *70*, 4–7. https://doi.org/10.1136/ard.2010.135376.
3. Ellis, I.o., Galea, M., Broughton, N., Locker, A., Blamey, R.w., and Elston, C.w. (1992). Pathological prognostic factors in breast cancer. II. Histological type. Relationship with survival in a large study with long-term follow-up. Histopathology *20*, 479–489. https://doi.org/10.1111/j.1365-2559.1992.tb01032.x.
4. Govindan, R., Page, N., Morgensztern, D., Read, W., Tierney, R., Vlahiotis, A., Spitznagel, E.L., and Piccirillo, J. (2006). Changing epidemiology of small-cell lung cancer in the United States over the last 30 years: analysis of the surveillance, epidemiologic, and end results database. J. Clin. Oncol. *24*, 4539–4544. https://doi.org/10.1200/JCO.2005.04.4859.
5. Patard, J.-J., Leray, E., Rioux-Leclercq, N., Cindolo, L., Ficarra, V., Zisman, A., De La Taille, A., Tostain, J., Artibani, W., Abbou, C.C., et al. (2005). Prognostic value of histologic subtypes in renal cell carcinoma: a multicenter experience. J. Clin. Oncol. *23*, 2763–2771. https://doi.org/10.1200/JCO.2005.07.055.
6. (1982). The World Health Organization histological typing of lung tumours. Second edition. Am. J. Clin. Pathol. *77*, 123–136. https://doi.org/10.1093/ajcp/77.2.123.
7. Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752. https://doi.org/10.1038/35021093.
8. Carlson, R.W., Scavone, J.L., Koh, W.-J., McClure, J.S., Greer, B.E., Kumar, R., McMillian, N.R., and Anderson, B.O. (2016). NCCN Framework for Resource Stratification: A Framework for Providing and Improving Global Quality Oncology Care. J. Natl. Compr. Cancer Netw. *14*, 961–969. https://doi.org/10.6004/jnccn.2016.0103.
9. Nowell, P.C. (1976). The Clonal Evolution of Tumor Cell Populations. Science *194*, 23–28. https://doi.org/10.1126/science.959840.
10. 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.

11. 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.

12. Bhandari, V., Hoey, C., Liu, L.Y., Lalonde, E., Ray, J., Livingstone, J., Lesurf, R., Shiah, Y.-J., Vujcic, T., Huang, X., et al. (2019). Molecular landmarks of tumor hypoxia across cancer types. Nat. Genet. *51*, 308–318. https://doi.org/10.1038/s41588-018-0318-2.

13. Bhandari, V., Li, C.H., Bristow, R.G., and Boutros, P.C.; PCAWG Consortium (2020). Divergent mutational processes distinguish hypoxic and normoxic tumours. Nat. Commun. *11*, 737. https://doi.org/10.1038/s41467-019-14052-x.

14. Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352. https://doi.org/10.1038/nature10983.

15. Parsons, D.W., Jones, S., Zhang, X., Lin, J.C.-H., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.-M., Gallia, G.L., et al. (2008). An Integrated Genomic Analysis of Human Glioblastoma Multiforme. Science *321*, 1807–1812. https://doi.org/10.1126/science.1164382.

16. Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. Nat. Med. *21*, 1350–1356. https://doi.org/10.1038/nm.3967.

17. Roepman, P., Schlicker, A., Tabernero, J., Majewski, I., Tian, S., Moreno, V., Snel, M.H., Chresta, C.M., Rosenberg, R., Nitsche, U., et al. (2014). Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. Int. J. Cancer *134*, 552–562. https://doi.org/10.1002/ijc.28387.

18. Osborne, C.K., Yochmowitz, M.G., Knight, W.A., and McGuire, W.L. (1980). The value of estrogen and progesterone receptors in the treatment of breast cancer. Cancer *46*, 2884–2888. https://doi.org/10.1002/1097-0142(19801215)46:12+<2884::aid-cncr2820461429>3.0.co;2-u.

19. Siev, M., Renson, A., Tan, H.-J., Rose, T.L., Kang, S.K., Huang, W.C., and Bjurlin, M.A. (2020). Prognostic Value of Histologic Subtype and Treatment Modality for T1a Kidney Cancers. Kidney Cancer *4*, 49–58. https://doi.org/10.3233/kca-190072.

20. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421. https://doi.org/10.1038/nature12477.

