Cancer heterogeneity is defined by normal cellular trade-offs

Corey Weistuch,[1] Kevin A. Murgas,[2] Jiening Zhu,[3] Larry Norton,[4]
Ken A. Dill,[5] Joseph O. Deasy,[1] and Allen R. Tannenbaum[6,3]

[1]*Memorial Sloan Kettering Cancer Center, Department of Medical Physics*
[2]*Stony Brook University, Department of Biomedical Informatics*
[3]*Stony Brook University, Department of Applied Mathematics and Statistics*
[4]*Memorial Sloan Kettering Cancer Center, Department of Medicine*
[5]*Stony Brook University, Department of Biomedical Engineering*
[6]*Stony Brook University, Department of Computer Science*

(Dated: April 12, 2023)

Gene expression predicts tumor characteristics such as resistance to anticancer therapy. However, generalizing these predictors to multiple cancer types and data sets to motivate new therapeutic strategies has proven difficult. Here, we present a nonnegative matrix factorization (NMF) approach that decomposes gene expression into a universal set of "archetype" fingerprints. By restricting our analysis to five well-defined biological pathways, we show that trade-offs between normal tissues constrain oncogenic heterogeneity. Thus, the resulting six archetypes unify gene expression variation across 54 tissue types, 1504 cancer cell lines, and 1770 patient samples. The archetype mixtures correlate with cancer cell line sensitivity to several common anticancer therapies, even among cancers of the same type. They also explain subtype-specific breast cancer characteristics and define poor prognostic subgroups in breast, colorectal, and pancreatic cancers. Overall, the approach offers an evolvable resource for understanding commonalities across cancers, which could eventually lead to more robust therapeutic strategies.

INTRODUCTION

Tumors, even those derived from the same tissue type, can exhibit very different gene expression profiles and treatment responses. However, recent evidence has emerged that these patterns often follow similar rules across different cancers [1–3]. For example, many cancers exhibit subtypes that are relatively enriched for either glycolytic or oxidative metabolism (the "Warburg effect" [4, 5]). However, three fundamental questions remain unanswered. What do these patterns represent? How do we identify these patterns from molecular data? And how might we exploit them in the clinic?

A major aspect of cancer heterogeneity can be understood in terms of trade-offs. Trade-offs are ubiquitous in nature and explain, for example, why the human body requires a diversity of specialized tissues and why cancers adapt to different environmental conditions [5–9]. The Warburg effect exemplifies the concept of an oncogenic trade-off, or a constraint involving the gain of a tumorigenic quality in lieu of another [4, 10]. The Warburg phenotype is observed as an increase in glycolysis in exchange for reduced oxidative phosphorylation, which can promote tumor growth. This effect has been described in breast cancer across molecular subtypes, where basal-like tumors exhibit greater glycolytic metabolism compared to luminal-like breast tumors [1]. Another key oncogenic trade-off includes the migration/proliferation dichotomy, where typically proliferative cancer cells can adapt to a low-proliferation and high-motility state in response to hypoxia, which then may support the development of invasive metastatic disease [11].

Recently developed mathematical tools have enabled trade-offs to be identified in high-dimensional settings, such as cancer gene expression data. Here, the data are modeled as mixtures of a set of fully-specialized extrema called *archetypes* [12]. As a model, archetypes can be used to identify distinct patterns of cellular features associated with biological trade-offs and are capable of describing heterogeneity across numerous cell types [13, 14]. In cancer, these archetypes represent unique gene expression programs associated with enrichment for different biological pathways and distinct cancer hallmarks [8, 10, 14]. Improving our understanding of cancer archetypes could, in turn, enable future therapeutic strategies that exploit oncogenic trade-offs to improve outcomes and limit tumor evolution [10, 15, 16].

While archetypal analysis is fairly mature, its application to biology has not been fully realized [12]. Archetypal analysis typically involves a dimensionality-reduction technique such as nonnegative matrix factorization (NMF) or principal component analysis (PCA) [13, 17, 18]. Biological applications of archetypal analysis have relied primarily on a combination of PCA and convex hull optimization, which, when applied to nonnegative data such as RNA sequencing read counts, can produce spurious results and fail to identify archetypes even in highly heterogeneous cancer samples [13, 14, 19–23]. By contrast, NMF is widely used in many fields to identify latent archetypes in data by enforcing a nonnegativity constraint, providing an interpretable, low-dimensional approximation [17, 18, 24–26]. In previous biological analyses, NMF has been demonstrated as a robust tool for the classification of cancer subtypes and identifying context-dependent gene expression pro-

files [27–30].

The aim of this study is to leverage the archetypal analysis technique and the availability of multiple complementary gene expression datasets to comprehensively resolve cancer trade-offs. Our overarching strategy makes use of normalized nonnegative matrix factorization (N-NMF) to quantify the mixture of archetypes in a given gene expression sample (Fig. 1). We train our method on a variety of normal tissue samples to place the heterogeneity of cancer gene expression within the framework of normal cellular variation. In doing so, we demonstrate with two publicly available cancer data sets that multigene functional trade-offs are crucial to understanding the effects of common oncogenic mutations, the efficacy of chemotherapy, and the survival outcomes of patient subgroups across a range of cancer types. Ultimately, this study demonstrates a simple framework for informative and interpretable archetypal analysis of cancer gene expression data with potential for translational impact in oncology.

RESULTS

Archetype Analysis Captures Tissue-Specific Heterogeneity in Gene Expression

Previous applications of archetype analysis found that healthy cells lie closer to trade-off boundaries (i.e., they are more specialized) than tumors [19]. Therefore, we utilized the 54 distinct, healthy tissue samples from the Genotype-Tissue Expression (GTEx) project to cover the breadth of gene expression space and to provide an enhanced signal for resolving archetypes [31]. GTEx was established to characterize the tissue-specific determinants of human traits and diseases and provides expression levels for 44,219 different genes. Here, however, we honed in on commonly enriched pathways in cancer by utilizing 780 genes from five established biological pathways in MSigDB: apoptosis, DNA repair, glycolysis (combined with gluconeogenesis), hypoxia, and oxidative phosphorylation [32, 33]. This was essential for removing many tissue-specific genes that would hinder the generalizability of our archetypes to multiple types of cancer [34].

Following our general approach for archetype mixture analysis (see Methods: N-NMF), we identified six archetypes of pan-cellular gene expression (see SI Fig. 1), each associated with enrichment for different biological pathways (Fig. 2). Each tissue contained a different mixture of archetypes (Fig. 2a), with related tissues, such as the constituents of the cerebrum, having more similar archetype scores. These scores also reflected known functional trade-offs (Fig. 2b) across tissues, such as the sizable oxidative requirements of the brain and heart (Archetypes 4 and 6, [35]), the task of glycogen production in the liver (Archetype 5), and the protection and regulation of genetic material in the testis (Archetype 2). In addition, some tissues expressed multiple archetypes and thus balance several functions. The cerebellum, for example, is both resilient to aging (like the testis [36, 37]) and oxygen-demanding (like the brain, [38]), while the kidney is responsible for multiple broad cellular functions, expresses about 70% of the genes in the human body, and is enriched for far fewer proteins compared to most other tissues [39]. Overall, we observed complete agreement between the functional properties of each tissue and their relative mixture of archetypes.

