

The Cancer States and Archetypes (CSA) and Celsus Data Analysis Methodologies

(Summer Student Internships 2024)

Pablo Tamayo

UC San Diego Moores Cancer Center and Center for Novel Therapeutics
Division of Genomics and Precision Medicine, UC San Diego School of Medicine

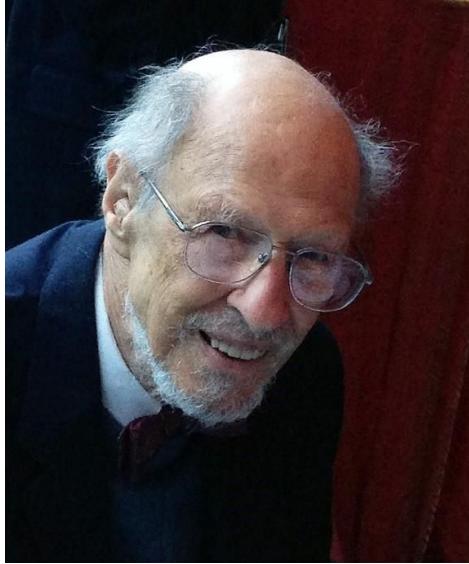
UCSD Center for Cancer Target Discovery and Development (CTD2)
Moores Cancer Center Computational Cancer Analysis Laboratory (CCAL)
California Initiative to Advance Precision Medicine UCSD project

Cancer Program, Broad Institute of MIT/Harvard



UCSD Center for Cancer Target
Discovery and Development





Fools ignore complexity.

Pragmatists suffer it.

Some can avoid it.

Geniuses remove it.

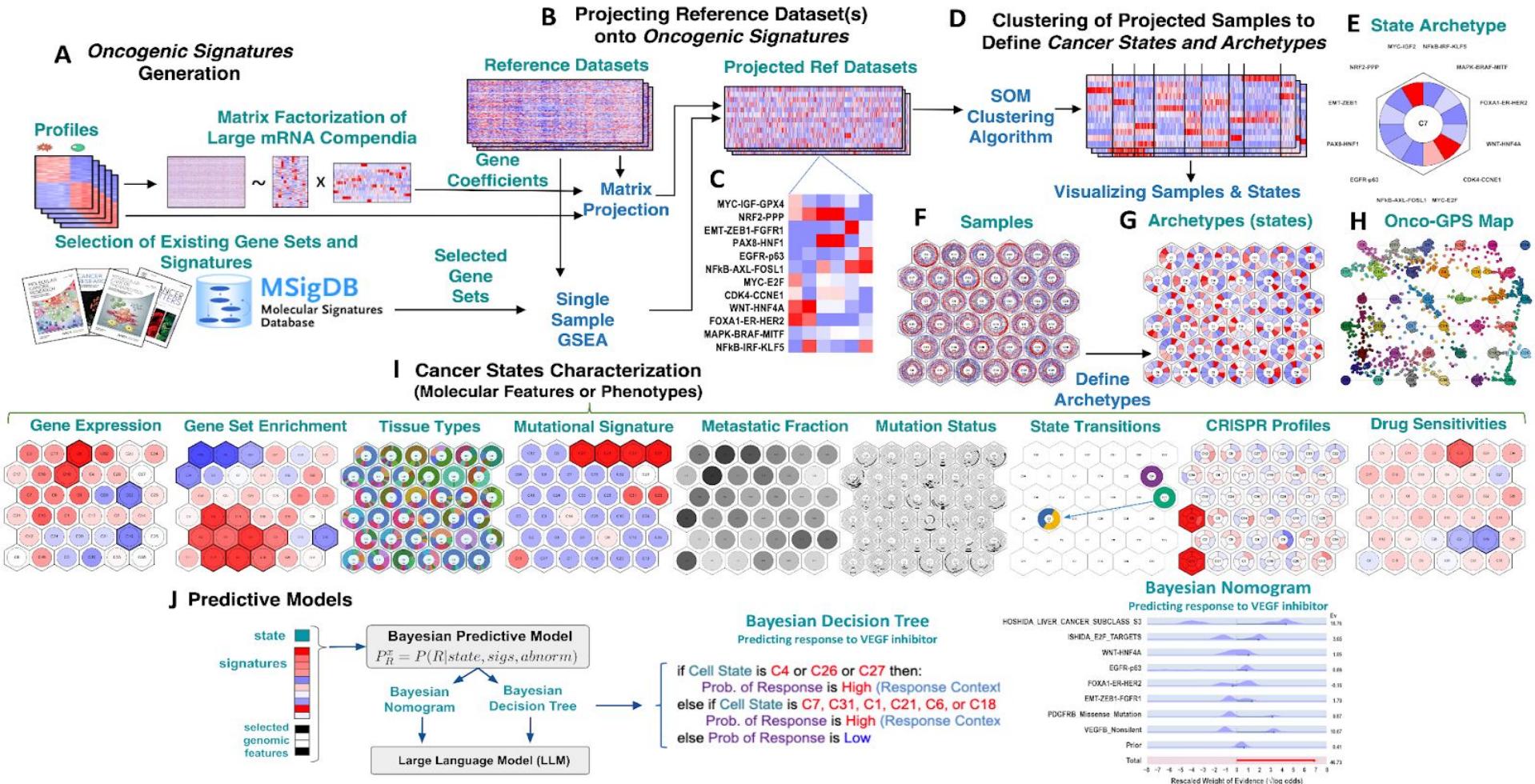
—Alan Perlis (1922-1990)

What is the CSA Framework?