21. Shen, R., Olshen, A.B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics *25*, 2906–2912. https://doi.org/10.1093/bioinformatics/btp543.

22. 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., 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.

23. Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics *26*, i237–i245. https://doi.org/10.1093/bioinformatics/btq182.

24. Voillet, V., Besse, P., Liaubet, L., San Cristobal, M., and González, I. (2016). Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. BMC Bioinf. *17*, 402. https://doi.org/10.1186/s12859-016-1273-5.

25. Lin, D.-Y., Zeng, D., and Couper, D. (2020). A general framework for integrative analysis of incomplete multiomics data. Genet. Epidemiol. *44*, 646–664. https://doi.org/10.1002/gepi.22328.

26. Dvinge, H., Git, A., Gräf, S., Salmon-Divon, M., Curtis, C., Sottoriva, A., Zhao, Y., Hirst, M., Armisen, J., Miska, E.A., et al. (2013). The shaping and functional consequences of the microRNA landscape in breast cancer. Nature *497*, 378–382. https://doi.org/10.1038/nature12108.

27. Pereira, B., Chin, S.-F., Rueda, O.M., Vollan, H.-K.M., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat. Commun. *7*, 11479. https://doi.org/10.1038/ncomms11479.

28. Fox, N.S., Haider, S., Harris, A.L., and Boutros, P.C. (2019). Landscape of transcriptomic interactions between breast cancer and its microenvironment. Nat. Commun. *10*, 3116. https://doi.org/10.1038/s41467-019-10929-z.

29. Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573. https://doi.org/10.1093/bioinformatics/btq170.

30. Zhang, W., Liu, Y., Sun, N., Wang, D., Boyd-Kirkup, J., Dou, X., and Han, J.-D.J. (2013). Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer. Cell Rep. *4*, 542–553. https://doi.org/10.1016/j.celrep.2013.07.010.

31. Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J. Clin. Oncol. *27*, 1160–1167. https://doi.org/10.1200/JCO.2008.18.1370.

32. Pusztai, L., Mazouni, C., Anderson, K., Wu, Y., and Symmans, W.F. (2006). Molecular classification of breast cancer: limitations and potential. Oncol. *11*, 868–877. https://doi.org/10.1634/theoncologist.11-8-868.

33. Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., and Pietenpol, J.A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J. Clin. Invest. *121*, 2750–2767. https://doi.org/10.1172/JCI45014.

34. Weigelt, B., Baehner, F.L., and Reis-Filho, J.S. (2010). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. J. Pathol. *220*, 263–280. https://doi.org/10.1002/path.2648.

35. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

36. Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70. https://doi.org/10.1016/s0092-8674(00)81683-9.

37. Cancer Genome Atlas Research Network (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature *499*, 43–49. https://doi.org/10.1038/nature12222.

38. Davis, C.F., Ricketts, C.J., Wang, M., Yang, L., Cherniack, A.D., Shen, H., Buhay, C., Kang, H., Kim, S.C., Fahey, C.C., et al. (2014). The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell *26*, 319–330. https://doi.org/10.1016/j.ccr.2014.07.014.

39. Cancer Genome Atlas Research Network; Linehan, W.M., Spellman, P.T., Ricketts, C.J., Creighton, C.J., Fei, S.S., Davis, C., Wheeler, D.A., Murray, B.A., Schmidt, L., et al. (2016). Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. N. Engl. J. Med. *374*, 135–145. https://doi.org/10.1056/NEJMoa1505917.

40. Cancer Genome Atlas Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519–525. https://doi.org/10.1038/nature11404.

41. Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550. https://doi.org/10.1038/nature13385.

42. Yuan, Y., Liu, L., Chen, H., Wang, Y., Xu, Y., Mao, H., Li, J., Mills, G.B., Shu, Y., Li, L., and Liang, H. (2016). Comprehensive Characterization of Molecular Differences in Cancer between Male and Female Patients. Cancer Cell *29*, 711–722. https://doi.org/10.1016/j.ccell.2016.04.001.

43. Li, C.H., Haider, S., Shiah, Y.-J., Thai, K., and Boutros, P.C. (2018). Sex Differences in Cancer Driver Genes and Biomarkers. Cancer Res. *78*, 5527–5537. https://doi.org/10.1158/0008-5472.CAN-18-0362.