Archetypes Associate Molecular Pathways with Cancer Characteristics

We next compiled expression levels, drug sensitivities (of 24 anti-cancer therapies), and mutation profiles from the Cancer Cell Line Encyclopedia (CCLE) [40, 41], and observed that the previous archetypes also predict pan-cancer disease characteristics (Figure 3). Each archetype, for example, was significantly associated with a distinct set of drug sensitivities (Fig. 3a) and mutations (Fig. 3b). The observed patterns of drug sensitivities naturally follow from the pathway-specific characteristics of each archetype: Archetype 1 is associated with apoptosis and sensitivity to IAP inhibitors (LBW242), Archetype 2 with DNA repair and sensitivity to topoisomerase inhibitors (Irinotecan and Topotecan), and Archetype 5 with glycolysis/gluconeogenesis and sensitivity to EGFR inhibitors (AZD0530, Lapatinib, and Erlotinib). Similarly, TP53 mutations disrupt cellular DNA repair and thus modulate Archetype 2. The mixture of archetypes in a given cancer type also depends, unsurprisingly, on the tissue lineage of origin (Fig. 3c). However, only select cancer lineages, such as those of the thyroid and liver, manifested archetypes consistent with their normal counterparts. Furthermore, recent studies have observed similar patterns of tumor initiation and proliferation in cancers of the pancreas, stomach, kidney, and liver [42]. Thus, the observed classification of different cancers in terms of their archetype mixtures warrants further study.

To test whether the previous sensitivity findings (Figs. 3a,b) were simply a result of known tissue-specific cancer characteristics (Fig. 3c), we repeated the previous analyses broken down by individual cancer lineages. SI Figure 2 shows the positive association between Irinotecan sensitivity and Archetype 2 (a) and Lapatinib sensitivity and Archetype 5 (b) across a few well-sampled cancer lineages. By contrast, mutations (SI Fig. 3) had a comparably smaller effect (a) and were poorly predictive of archetype scores as determined by a multilinear regression analysis (b). Thus, the effect size, compounded by the rarity of many of the mutations examined, was determined to be too small to be resolvable in individual cancer lineages. Overall, we found that archetype mix-

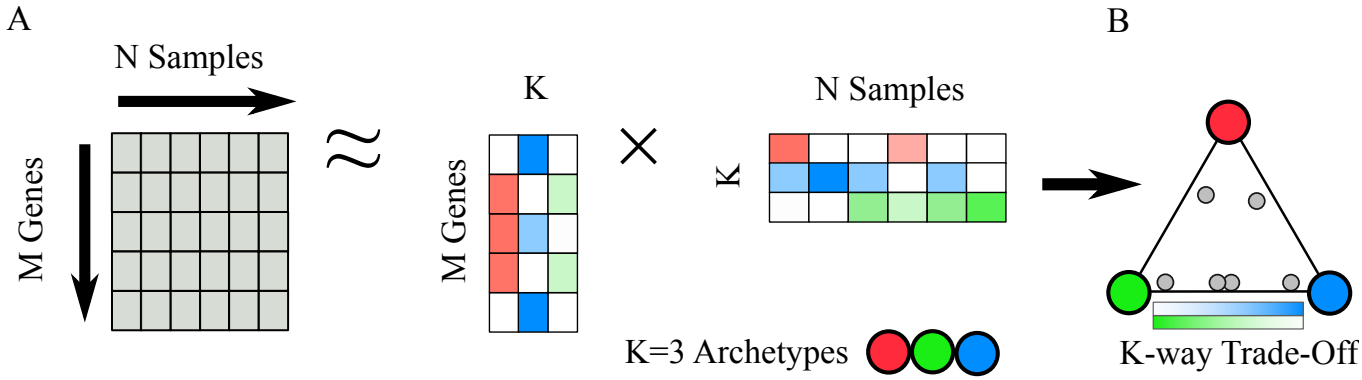


FIG. 1: **Normalized Nonnegative Matrix Factorization (N-NMF) identifies biological trade-offs**. a. N-NMF clusters the data (both the N samples and the M genes) into K (here 3) different groups based on their similarity to representative signatures called "archetypes". Unlike traditional clustering, each sample and gene can belong to multiple groups; the normalized gene and sample group weights, given by inferred matrices W and H respectively, are chosen to best approximate the original data (see Methods: Choosing the number of archetypes). b. The sample group weights (H) thus provide the coordinates (black) for a K-way archetype trade-off.

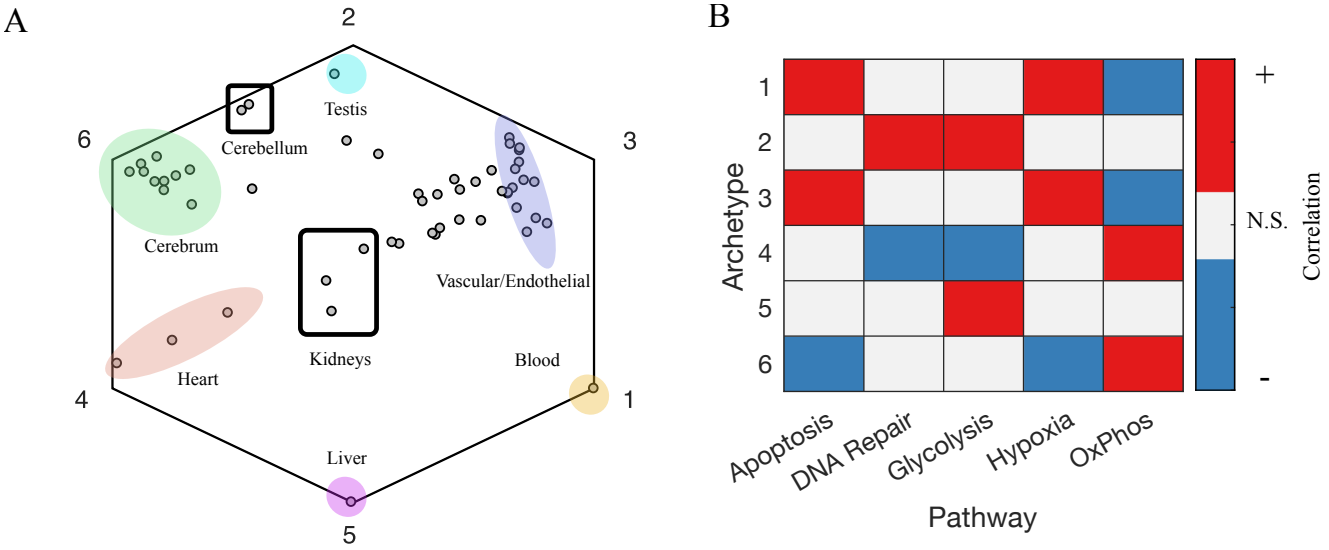


FIG. 2: **Six archetypes characterize gene expression heterogeneity across healthy tissues ($N = 54$)**. a. The median expression of each tissue (grey circle) from GTEx projected onto the six archetypes derived by N-NMF. Each archetype corresponds to a vertex of a centered, unit hexagon. Thus, tissues favoring a single archetype (e.g., liver) lie on the vertices, while tissues expressing a mixture of archetypes lie either on the edges (cerebellum) or in the center (kidneys). b. Hallmark expression pathways associated with each archetype ($p < 0.05$, t-test)

ture scores reliably predict some anti-cancer drug sensitivities, even within individual cancer types, and were only weakly associated with individual mutations.