- The *Cancer States and Archetypes* (CSA) analysis framework is an evolution and integration of many of our published genomic analysis methodologies, including Non-Negative Matrix Factorization,^{1–3} Onco-GPS,⁴ REVEALER,^{5,6} GSEA,^{7–9} SOM,¹⁰ Bayesian Nomograms,¹¹ SWNE,¹² DiSCoVER,¹³ and new algorithms for matrix factorization, visualization, and predictive modeling (manuscripts in preparation) we have developed in the last 10 years.
- It comprises over 20 integrated methods and algorithms that generate a unified *state representation model* that provides classification, comparative analysis, and predictive modeling methods for cancer samples, including primary tumors, cell lines, organoids, PDXs, etc.
- We have successfully applied different aspects of this framework to delineate and study specific instances of oncogenic states in a variety of studies across cancer types, including GIST,¹⁴ NSCLC,¹⁵ melanoma,¹⁶ medulloblastomas,^{17,18} prostate,^{19,20} uveal melanoma,²¹ HNSC,^{22–27} neuroblastoma,²⁸ MDS,²⁹ ovarian cancer,^{30–35} and therapy-resistant tumors.^{36–41}

Generating a Cellular States Model: Method Overview

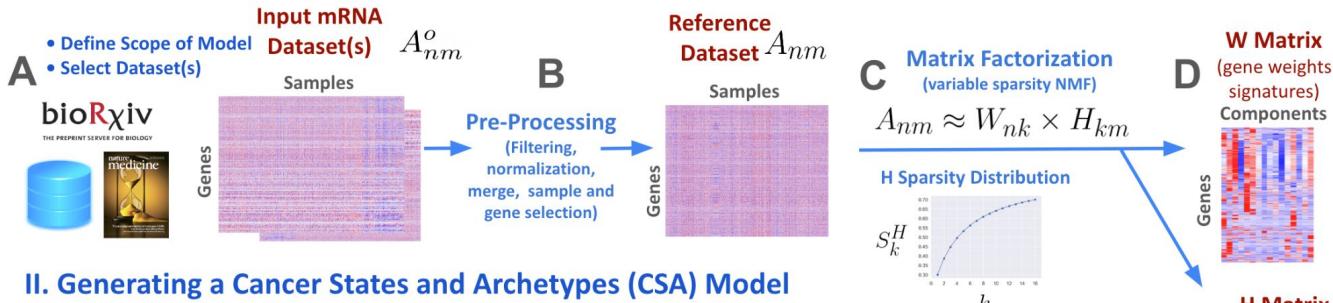
Cancer States and Archetypes Framework: Defining and Characterizing Cancer States



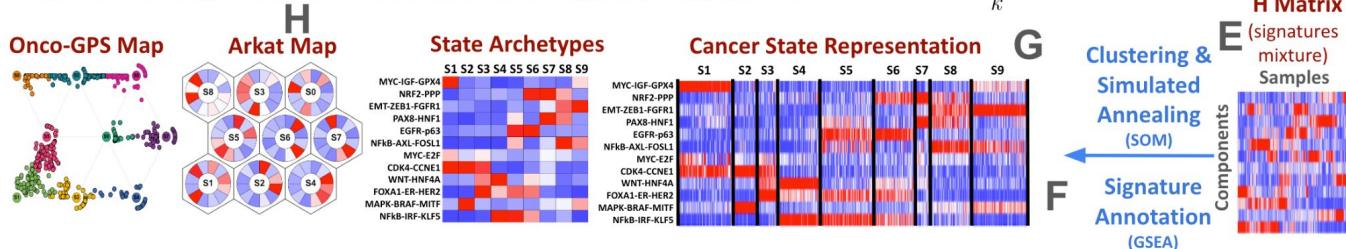
Generating a Cellular State Model

36-State Cellular State Model using Broad's CCLE 1000+ solid tumor cell lines

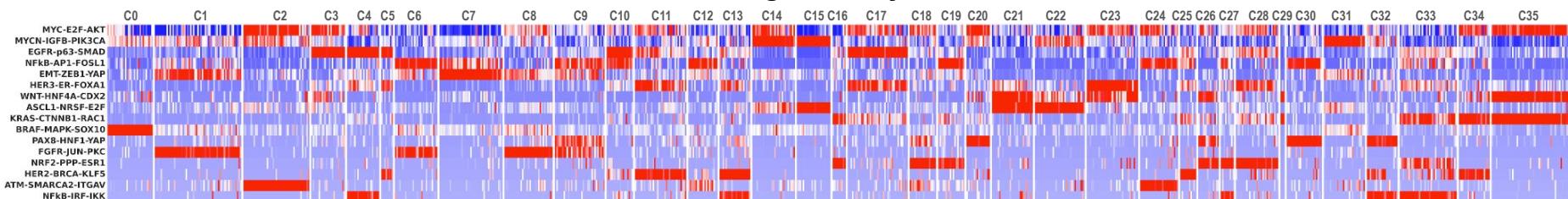
I. Matrix Factorization and Generation of Transcriptional Signatures



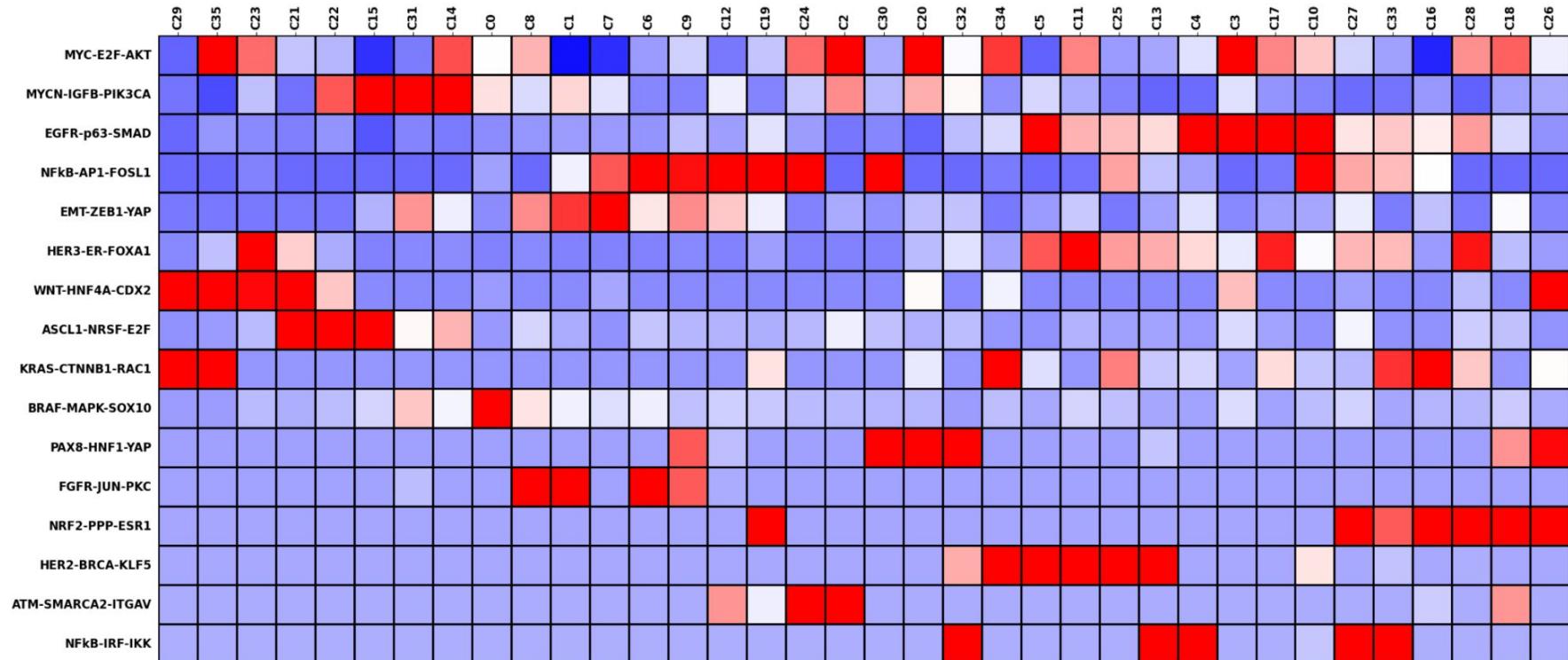
II. Generating a Cancer States and Archetypes (CSA) Model