44. Li, C.H., Prokopec, S.D., Sun, R.X., Yousif, F., Schmitz, N., and PCAWG Tumour Subtypes and Clinical Translation; and Boutros, P.C.; PCAWG Consortium (2020). Sex differences in oncogenic mutational processes. Nat. Commun. *11*, 4330. https://doi.org/10.1038/s41467-020-17359-2.

45. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods *11*, 333–337. https://doi.org/10.1038/nmeth.2810.

46. Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K.S., and Hilsenbeck, S.G. (2018). A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics *19*, 71–86. https://doi.org/10.1093/biostatistics/kxx017.

47. Mo, Q., Wang, S., Seshan, V.E., Olshen, A.B., Schultz, N., Sander, C., Powers, R.S., Ladanyi, M., and Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc. Natl. Acad. Sci. USA *110*, 4245–4250. https://doi.org/10.1073/pnas.1208949110.

48. Quon, G., Haider, S., Deshwar, A.G., Cui, A., Boutros, P.C., and Morris, Q. (2013). Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. Genome Med. *5*, 29. https://doi.org/10.1186/gm433.

49. Anghel, C.V., Quon, G., Haider, S., Nguyen, F., Deshwar, A.G., Morris, Q.D., and Boutros, P.C. (2015). ISOpureR: an R implementation of a computational purification algorithm of mixed tumour profiles. BMC Bioinf. *16*, 156. https://doi.org/10.1186/s12859-015-0597-x.

50. Jackson, C.H. (2016). flexsurv: A Platform for Parametric Survival Modeling in R. J. Stat. Software *70*, i08. https://doi.org/10.18637/jss.v070.i08.

51. P'ng, C., Green, J., Chong, L.C., Waggott, D., Prokopec, S.D., Shamsi, M., Nguyen, F., Mak, D.Y.F., Lam, F., Albuquerque, M.A., et al. (2019). BPG: Seamless, automated and interactive visualization of scientific data. BMC Bioinf. *20*, 42. https://doi.org/10.1186/s12859-019-2610-2.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
METABRIC Breast Cancer Dataset	European Genome-Phenome Archive[14]	EGAS00000000083
TCGA Dataset	Broad GDAC Firehose	https://gdac.broadinstitute.org/
Software and algorithms		
iSubGen(v1.0.2)	This paper	https://cran.r-project.org/web/packages/iSubGen and https://github.com/uclahs-cds/package-iSubGen and https://doi.org/10.5281/zenodo.13852306
ConsensusClusterPlus (v1.8.1)	Wilkerson and Hayes, 2010[29]	https://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html
keras: R Interface to 'Keras' (v2.15.0)	https://github.com/rstudio/keras/tree/r2	https://cran.r-project.org/web/packages/keras
tensorflow: R Interface to 'TensorFlow' (v2.16.0)	https://github.com/rstudio/tensorflow	https://cran.r-project.org/web/packages/tensorflow

METHOD DETAILS

METABRIC breast cancer dataset

The METABRIC cohort contains 1,991 patients each with a primary fresh frozen breast cancer specimen.[14,26,27] METABRIC annotation includes overall survival and PAM50 subtype classifications. Six patients had subtype classification NC (not classified) and 211 patients had subtype classification of normal-like breast cancer were excluded in the main analysis. For one supplementary analysis, we included the normal-like breast cancer samples for integrative subtype generation (Figures S3A–S3F). The cohort also has mRNA profiles from Illumina HT-12 v3 microarrays for 144 adjacent normal breast tissue samples from a subset of the patients with breast cancer samples. The METABRIC dataset includes mRNA abundance, CNAs, miRNA and SNVs. The relative mRNA abundances of 19,877 genes were profiled using Illumina HT-12 v3 microarrays for 1,988 patients. CNA data covers 18,852 genes profiled using Affymetrix SNP 6.0 microarrays for 1,989 patients. There are 823 relative miRNA abundances profiles using Agilent Human miRNA Microarray 2.0 for 1,285 patients.[26] The METABRIC cohort also included targeted sequencing data covering 173 genes frequently mutated in breast cancer (i.e., candidate driver genes) with somatic SNV calls.[27] The mRNA abundance was deconvolved into tumor cell and tumor adjacent cell mRNA abundance using ISOpure.[28,48,49] TC/TAC deconvolution was performed for all patients in the training cohort and all patients in the testing cohort. We used the training and testing cohort divisions from the METABRIC paper. Subtypes were discovered using only the training cohort.