Archetype Mixtures Reveal Non-Canonical Associations within Molecular Subtypes of Breast Cancer

Individual cancers are often subclassified based on their genomic characteristics, pathology, and treatment responses. In contrast to archetypes, however, this local approach to tumor classification can limit our ability to translate insights between cancer types. Thus, we next tested whether such differences can be more broadly understood in terms of our common archetype framework. Breast cancer provides a model system for such an analysis as it is well-studied and well-classified into five standard subtypes [43]: normal-like, luminal A and B, basal, and HER2+. To augment the limited breast cell lines from CCLE ($N = 63$), however, we first projected the breast patient expression data ($N = 1111$) from the Cancer Genome Atlas (TCGA) onto the previous archetypes

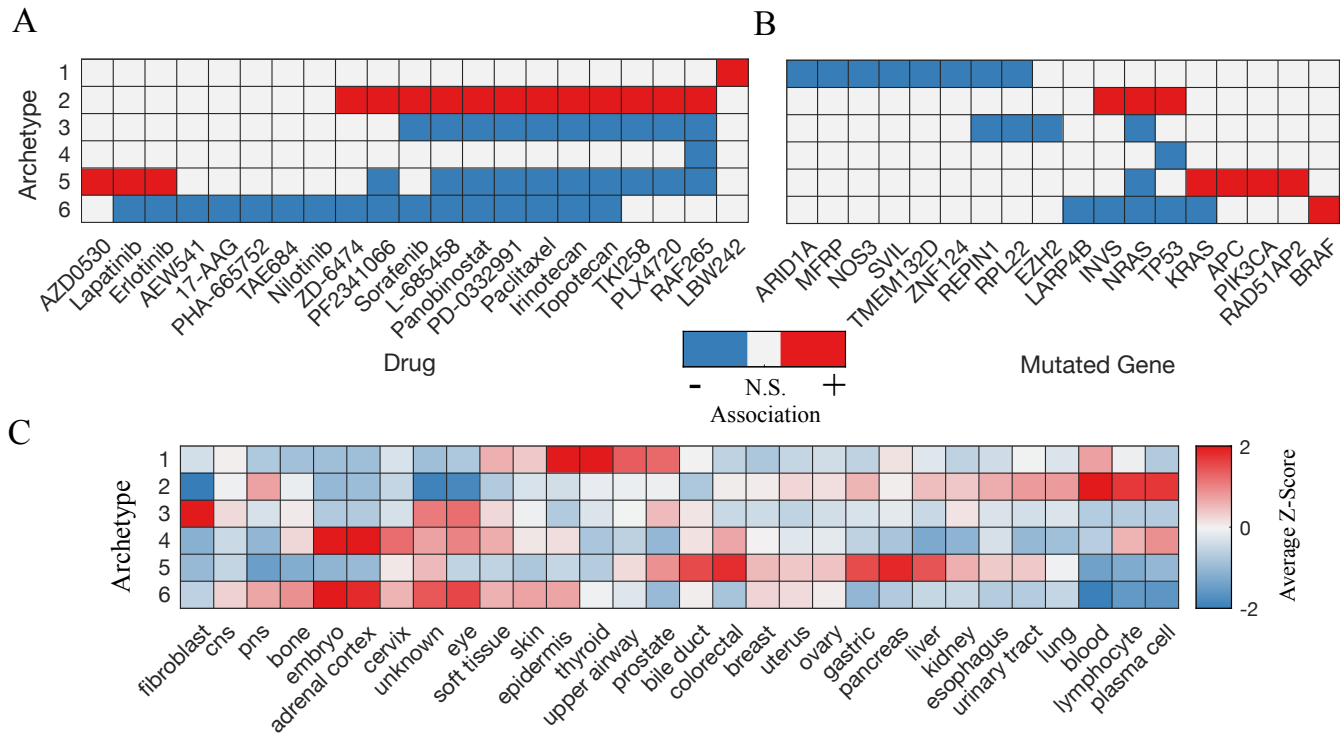


FIG. 3: **Archetype mixtures quantify oncogenic characteristics across cancer cell lines**. Archetype mixture scores were computed by aligning CCLE expression profiles to GTEx and projecting the resulting expression levels onto the six previously-identified archetypes (see SI Figure 1). Shown are the significant (Spearman) correlations of CCLE ($N = 1405$, $p < 0.05$, t-test) archetype scores with **(a)** the sensitivities (quantified using activity areas [40]) of 21/24 screened anticancer drugs and **(b)** 18/73 recurrent (TCGA hotspot) cancer mutations. Positive (red) associations depict drug sensitivity and high mutation frequency, respectively; negative (blue) associations relay the opposite trend. **(c).** Average Z-score (across all cell lines from the same tissue type) of each archetype. While several lineages (and thus cancer types) are preferentially associated with specific archetypes (red), these associations are less well defined compared to healthy tissues (see Fig. 2).

([44–47], see Methods: N-NMF). This provided both an enhanced sample size and a validation of our comparisons in cell lines and direct patient samples.

While the subtypes provided in each data set differ slightly, we observed common subtype-specific patterns in both Archetype 2 and 5 (Figure 4). Compared to other breast subtypes (particularly luminal A, [1]), basal tumors were enriched for glycolysis and thus associated with higher levels of Archetype 2 (Figs. 4a,b). Since luminal B tumors have characteristics of both luminal A and basal subtypes, they intuitively had intermediate Archetype 2 scores [48]. HER2+ tumors, on the other hand, overexpress HER2, a protein related to EGFR, potentially explaining why they were enriched for Archetype 5 compared to other breast cancer subtypes (Figs. 4c,d). Deviating from the standard breast classification scheme, CCLE further splits basal breast cancer into types A and B, with basal A being described as having more luminal-like properties [49]. However, comparing Archetype 5 (Fig. 4c) suggests that basal A tumors also have characteristics in line with HER2+ positive tumors and that these features are distinct from the canonical luminal-to-basal spectrum [49]. More broadly, the substantial variability in archetype mixture scores within individual subtypes suggests that the current classification scheme may group together cancers with very different characteristics that could be further distinguished using archetypes.

Archetypes Delineate Cohorts with Distinct Survival Advantages

As archetypes encode information about tumor drug sensitivities and enriched biological pathways, we hypothesized that could also explain, in part, differences in patient survival within specific cancers. Figure 5 shows the Kaplan-Meier survival curves of these distinct survival groups found for three different TCGA cohorts: Breast Invasive Carcinoma (5a, $N = 1111$), Colon Ade-