Cell Line Profiles Organized by Cell State

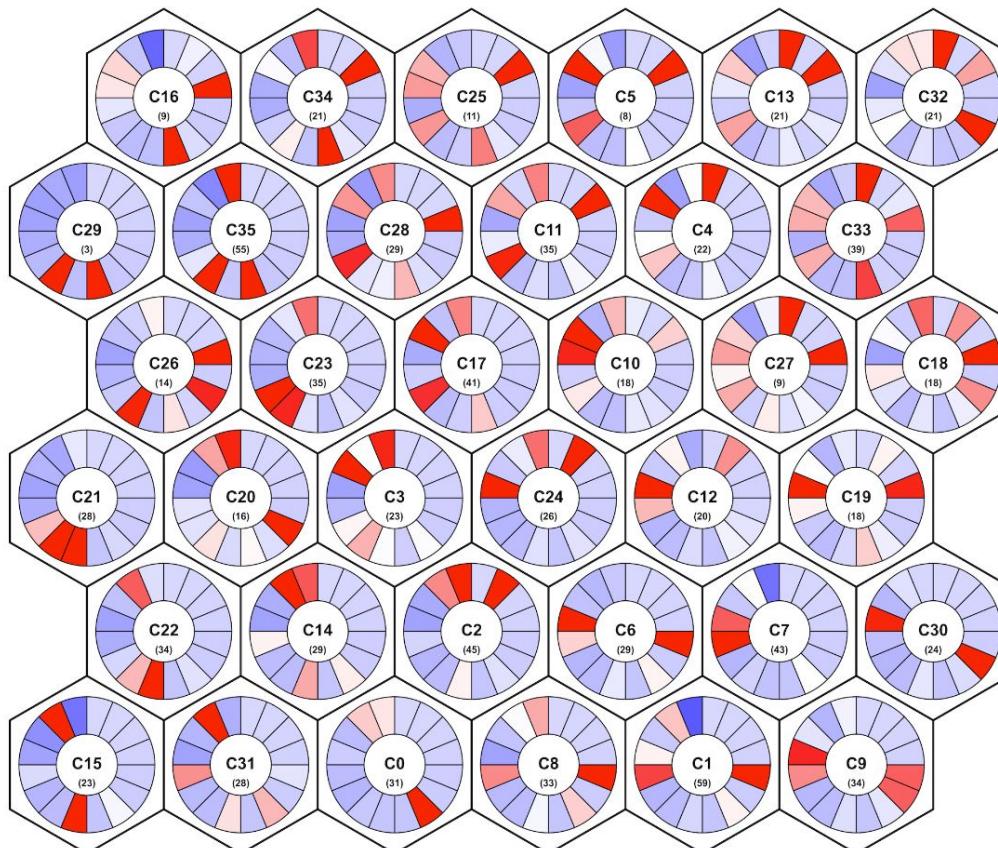


36-State Cellular States Model (linear heatmap version)



36-State Cellular States Model (hex layout version)