TCGA data

TCGA datasets were downloaded from Broad GDAC Firehose (https://gdac.broadinstitute.org/), release 2016-01-28. We used the mRNA abundance, CNAs, SNVs for the TCGA samples. The mRNA abundances of 20,531 genes were profiled using exome sequencing. CNA data covers 24,776 genes profiled using Affymetrix SNP 6.0 microarrays. There are 18,152 genes with a mutation for the SNV data. Per patient trinucleotide mutational signatures calls were also downloaded. miRNA was downloaded from the GDC Data Portal (https://portal.gdc.cancer.gov/), data Release 25.0 – July 22, 2020. There were 1,881 miRNA abundance profiled through sequencing.

For breast cancer, we excluded samples classified as normal-like PAM50 subtype and used 777 patients with four available data types: mRNA, CNAs, SNVs and trinucleotide mutational signature. Patients were randomly equally split into training and testing cohorts, stratifying by subtype.

For kidney cancer, we used 241 *kidney renal* papillary cell *carcinomas* (KIRP), 267 kidney renal clear cell carcinomas (KIRC) and 49 chromophobe renal cell carcinomas (KICH).[37–39] We only used patients with all six data types. Patients were randomly divided per subtype to create equally sized training and testing cohorts.

For lung cancer, we used 1,021 patients from the TCGA lung cancer cohorts,[40,41] which is a combination of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). Using random sampling per subtype, we divided the patients into a training cohort of 512 and a testing cohort of 509 patients.

For the pan-cancer cohort, twelve TCGA datasets with more than 200 patients with the data types were chosen: BLCA, BRCA, HNSC, KIRC, KIRP, LGG, LIHC, LUAD, PRAD, SKCM, STES, THCA. From each cancer type, we equally divided up to 300 randomly selected patients in two non-overlapping subsets. In total each subset had 1,709 patients.

Independent reduced features

The reduced feature matrix is a matrix where each row is a patient and each column is an IRF. For each data type, an autoencoder was created using the keras (v2.1.5) and tensorflow (v1.10) packages in R. RNA profiles were scaled before inputting to the autoencoder. The autoencoders were trained with mean squared error loss function, Adam optimization and tanh as the activation function. Each autoencoder had three hidden layers with fifteen nodes, two nodes, fifteen nodes respectively. We tested one, three and five hidden layers with various node sizes (1, 2, 5, 15, 30, 25, 50, 100, 200). The IRFs were then extracted from the layer with the bottleneck layer (here the layer with two nodes). These IRFs for each data type were combined into a matrix where each column corresponded to a node in the bottleneck layer from the autoencoders. There were two columns from each data type.

Consensus integrative similarities

The pairwise comparison matrix is a matrix where each row is a patient, and each column is a pair of data types. The entries in the matrix are correlations or consensus correlations. To calculate the correlation for a patient and a pair of data types, the similarities between that patient and each of the other patients in the cohort were calculated for each of the data types. These similarities were then correlated between two data types. This was repeated for each patient and each pair of data types using Spearman's correlations. The similarity metric varied depending on the molecular data type. For CNA, SNV, trinucleotide mutational signatures data types, we used Jaccard distance as the similarity metric. For mRNA, miRNA and methylation data types, we used 1 - Pearson's correlation as the similarity metric. To create CISs, patients were correlated with bootstrapping and each CIS was the median correlation from the sub-sampled repetitions. For each bootstrap, 80% of the patients were sampled without replacement and all the patients were individually correlated to that 0.8 subset of patients. This was repeated 10 times and the median of the correlations for each patient and data type pair.

Integrative subtype generation (iSubGen)

There are four steps to creating subtypes with iSubGen. (1) Create a pairwise comparison matrix which assesses the relationships between patient similarities in a pairwise approach such as CISs. (2) Create a reduced feature matrix to assess the main pattern of each independent matrix. (3) Combine the pairwise similarity matrix and the reduced feature matrix with appropriate re-weighting. (4) Perform pattern discovery on the combined matrix. We used consensus clustering (v1.8.1)[29] with a seed of 17, with 1000 clustering repetitions and Euclidean distance metric and hierarchical clustering with Ward linkage. The number of clusters was determined using the consensus cluster results, including the consensus matrix and cumulative distribution functions and association with CISs and clinical features.