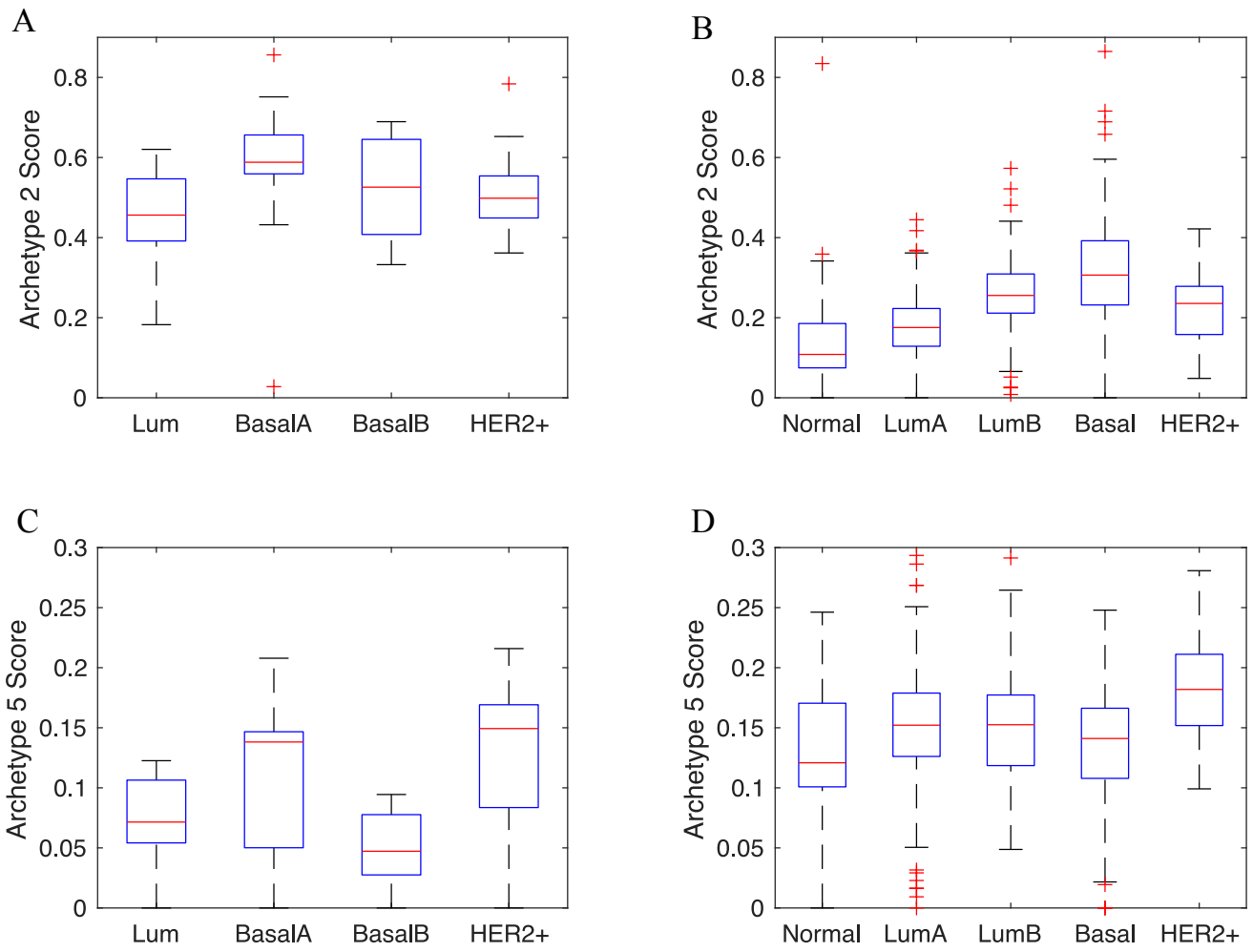


FIG. 4: **Archetype composition varies across breast cancer subtypes**. Shown are box plots of the Archetype 2 (a,b) and Archetype 5 (c,d) scores across breast cancer cell lines in CCLE (a,c) and primary tumors in TCGA (b,d). Red asterisks denote outlier samples with archetype scores beyond 1.5 times the interquartile range away from the box.

nocarcinoma (5b, $N = 481$), and Pancreatic Adenocarcinoma (5c, $N = 178$).

We found the best separation to be achieved using a different archetype (2, 4, and 6, respectively) for each cohort. However, we note that these three archetypes are intimately related: Archetype 2 is associated with glycolysis, while 4 and 6 are associated with oxidative phosphorylation. Thus, the survival groups quantify the observation that Warburg (Archetype 2) tumors tend to be far more aggressive than oxidative tumors (Archetypes 4 and 6) [4, 7]. As seen for breast cancer (5a), intermediate archetype scores can also reveal novel vulnerable groups. Luminal B breast cancer, associated with an intermediate phenotype between luminal A (oxidative) and basal (Warburg), is typically associated with worse patient outcomes [48]. However, we observed poorer survival in this intermediate archetype group even when luminal B samples were removed (SI Figure 4a). More strikingly, these survival differences were more significant than even those between the original PAM50 subgroups (SI Figure 4b). Thus, archetype mixture analysis may provide a complementary and, in some ways, improved classification scheme for breast cancer.

DISCUSSION AND CONCLUSIONS

We have presented a novel strategy for resolving the trade-offs between multi-gene signatures, also known as archetypes, and how they affect particular tumor characteristics. According to our analysis, six archetypes model gene expression heterogeneity in healthy and cancer cells.

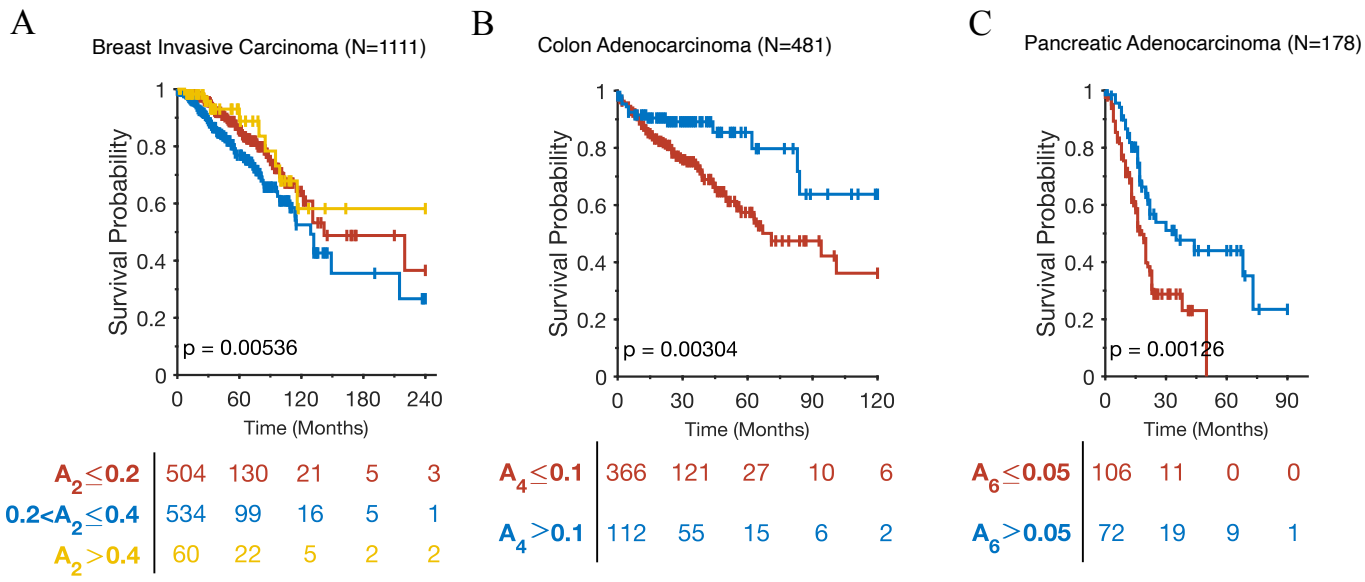


FIG. 5: **Archetype mixture scores distinguish groups with poor survival in (a) Breast, (b) Colon, and (c) Pancreatic cancers from TCGA.** Kaplan-Meier survival curves were separated according to bins of Archetype score. Tables below indicate the number of surviving patients at each time point. Significance was determined using a log-rank test.

In contrast to previous studies, we demonstrated our model on three independent data sets to show its applicability in quantifying the heterogeneity of healthy tissues and different cancers. Moreover, despite the fact that the current study's goal was purely exploratory, we discovered that the archetypes identified by our method were strongly associated with various tumor characteristics, such as drug sensitivities and cancer subtypes. This, we underscore, may enable broad and direct application to future gene expression studies.

Several predictions of our unsupervised archetype framework have also been confirmed by recent clinical studies. Sorafenib, for example, is predicted to be effective only against tumors expressing Archetype 2. This comes as no surprise given that Sorafenib is known to provide little clinical efficacy against melanoma, efficacy at high doses against hepatocellular carcinoma and kidney cancer, and efficacy at lower doses in acute myeloid leukemia, which is in agreement with the observed Archetype 2 scores [50–54]. Given this agreement, archetypal analysis may then be able to provide useful a priori information for clinical therapeutic strategy and future clinical trials.

A few archetypes we found are weakly associated with gene mutations (such as BRAF and PIK3CA) known to confer specific drug resistances [55]. One possible explanation is that these mutations drive a tumor away from its default archetype (the one being targeted) and into a new archetype [14]. Consequently, such mutations may inadvertently sensitize tumors to drugs associated with their new archetype composition, a phenomenon referred to as collateral drug sensitivity [56, 57]. While such phenomena were not detectable in the current study, a future experiment could target drug sensitivity genes in cell lines to more directly examine the resulting archetype perturbations.