State Archetypes



Cellular State C33 Archetype

MYCN-IGFB-PIK3CA

ATM-SMARCA2-ITGAV

EGFR-p63-SMAD

HER2-BRCA-KLF5

NFkB-AP1-FOSL1

NRF2-PPP-ESR1

EMT-ZEB1-YAP

FGFR-JUN-PKC

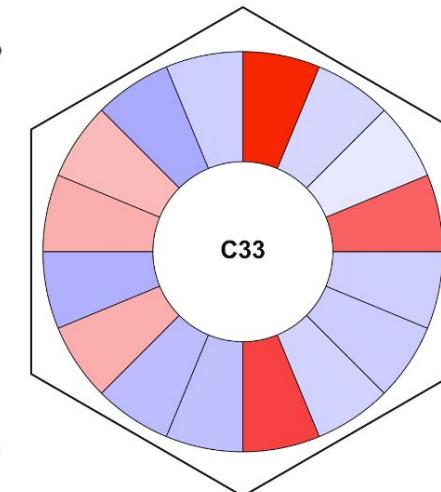
HER3-ER-FOXA1

PAX8-HNF1-YAP

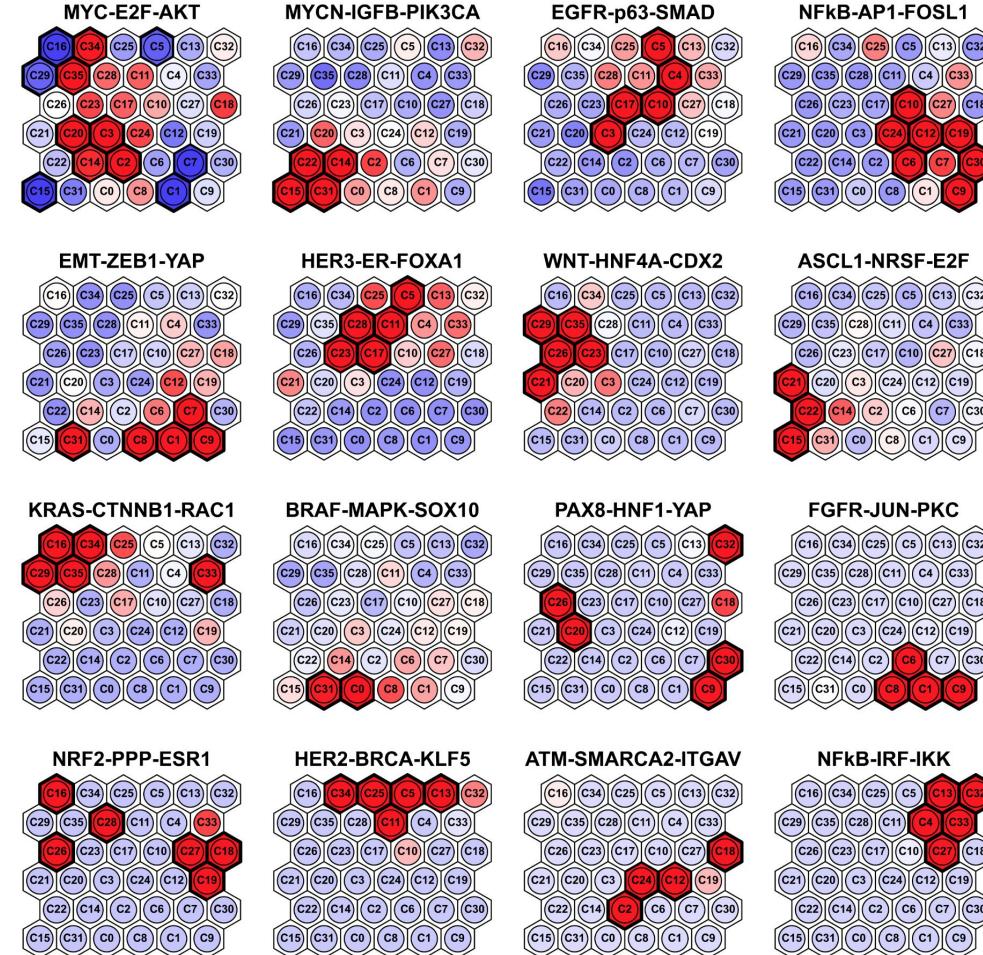
WNT-HNF4A-CDX2

BRAF-MAPK-SOX10

ASCL1-NRSF-E2F KRAS-CTNNB1-RAC1



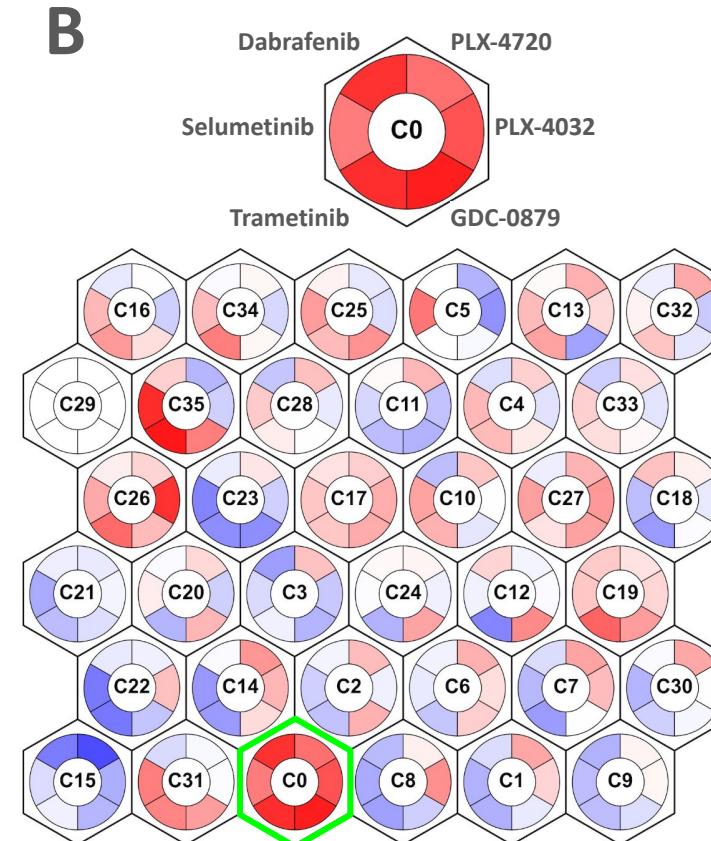
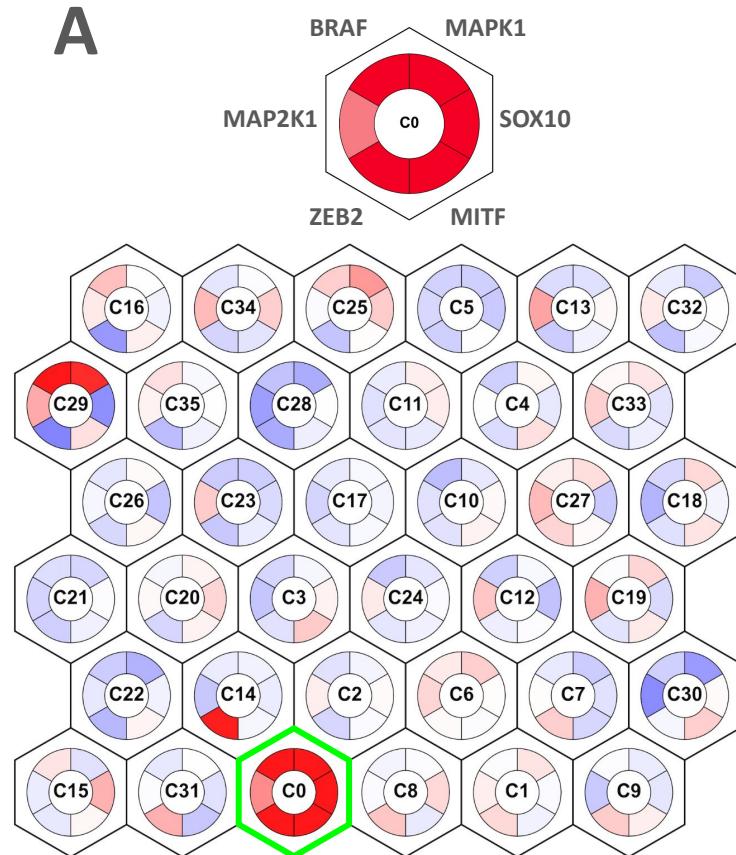
Transcriptional signatures patterns across 36 cellular states of the model