QUANTIFICATION AND STATISTICAL ANALYSIS

Survival associations of CISs

We created a receiver operating characteristic curve using the CISs for each pair of data types. For each pair of data types, the CISs were dichotomized at every possible threshold and agreement with overall survival was assessed. For further examples of the survival associations, we created Kaplan-Meier curves and assessed the survival association using log rank tests for $CIS_{TAC\ mRNA\ -\ miRNA}$. CIS dichotomization threshold was chosen to maximize the harmonic mean of true positive and false positive rates for predicting five-year overall survival using all the patients in the training cohort.

Random forest classifiers

We created random forest classifiers predicting breast cancer subtype or pan-cancer cancer type from CISs using the randomForest (v4.6-14) R package. Receiver Operating Characteristic (ROC) curves and the area under ROC curve (AUROC or AUC) were calculated using the pROC (v 1.18.2) R package.

Breast cancer integrative multi-omics subtypes

iSubGen was run on the 684 patients in the training cohort with all five data types: CNA, SNV, miRNA abundance, TC mRNA abundance and TAC mRNA abundance. TC mRNA, TAC mRNA and miRNA features were *Z*-scaled per feature before autoencoder training. For each data type, the autoencoder had three hidden layers with fifteen nodes, two nodes, fifteen nodes respectively. Therefore there were two independent reduced features (IRFs) per data type. Consensus clustering was performed for 2 to 10 subtypes with 0.8 sub-sampling of features and patients. A weighting of 1:8 for CISs to independent reduced features was selected. The hyperparameter of five for subtype number was selected by visual assessment of iSubGen subtypes with CIS, PAM50 subtypes and prognosis in the training cohort. There were 367 patients in the testing cohort with all five data types. Testing cohort independent reduced features were created using the trained autoencoders with TC mRNA, TAC mRNA and miRNA features scaled using the mean and standard deviations from the training cohort. Testing cohort CISs were calculated for each patient relative to the patients in the training cohort, not relative to the patients in the other patients in the testing cohort.

Breast cancer subtypes from mRNA of cancer hallmarks and pathways

iSubGen was run on the 996 patients in the training cohort with mRNA abundance. Thirteen gene sets were selected from the MSigDB hallmark gene sets collection and mRNA abundance for each gene set was used as a separate data type. mRNA features

were *Z*-scaled per feature per mRNA set before autoencoder training. For each data type, the autoencoder had three hidden layers with fifteen nodes, two nodes, fifteen nodes respectively. Therefore there were two IRFs per data type. Consensus clustering was performed for 2 to 10 subtypes with 0.8 sub-sampling of features and patients. A weighting of 1:4 for CISs to independent reduced features was selected. The hyperparameter of nine for subtype number was selected by visual assessment of iSubGen subtypes with CIS, PAM50 subtypes and prognosis in the training cohort.

Breast cancer subtypes combining integrative-omics features and mRNA sets

iSubGen was run on the 684 patients in the training cohort with all five data types: CNA, SNV, miRNA abundance, TC mRNA abundance and TAC mRNA abundance. Features were used as described from breast cancer integrative multi-omics subtypes and breast cancer subtypes from mRNA of cancer hallmarks and pathways. A weighting of 1:2 for CISs to IRFs was selected. Consensus clustering was performed for 2 to 18 subtypes with 0.2 sub-sampling of features and 0.8 sub-sampling of patients. The hyperparameter of ten for subtype number was selected by visual assessment of iSubGen subtypes with CIS, PAM50 subtypes and prognosis in the training cohort.

Kidney cancer subtypes

iSubGen was run on the 283 patients in the training cohort with all six data types: CNA, SNV, trinucleotide mutational signatures, methylation, mRNA abundance and miRNA abundance. mRNA and miRNA features were *Z*-scaled per feature before autoencoder training. For each data type, the autoencoder had three hidden layers with fifteen nodes, two nodes, fifteen nodes respectively. Therefore, there were two IRFs per data type. Consensus clustering was performed for 2 to 10 subtypes with 0.8 sub-sampling of features and patients. A weighting of 1:2 for CISs to independent reduced features was selected. The hyperparameter of five for subtype number was selected by visual assessment of iSubGen subtypes with CIS, TCGA kidney cancer types and prognosis in the training cohort. There were 274 patients in the testing cohort with all six data types that were classified using centroid classification. Testing cohort independent reduced features (IRFs) were created using the trained autoencoders with mRNA and miRNA features scaled using the mean and standard deviations from the training cohort. Testing cohort CISs were calculated for each patient relative to the patients in the training cohort, not relative to the patients in the other patients in the testing cohort.