The archetype mixture model was trained using healthy tissues to maximize signal-to-noise and functionally validate the archetypes against known cellular trade-offs. However, tumors may also exhibit additional clinically significant archetypes not observed in healthy tissues [10]. While we restricted our gene set to five hallmark pathways of broad relevance to cancer, additional archetypes are also likely contained in the remaining genes. However, even with these restrictions, our archetypes are in broad agreement with known cancer hallmarks [7] and tumor archetypes [10]. Furthermore, the utilization of N-NMF allows our framework to be easily augmented with additional archetypes and data constraints while preserving our current archetypes [26, 58].

Our framework can also be extended to probe intratumor heterogeneity through the use of single-cell gene expression data. Given the observed applicability to different sources of data, the current archetypes may be useful for identifying treatment-resistant cells [10]. Other components of the tumor microenvironment, such as immune cells and fibroblasts, could be identified as well [59, 60].

As previously discussed, N-NMF can identify archetypes while avoiding some of the methodological limitations of other methods. However, N-NMF requires that the inferred archetype weights sum to 1. This can introduce spurious gene anticorrelations into the data and make it more sensitive to differences in data pre-processing. As a result, archetypes should be independently validated through known biological and functional experiments, as is done here and in any unsupervised analysis.

In conclusion, our work identifies broad patterns of gene expression heterogeneity, observed across normal cells and different cancers, that predict cellular characteristics and patient outcomes. By doing so, we summarize the numerous genetic differences among cells with a few quantitative biomarkers that can be more easily understood and more systematically perturbed in a future clinical setting.

METHODS

Gene Expression Data

GTEx

GTEx (version 8) gene-level TPM-normalized expression data were downloaded from the GTEx Portal [31]. The data set, consisting of 54 distinct tissues, is one of the most comprehensive resources for studying tissue-specific gene expression. To minimize bias in our archetypes toward highly-expressed genes, however, we divided the expression of each gene by its standard deviation across tissues. These expression levels were then normalized across the genes in each tissue to prepare the data for archetype analysis.

CCLE

Gene-level TPM-normalized expression from the Cancer Cell Line Encyclopedia (CCLE, $N = 1405$) along with matched drug sensitivities ($N = 469$) and mutations ($N = 1250$) were downloaded from the public Dependency Map (DepMap) portal (version 22Q2) [40].

TCGA

Gene-level TPM-normalized expression data from The Cancer Genome Atlas (TCGA) within the breast (BRCA, $N = 1111$), colon (COAD, $N = 481$), and pancreatic (PAAD, $N = 178$) cancer cohorts were downloaded via the TCGAbiolinks R package [44–47]. Samples were restricted to primary tumors.

Normalized Nonnegative Matrix Factorization (N-NMF)

As already described, the key methodology used in this paper is *nonnegative matrix factorization* (NMF). In general, the heterogeneous data collected in medicine represent the integrated result of several inter-related variables or are combinations of several latent components, or factors. Such complex datasets must be decomposed into their underlying components to find key structures and extract hidden information. Thus, approximate low-rank matrix and tensor factorizations play a pivotal role in enhancing the data. This impacts many major medical data problems, including dimension reduction, discrimination, and clustering. Nonnegative matrix factorization, in particular, has led to numerous applications in biology, as previously discussed. We now give the mathematical details of the exact formulation we employed in the present work.

Normalized non-negative matrix factorization (N-NMF) is applied as a tool to establish a low-dimensional representation of the variability in gene expression profiles for a given dataset. The $N \times M$ data matrix V, representing N gene expression values of M samples, is approximated as the product of two low-rank matrices W and H:

$$V \approx WH, \tag{1}$$

where W is an $N \times k$ matrix representing coefficients of each gene contribution to k archetypes, and H is a $k \times M$ matrix representing weights of each archetype to approximate each sample of gene expression.

More rigorously, the matrices W and H are found by minimizing the error under the Frobenius norm:

$$\min_{W>0, H>0} ||V - WH||_F$$

We recall that for a general $m \times n$ matrix A, the *Frobenius norm* is the square root of the sum of the absolute squares of its elements.

We fit the NMF approximation with an iterative algorithm that alternates updating the W and H matrices for each iteration and exhibits monotonic convergence similar to an Expectation-Maximization (EM) algorithm [24]. The iterative algorithm applies the following update rules:

$$\begin{aligned} W' &= W \odot (VH' \odot \frac{1}{WHH'}), \\ H' &= H \odot (W'V \odot \frac{1}{W'WH}), \end{aligned} \tag{2}$$

where $\odot$ indicates element-wise multiplication. To avoid degeneracy in fitting, the H matrix is normalized after each iteration such that the coefficients of each archetype score sum to 1 for each sample.

After determining the ideal number of archetypes as rank k (see below), we randomly initialize the W and H matrices and train the archetype approximation over 1,000 iterations. Subsequently, the trained archetypes are used to define archetype scores on a new dataset (i.e., CCLE and TCGA data) with matched gene order, using the same iterative algorithm but keeping the W gene-archetype coefficient matrix fixed.

Choosing the number of archetypes

The optimal number of archetypes was chosen following the method of profile log-likelihood [61]. This models the unknown number of archetypes as a latent variable that can be directly optimized over. This method was selected due to its simplicity to implement and its superior performance compared to multiple other methods for selecting the number of factors in NMF [62]. For N-NMF, the minimum number of factors was set as $k = 2$ (due to the normalization constraint), and the maximum (a requirement of the approach) was chosen to be $k = 30$. The optimal number of archetypes was then determined as the number that maximized the resulting profile log-likelihood ($k = 6$ for GTEx, see SI Fig 1).

Statistical Analysis

For statistical comparison, we utilize the Spearman's rank correlation t-test, following the Benjamini-Hochberg procedure for multiple comparisons with a significance threshold of FDR< 0.05 [63].

SUPPLEMENTARY INFORMATION

Acknowledgments

Research presented here was funded by the following: the Marie-Josée Kravis Fellowship in Quantitative Biology (CW), the Laufer Center for Physical and Quantitative Biology (KAD), AFOSR grants FA9550-20-1-0029 and FA9550-23-1-0096 (ART), NIH grant R01-AG048769 (ART), a grant from Breast Cancer Research Foundation BCRF-17-193 (LN, JOD, and ART), Army Research Office grant W911NF2210292 (ART), and a grant from the Cure Alzheimer's Foundation (ART).