Some illustrative applications of the cellular states model

- I) Investigating state-based MAPK genetic and drug sensitivities
- II) Tracking cell state resistance transitions induced by a BRAF inhibitor
- III) Analysing vulnerabilities of selected cell states
- IV) A more in-depth analysis of a different set of selected cell states
- V) Multi-omic analysis across the 36 cell states

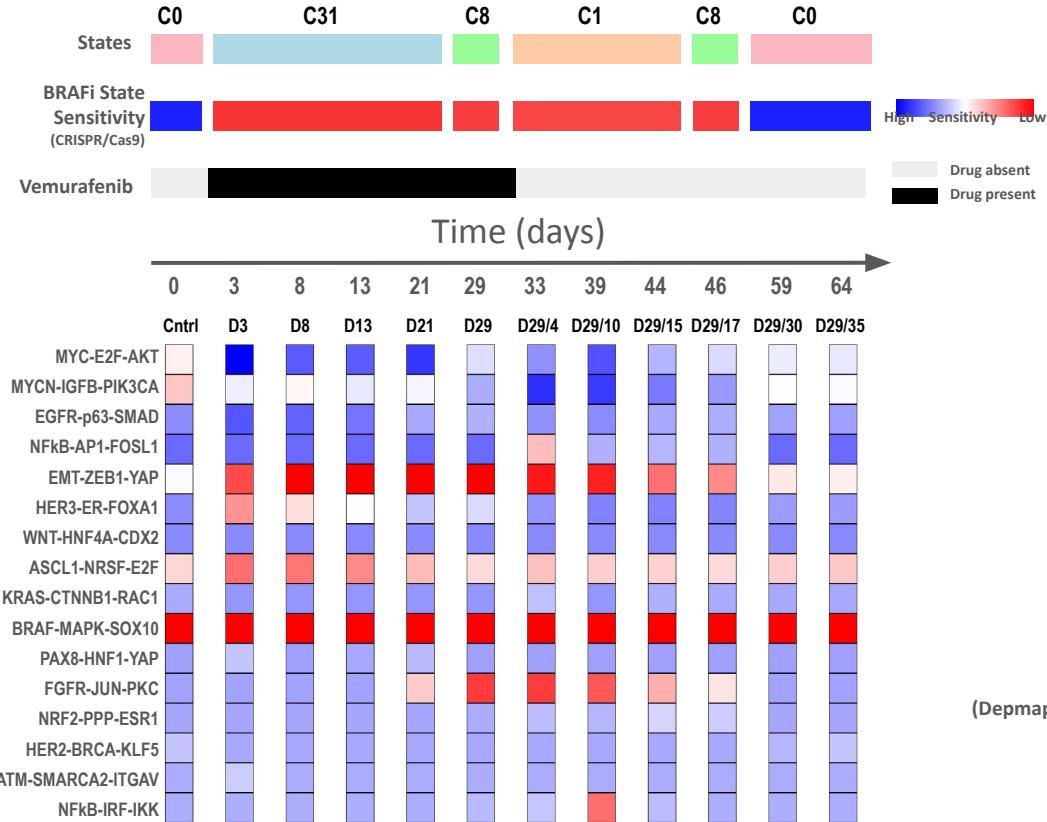
The CSA model of cellular states can be used to study the sensitivity to genetic or pharmacological perturbation across states by plotting the average sensitivity of each state's members on top of the hexagonal layout. In Fig xA we can see that the state most sensitive to CRISPR/Cas9 knockout of six genes in the MAPK pathway is state C0 (highlighted in green) presumably because of oncogenic pathway addiction and the react that this state is enriched with BRAF mutant samples. Fig xB shows the same type of analysis but this time for the sensitivity to six compounds targeting either *BRAF* or *MEK1*. In this case state C0 is also the most sensitive to those compounds but other states also appear to be partially sensitivity to a subset of the compounds.



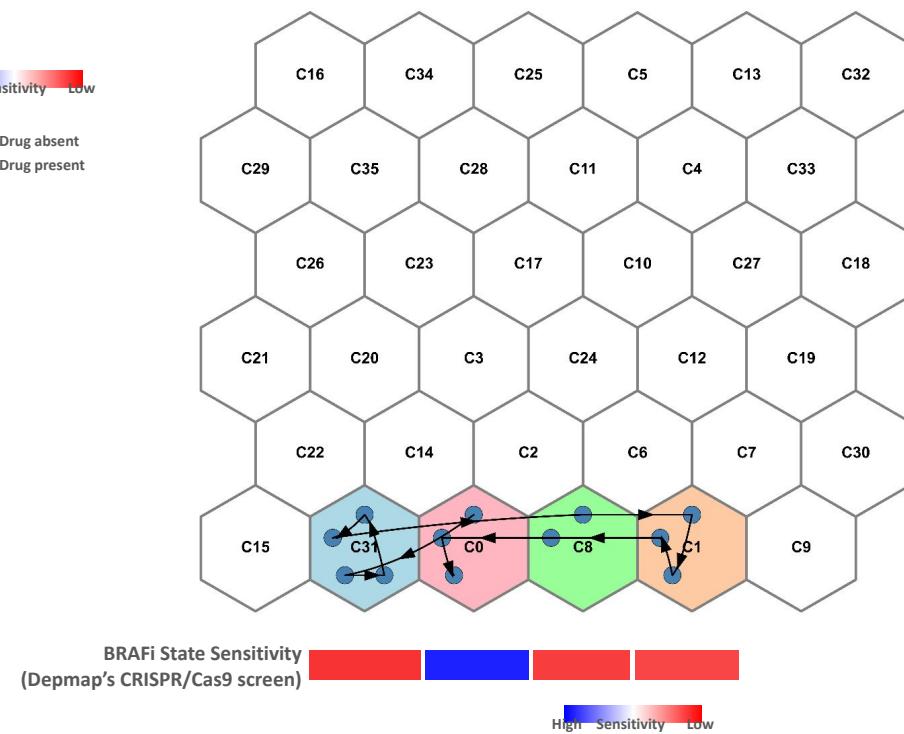
Tracking Cell State Transitions Induced by a Drug

BRAF-mutant melanoma cell line (M397) undergoing reversible cell-state changes induced by BRAF inhibition (vemurafenib, GEO Dataset: GSE134459, Yapeng et al.)