Lung cancer subtypes

iSubGen was run on the 512 patients in the training cohort with any of the six data types: CNA, SNV, trinucleotide mutational signatures, methylation, mRNA abundance and miRNA abundance. All patients had at least two data types. If missing data, NA was used in the matrix. mRNA and miRNA features were *Z*-scaled per feature and trinucleotide mutational signatures features were $\log_{10}$-scaled before autoencoder training. For each data type, the autoencoder had three hidden layers with fifteen nodes, two nodes, fifteen nodes respectively. Therefore, there were two IRFs per data type. Consensus clustering was performed for 2 to 10 subtypes with 0.5 sub-sampling of features and patients. Diana, instead of hclust, was used within consensus clustering because it can handle clustering missing data. A weighting of 1:8 for CISs to independent reduced features was selected. The hyperparameter of two for subtype number was selected based on the number of histological types.

Using a subset of 126 patients (63 LUAD, 63 LUSC) with all six data types, we ran iSubGen as we did with the cohort including patients with missing data. Since the cohort with missing data has equivalent numbers of LUAD and LUSC patients, we downsampled to have a cohort with equal proportion of each subtype with all data types. There were 63 LUSC patients with all data types so we randomly selected 63 LUAD patients from the 212 LUAD patients with all data types.

There were 279 patients in the testing cohort with all five data types that were classified using centroid classification. Testing cohort independent reduced features were created using the trained autoencoders with mRNA and miRNA features scaled using the mean and standard deviations from the training cohort and trinucleotide mutational signatures features were again $\log_{10}$-scaled. Testing cohort CISs were calculated for each patient relative to the patients in the training cohort, not relative to the patients in the other patients in the testing cohort.

Robustness comparison to other algorithms

For robustness testing across different sample sizes, samples were randomly selected from the METABRIC cohort. For concatenated integrative subtyping, the features from the data types were combined and clustered using consensus clustering for 2 to 10 subtypes with 0.5 sub-sampling of features and patients, seed of 17 and 1000 reps. For COCA subtyping, consensus clustering was performed for each data type and the clusters with the maximum median silhouette coefficient were selected. Clustering results from each data type were encoded as dummy variables, combined and the combined matrix was clustered using consensus clustering. The consensus clustering used 0.5 subsampling of patients, seed 17 of 1000 reps. For iClusterBayes, the tuned.iClusterBayes function was used with Gaussian type. Since the other algorithms cannot all handle missing data, the lung cancer subtyping was run on the subset of the cohort that has all data types. Silhouette scores were computed as the mean silhouette coefficient of clustering algorithms and was calculated using silhouette_score function of the scikit-learn (v1.1.1) library. The AIC scores of AFT models were calculated using R package flexsurv (v 2.2.2) flexsurvreg function.[50]

Pan-cancer subgroups

Two subsets were randomly created with the twelve TCGA datasets (BLCA, BRCA, HNSC, KIRC, KIRP, LGG, LIHC, LUAD, PRAD, SKCM, STES, THCA) with more than 200 patients for the six data types (mRNA, CNA, SNV, trinucleotide mutational signatures, methylation, miRNA). From each cancer type, we randomly selected 300 patients or all the patients and evenly divided them between the two subsets. There were 1,709 patients in the first subset and 1,709 patients in the second subset. Both subsets were independently run through iSubGen. mRNA features were *Z*-scaled per feature before autoencoder training. For each data type, the autoencoder had three hidden layers with fifteen nodes, two nodes, fifteen nodes respectively. Therefore, there were two IRFs per data type. Consensus clustering was performed for 2 to 30 subtypes with 0.8 sub-sampling of features and 0.1 sub-sampling of patients. A weighting of 1:4 for CISs to independent reduced features was selected. The hyperparameter of fourteen for subtype number was selected in both subsets by assessing association of the iSubGen groups with cancer types using adjusted Rand index in the training cohort.

Visualization

All plotting was performed in the R statistical environment (v3.4.3) using the lattice (v0.20-38), latticeExtra (v0.6-28), RColorBrewer (v1.1-2) and cluster (v2.0.7–1) packages via the BPG library (v5.9.8).[51]