Author contributions

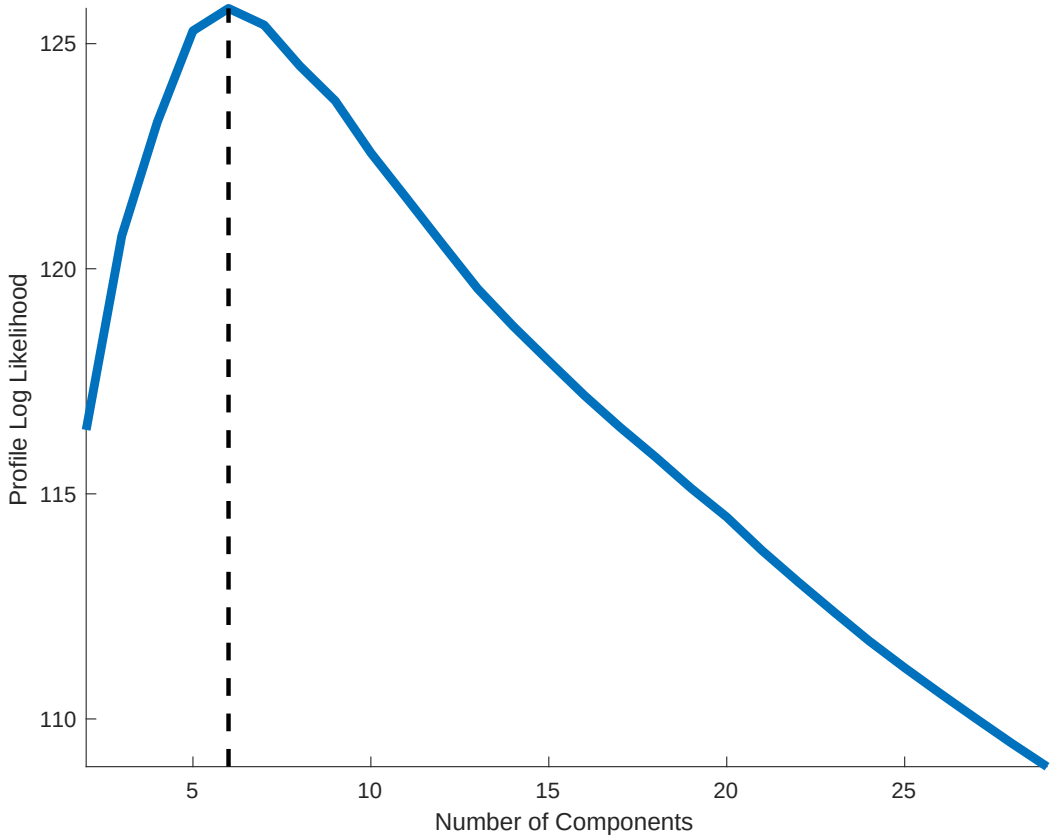
Conceptualization: C.W. and K.A.D; Methodology: C.W., K.A.M, and J.Z; Analysis and investigation: C.W. and K.A.M.; Writing: C.W. and K.A.M.; Editing: C.W., K.A.M., J.Z., L.N., K.A.D., J.O.D., A.R.T; Supervision: J.O.D. and A.R.T.; Funding acquisition: C.W., L.N. J.O.D., and A.R.T.

Competing interests

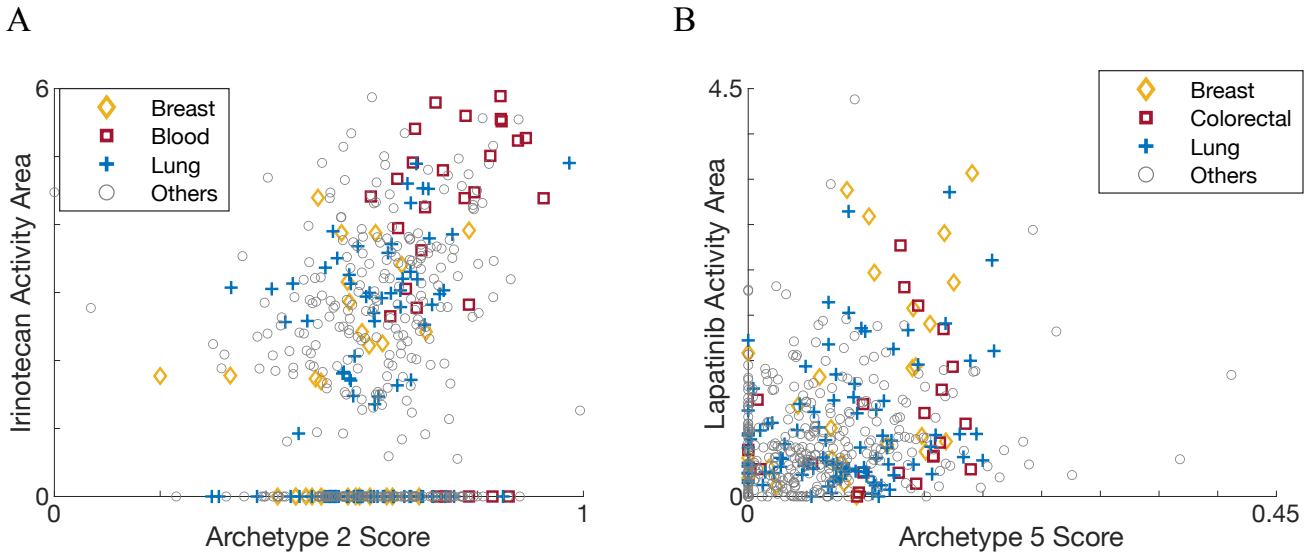
The authors declare no competing interests.

Data and code availability

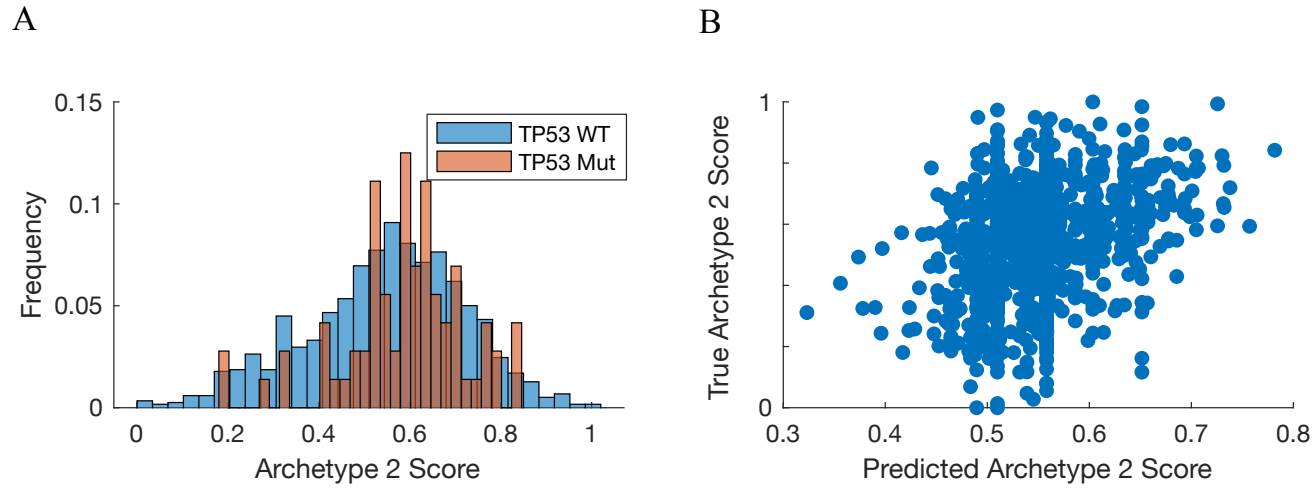
The source data and MATLAB code that support the findings of this study are available in Figshare with the identifier https://doi.org/10.6084/m9.figshare.22592737.



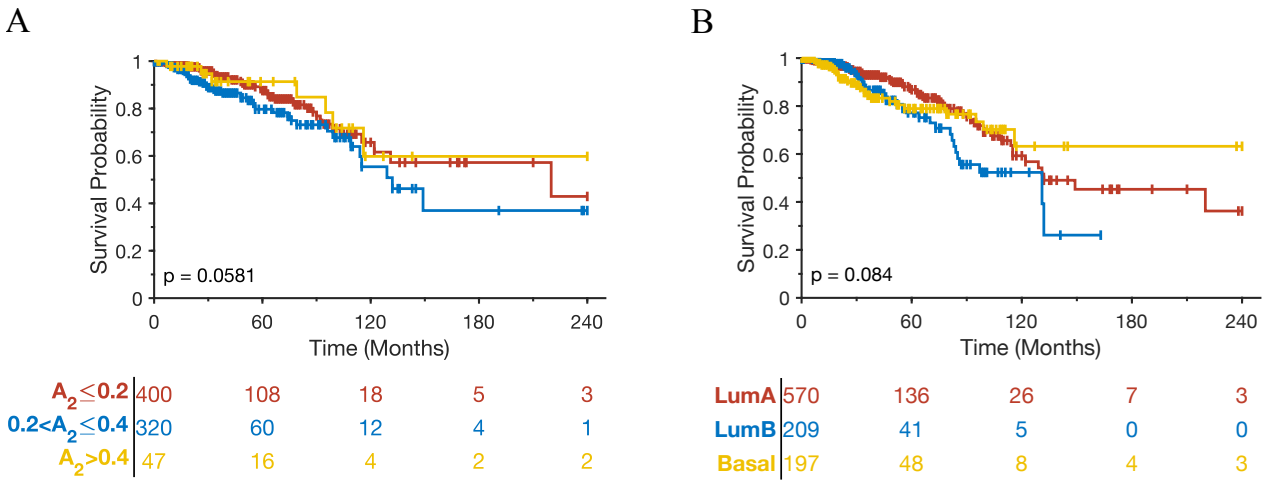
SI1: **Profile log-likelihood analysis to determine optimal N-NMF rank** k. Log-likelihood was computed over each k number of components to optimize the N-NMF algorithm when approximating the GTEx dataset. The black dotted line represents the maximum profile log-likelihood at the optimal $k = 6$.