Transcriptional Signature Profiles

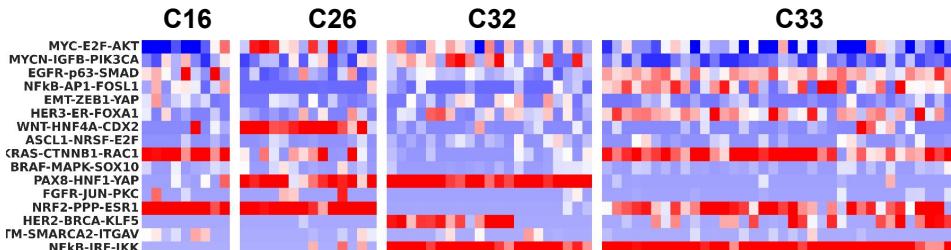


Cell State Transitions



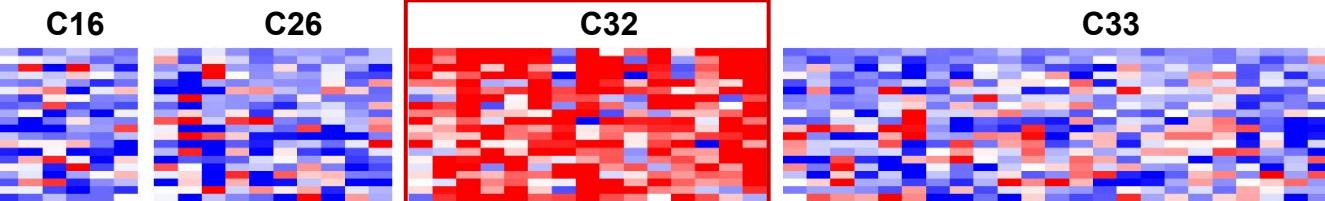
Finding Vulnerabilities to Genetic and Pharmacological Perturbations using Screening Data

CSA Cellular States Transcriptional Signatures

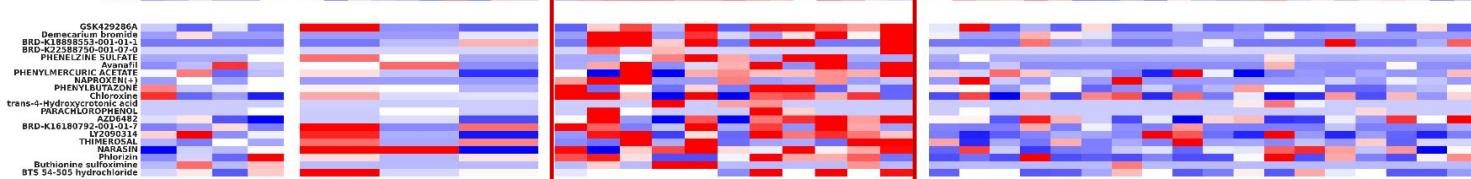


CRISPR/Cas9 Cancer Dependencies Map (DepMap)

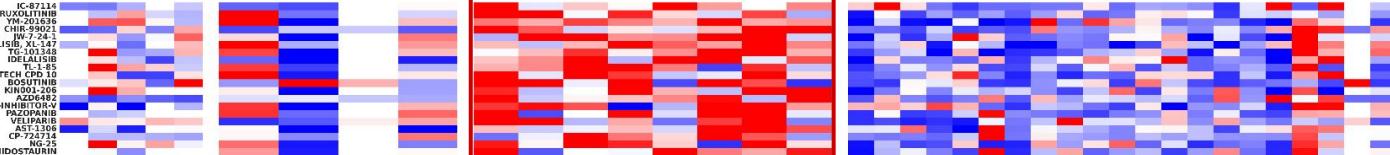
Ferroptosis control



PRISM drug repurposing resource

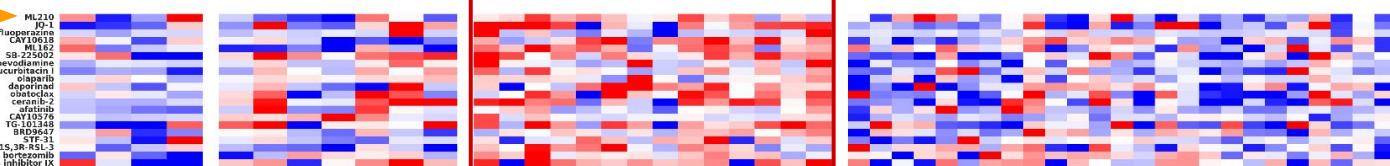


GDSCV Genomics of Drug Sensitivity in Cancer



CTRP Cancer Therapeutics Response Portal

Ferroptosis inducer



A more in-depth analysis of 4 GI-cancers relevant Cellular States

Transcriptional
Signatures

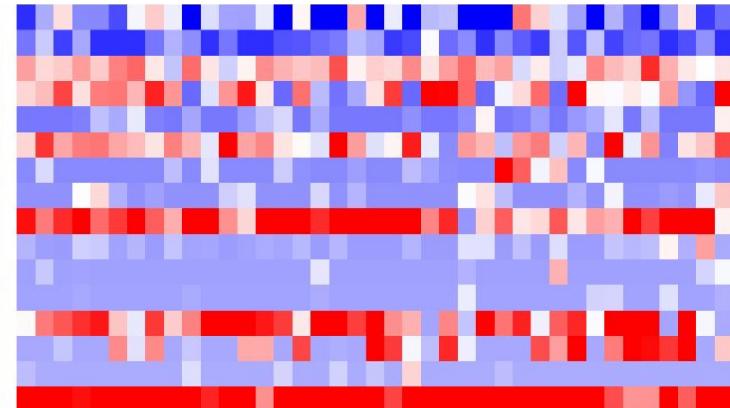
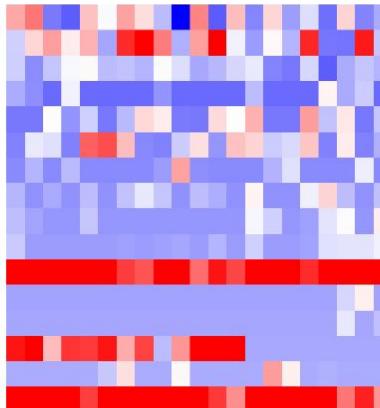
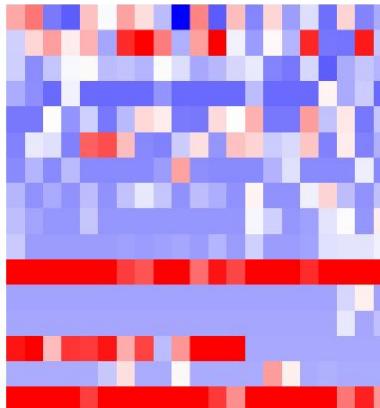
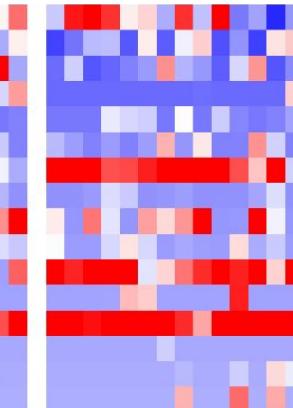
C16

C26

C32

C33

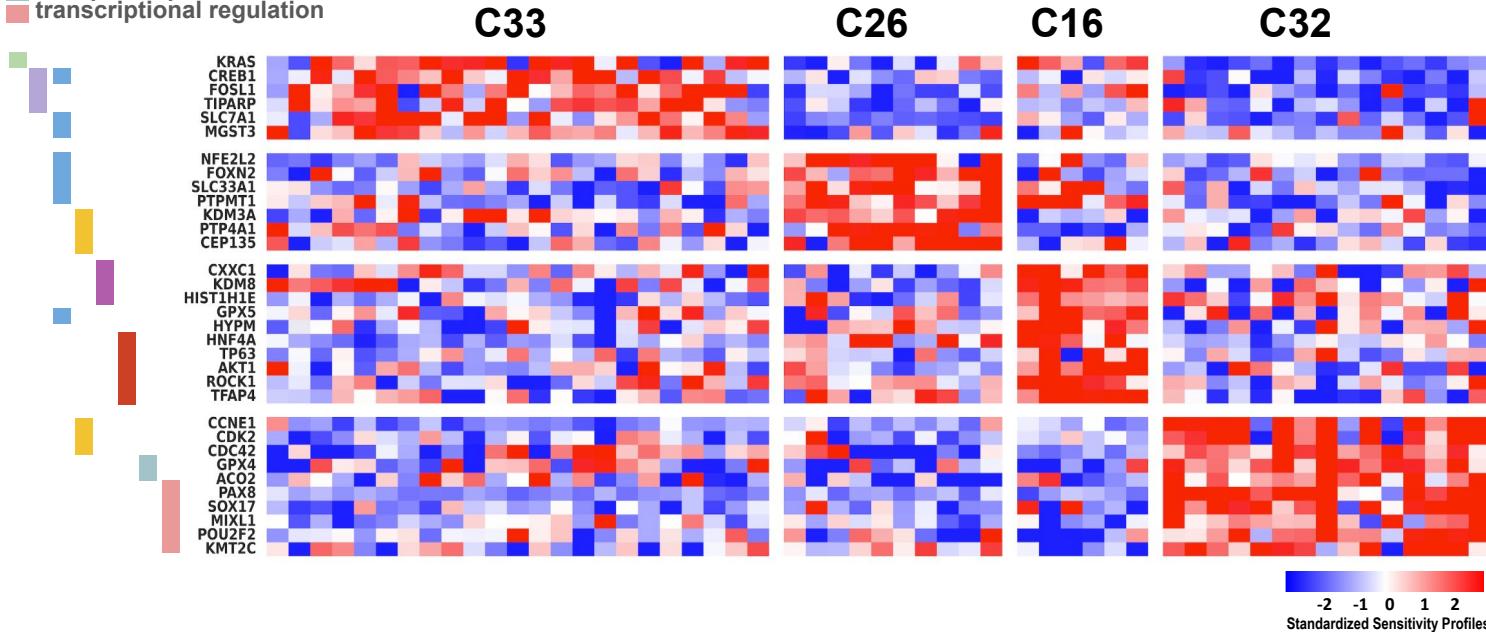
MYC-E2F-AKT
MYCN-IGFB-PIK3CA
EGFR-p63-SMAD
NFkB-AP1-FOSL1
EMT-ZEB1-YAP
HER3-ER-FOXA1
WNT-HNF4A-CDX2
ASCL1-NRSF-E2F
KRAS-CTNNB1-RAC1
BRAF-MAPK-SOX10
PAX8-HNF1-YAP
FGFR-JUN-PKC
NRF2-PPP-ESR1
HER2-BRCA-KLF5
ATM-SMARCA2-ITGAV
NFkB-IRF-IKK



-2 -1 0 1 2
Standardized Transcriptional
Signature Profile

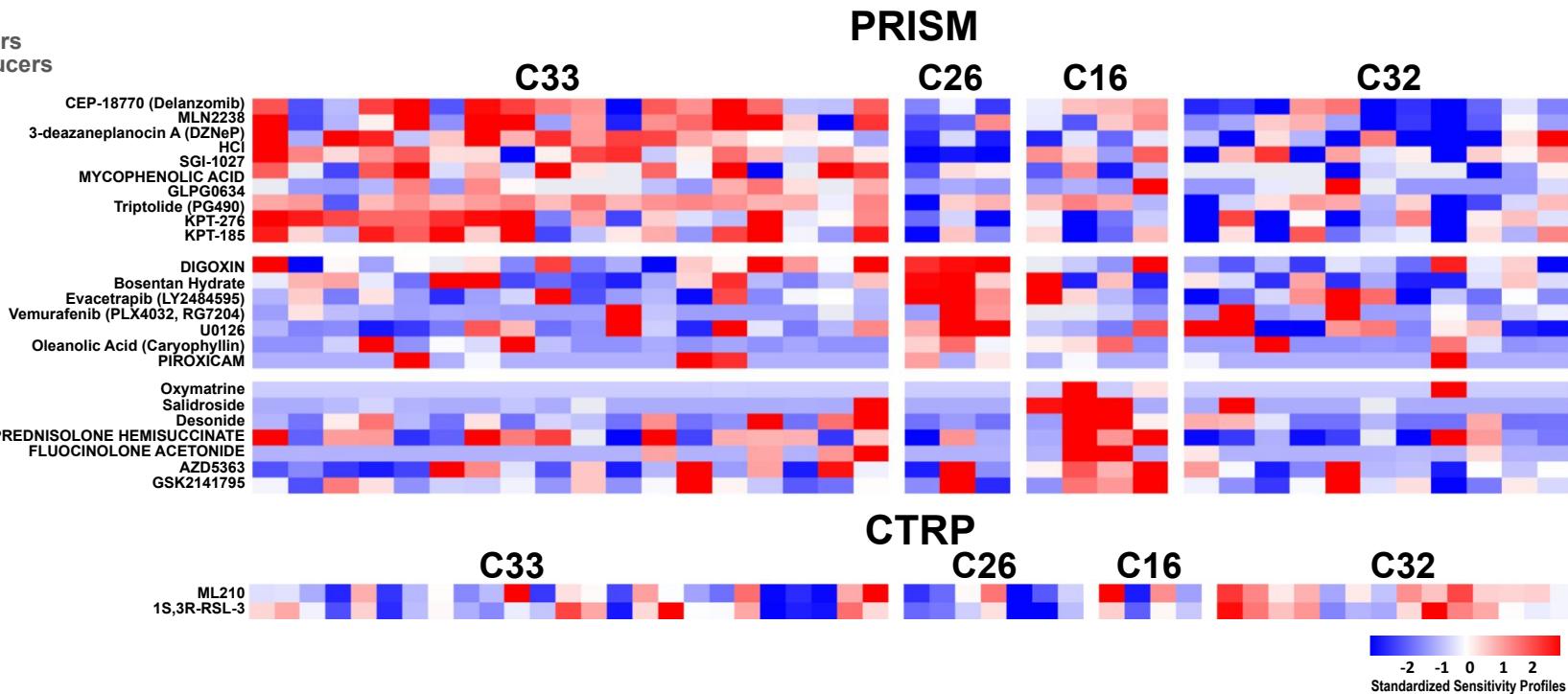
Selected state-specific genetic lethaliess (CRISPR) for Cellular States C33, C26, C16 and C32