SI2: **Archetypes predict anti-cancer drug sensitivities within different cancer types.** Shown are scatter plots of Archetype 2 score against Irinotecan activity area (a) and Archetype 5 score against Lapatinib activity area (b) across several representative cancer types in CCLE.



SI3: **Mutations are weakly associated with archetype scores.** (a). Histograms of archetype 2 scores in TP53 mutated (orange) and wild type (blue) cell lines. (b). Scatter plot of true versus predicted archetype 2 scores using mutation status alone. Predicted values were determined from a standard multivariate linear regression on 73 recurrent hotspot mutations annotated as part of CCLE ($R^2 = 0.0967$).



SI4: **Archetypes refine survival groups in breast cancer.** Kaplan-Meier curves of basal and luminal breast cancers stratified by Archetype 2 (a) compared to those stratified by breast cancer subtype (b). Significance was determined using a log-rank test.

REFERENCES

[1] Paola Lunetti, Mariangela Di Giacomo, Daniele Vergara, Stefania De Domenico, Michele Maffia, Vincenzo Zara, Loredana Capobianco, and Alessandra Ferramosca. Metabolic reprogramming in breast cancer results in distinct mitochondrial bioenergetics between luminal and basal subtypes. *The FEBS journal*, 286(4):688–709, 2019.

[2] Xinxin Peng, Zhongyuan Chen, Farshad Farshidfar, Xiaoyan Xu, Philip L Lorenzi, Yumeng Wang, Feixiong Cheng, Lin Tan, Kamalika Mojumdar, Di Du, et al. Molecular characterization and clinical relevance of metabolic expression subtypes in human cancers. *Cell reports*, 23(1):255–269, 2018.

[3] Agustín González-Reymúndez and Ana I Vázquez. Multi-omic signatures identify pan-cancer classes of tumors beyond tissue of origin. *Scientific reports*, 10(1):1–13, 2020.

[4] Maria V Liberti and Jason W Locasale. The warburg effect: how does it benefit cancer cells? *Trends in biochemical sciences*, 41(3):211–218, 2016.

[5] Mehdi Damaghi, Jeffrey West, Mark Robertson-Tessi, Liping Xu, Meghan C Ferrall-Fairbanks, Paul A Stewart, Erez Persi, Brooke L Fridley, Philipp M Altrock, Robert A Gatenby, et al. The harsh microenvironment in early breast cancer selects for a warburg phenotype. *Proceedings of the National Academy of Sciences*, 118(3):e2011342118, 2021.

[6] Douglas J Futuyma and Gabriel Moreno. The evolution of ecological specialization. *Annual review of Ecology and Systematics*, pages 207–233, 1988.

[7] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[8] Douglas Hanahan. Hallmarks of cancer: new dimensions. *Cancer discovery*, 12(1):31–46, 2022.

[9] Mel Greaves and Carlo C Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012.

[10] Jean Hausser and Uri Alon. Tumour heterogeneity and the evolutionary trade-offs of cancer. *Nature Reviews Cancer*, 20(4):247–257, 2020.

[11] Haralampos Hatzikirou, David Basanta, Matthias Simon, K Schaller, and Andreas Deutsch. 'go or grow': the key to the emergence of invasion in tumour progression? *Mathematical medicine and biology: a journal of the IMA*, 29(1):49–65, 2012.

[12] Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

[13] Oren Shoval, Hila Sheftel, Guy Shinar, Yuval Hart, Omer Ramote, Avi Mayo, Erez Dekel, Kathryn Kavanagh, and Uri Alon. Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space. *Science*, 336(6085):1157–1160, 2012.

[14] Jean Hausser, Pablo Szekely, Noam Bar, Anat Zimmer, Hila Sheftel, Carlos Caldas, and Uri Alon. Tumor diversity and the trade-off between universal cancer tasks. *Nature Communications*, 10(1):5423, 2019.

[15] Joseph M Chan, Samir Zaidi, Jillian R Love, Jimmy L Zhao, Manu Setty, Kristine M Wadosky, Anuradha Gopalan, Zi-Ning Choo, Sitara Persad, Jungmin Choi, et al. Lineage plasticity in prostate cancer depends on jak/stat inflammatory signaling. *Science*, 377(6611):1180–1191, 2022.

[16] Sarah M Groves, Geena V Ildefonso, Caitlin O McAtee, Patricia MM Ozawa, Abbie S Ireland, Philip E Stauffer, Perry T Wasdin, Xiaomeng Huang, Yi Qiao, Jing Shan Lim, et al. Archetype tasks link intratumoral heterogeneity to plasticity and cancer hallmarks in small cell lung cancer. *Cell Systems*, 13(9):690–710, 2022.

[17] Anil Damle and Yuekai Sun. A geometric approach to archetypal analysis and nonnegative matrix factorization. *Technometrics*, 59(3):361–370, 2017.

[18] Hamid Javadi and Andrea Montanari. Nonnegative matrix factorization via archetypal analysis. *Journal of the American Statistical Association*, 115(530):896–907, 2020.

[19] Yuval Hart, Hila Sheftel, Jean Hausser, Pablo Szekely, Noa Bossel Ben-Moshe, Yael Korem, Avichai Tendler, Avraham E Mayo, and Uri Alon. Inferring biological tasks using pareto analysis of high-dimensional data. *Nature methods*, 12(3):233–235, 2015.

[20] Yael Korem, Pablo Szekely, Yuval Hart, Hila Sheftel, Jean Hausser, Avi Mayo, Michael E Rothenberg, Tomer Kalisky, and Uri Alon. Geometry of the gene expression space of individual cells. *PLoS computational biology*, 11(7):e1004224, 2015.

[21] Mengyi Sun and Jianzhi Zhang. Rampant false detection of adaptive phenotypic optimization by parti-based pareto front inference. *Molecular biology and evolution*, 38(4):1653–1664, 2021.

[22] Corey Weistuch, Jiening Zhu, Joseph O Deasy, and Allen R Tannenbaum. The maximum entropy principle for compositional data. *BMC bioinformatics*, 23(1):1–13, 2022.

[23] Miri Adler, Avichai Tendler, Jean Hausser, Yael Korem, Pablo Szekely, Noa Bossel, Yuval Hart, Omer Karin, Avi Mayo, and Uri Alon. Controls for phylogeny and robust analysis in pareto task inference. *Molecular biology and evolution*, 39(1):msab297, 2022.

[24] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[25] Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

[26] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on knowledge and data engineering*, 25(6):1336–1353, 2012.

[27] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, 101(12):4164–4169, 2004.

[28] Richard A Moffitt, Raoud Marayati, Elizabeth L Flate, Keith E Volmar, S Gabriela Herrera Loeza, Katherine A Hoadley, Naim U Rashid, Lindsay A Williams, Samuel C Eaton, Alexander H Chung, et al. Virtual microdissection identifies distinct tumor-and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature genetics*, 47(10):1168–1178, 2015.

[29] Gabriela S Kinker, Alissa C Greenwald, Rotem Tal, Zhanna Orlova, Michael S Cuoco, James M McFarland, Allison Warren, Christopher Rodman, Jennifer A Roth, Samantha A Bender, et al. Pan-cancer single-cell rna-

seq identifies recurring programs of cellular heterogeneity. *Nature genetics*, 52(11):1208–1218, 2020.

[30] Dalia Barkley, Reuben Moncada, Maayan Pour, Deborah A Liberman, Ian Dryg, Gregor Werba, Wei Wang, Maayan Baron, Anjali Rao, Bo Xia, et al. Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nature Genetics*, 54(8):1192–1201, 2022.

[31] GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[32] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[33] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[34] Katherine A Hoadley, Christina Yau, Toshinori Hinoue, Denise M Wolf, Alexander J Lazar, Esther Drill, Ronglai Shen, Alison M Taylor, Andrew D Cherniack, Vésteinn Thorsson, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304, 2018.

[35] Charles B Cairns, James Walther, Alden H Harken, and Anirban Banerjee. Mitochondrial oxidative phosphorylation thermodynamic efficiencies reflect physiological organ roles. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 274(5):R1376–R1383, 1998.

[36] Hunter B Fraser, Philipp Khaitovich, Joshua B Plotkin, Svante Pääbo, and Michael B Eisen. Aging and gene expression in the primate brain. *PLoS biology*, 3(9):e274, 2005.

[37] Xichen Nie, Sarah K Munyoki, Meena Sukhwani, Nina Schmid, Annika Missel, Benjamin R Emery, Jan-Bernd Stukenborg, Artur Mayerhofer, Kyle E Orwig, Kenneth I Aston, et al. Single-cell analysis of human testis aging and correlation with elevated body mass index. *Developmental Cell*, 57(9):1160–1176, 2022.

[38] Clare Howarth, Padraig Gleeson, and David Attwell. Updated energy budgets for neural computation in the neocortex and cerebellum. *Journal of Cerebral Blood Flow & Metabolism*, 32(7):1222–1232, 2012.

[39] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[40] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[41] Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, Barbara A Weir, Gregory Kryukov, Glenn S Cowley, Stanley Gill, William F Harrington, Sasha Pantel, John M Krill-Burger, et al. Defining a cancer dependency map. *Cell*, 170(3):564–576, 2017.

[42] Spencer G Willet, Mark A Lewis, Zhi-Feng Miao, Dengqun Liu, Megan D Radyk, Rebecca L Cunningham, Joseph Burclaff, Greg Sibbel, Hei-Yong G Lo, Valerie Blanc, et al. Regenerative proliferation of differentiated cells by mtorc 1-dependent paligenosis. *The EMBO journal*, 37(7):e98311, 2018.

[43] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160, 2009.

[44] Brigham & Women's Hospital & Harvard Medical School Chin Lynda 9 11 Park Peter J. 12 Kucherlapati Raju 13, Genome data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, Institute for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[45] Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330, 2012.

[46] Benjamin J Raphael, Ralph H Hruban, Andrew J Aguirre, Richard A Moffitt, Jen Jen Yeh, Chip Stewart, A Gordon Robertson, Andrew D Cherniack, Manaswi Gupta, Gad Getz, et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell*, 32(2):185–203, 2017.

[47] Antonio Colaprico, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, Isabella Castiglioni, et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, 44(8):e71–e71, 2016.

[48] Chad J Creighton. The molecular profile of luminal b breast cancer. *Biologics: Targets and Therapy*, pages 289–297, 2012.

[49] Xiaofeng Dai, Hongye Cheng, Zhonghu Bai, and Jia Li. Breast cancer cell line classification and its relevance with breast tumor subtyping. *Journal of Cancer*, 8(16):3131, 2017.

[50] T Eisen, T Ahmad, KT Flaherty, M Gore, S Kaye, R Marais, I Gibbens, S Hackett, M James, LM Schuchter, et al. Sorafenib in advanced melanoma: a phase ii randomised discontinuation trial analysis. *British journal of cancer*, 95(5):581–586, 2006.

[51] Bernard Escudier, Tim Eisen, Walter M Stadler, Cezary Szczylik, Stéphane Oudard, Michael Siebels, Sylvie Negrier, Christine Chevreau, Ewa Solska, Apurva A Desai, et al. Sorafenib in advanced clear-cell renal-cell carcinoma. *New England Journal of Medicine*, 356(2):125–134, 2007.

[52] Josep M Llovet, Sergio Ricci, Vincenzo Mazzaferro, Philip Hilgard, Edward Gane, Jean-Frédéric Blanc, Andre Cosme De Oliveira, Armando Santoro, Jean-Luc Raoul, Alejandro Forner, et al. Sorafenib in advanced hepatocellular carcinoma. *New England journal of medicine*, 359(4):378–390, 2008.

[53] Tao Liu, Vijay Ivaturi, Philip Sabato, Jogarao VS Gobburu, Jacqueline M Greer, John J Wright, B Douglas Smith, Keith W Pratz, Michelle A Rudek, and ETCTN-

6745 study team. Sorafenib dose recommendation in acute myeloid leukemia based on exposure-flt3 relationship. *Clinical and translational science*, 11(4):435–443, 2018.

[54] Maria Larrosa-Garcia and Maria R Baer. Flt3 inhibitors in acute myeloid leukemia: current status and future directions. *Molecular cancer therapeutics*, 16(6):991–1001, 2017.

[55] Neil Vasan, José Baselga, and David M Hyman. A view on drug resistance in cancer. *Nature*, 575(7782):299–309, 2019.

[56] Kristen M Pluchino, Matthew D Hall, Andrew S Goldsborough, Richard Callaghan, and Michael M Gottesman. Collateral sensitivity as a strategy against cancer multidrug resistance. *Drug Resistance Updates*, 15(1-2):98–105, 2012.

[57] Ahmet Acar, Daniel Nichol, Javier Fernandez-Mateos, George D Cresswell, Iros Barozzi, Sung Pil Hong, Nicholas Trahearn, Inmaculada Spiteri, Mark Stubbs, Rosemary Burke, et al. Exploiting evolutionary steering to induce collateral drug sensitivity in cancer. *Nature communications*, 11(1):1–14, 2020.

[58] Russell Z Kunes, Thomas Walle, Tal Nawy, and Dana Pe'er. Supervised discovery of interpretable gene programs from single-cell data. *bioRxiv*, 2022.

[59] Elham Azizi, Ambrose J Carr, George Plitas, Andrew E Cornish, Catherine Konopacki, Sandhya Prabhakaran, Juozas Nainys, Kenmin Wu, Vaidotas Kiseliovas, Manu Setty, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell*, 174(5):1293–1308, 2018.

[60] Aimy Sebastian, Nicholas R Hum, Kelly A Martin, Sean F Gilmore, Ivana Peran, Stephen W Byers, Elizabeth K Wheeler, Matthew A Coleman, and Gabriela G Loots. Single-cell transcriptomic analysis of tumor-derived fibroblasts and normal tissue-resident fibroblasts reveals fibroblast heterogeneity in breast cancer. *Cancers*, 12(5):1307, 2020.

[61] Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, 2006.

[62] José M Maisog, Andrew T DeMarco, Karthik Devarajan, Stanley Young, Paul Fogel, and George Luta. Assessing methods for evaluating the number of components in non-negative matrix factorization. *Mathematics*, 9(22):2840, 2021.

[63] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.