KRAS
NFkB
oxidative stress response
cell cycle regulation
chromatin remodeling
epithelial differentiation
ferroptosis prevention
transcriptional regulation



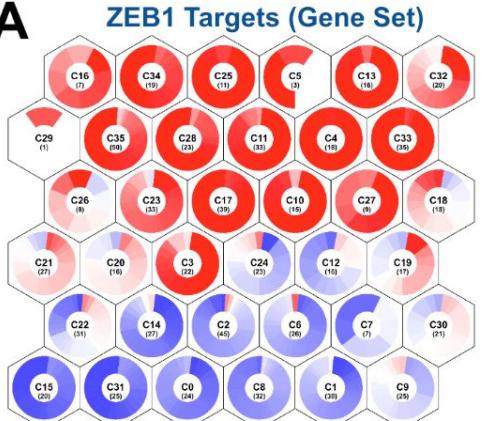
Selected state-specific drug sensitivities for Cellular States C33, C26, C16 and C32

- █ proteasome inhibitors
- █ epigenetic modulators
- █ anti-inflammatory
- █ XPO1 inhibitors
- █ ion channel modulators
- █ MAPK inhibitors
- █ antioxidants
- █ corticosteroids
- █ AKT1-3 inhibitors
- █ ferroptosis inducers

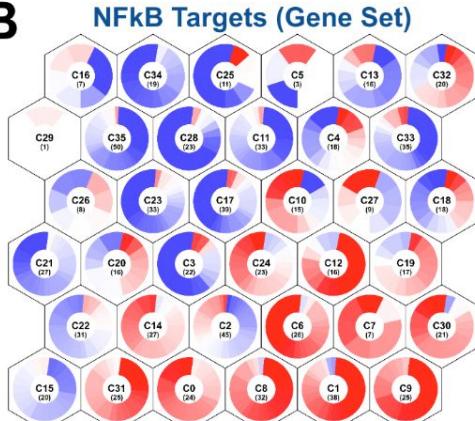


Multi-omic analysis across the 36 cell states

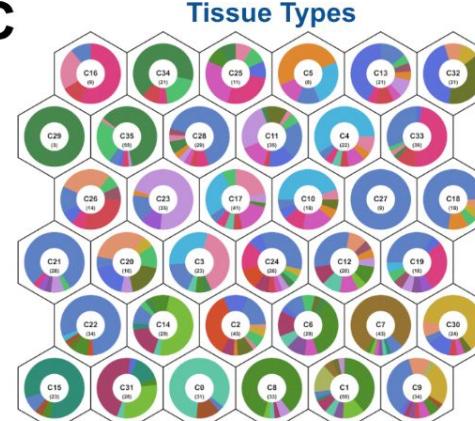
A



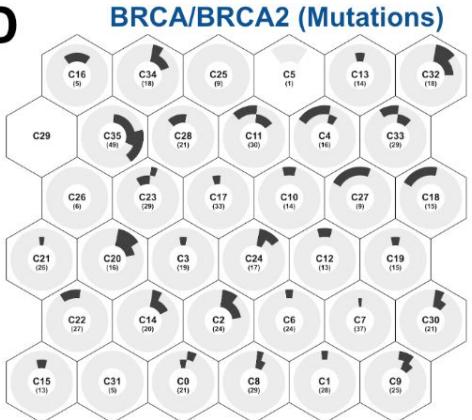
B



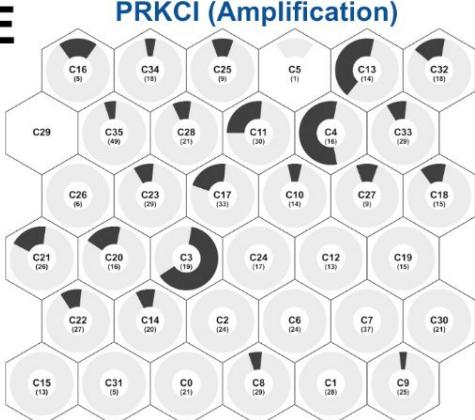
C



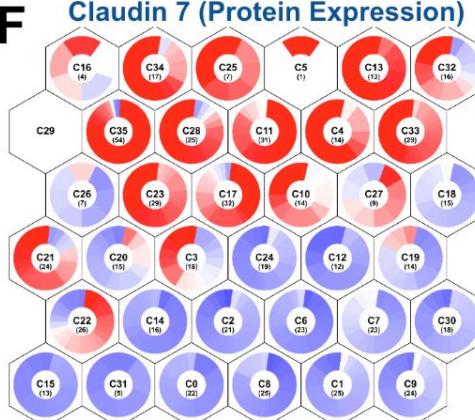
D



E



F



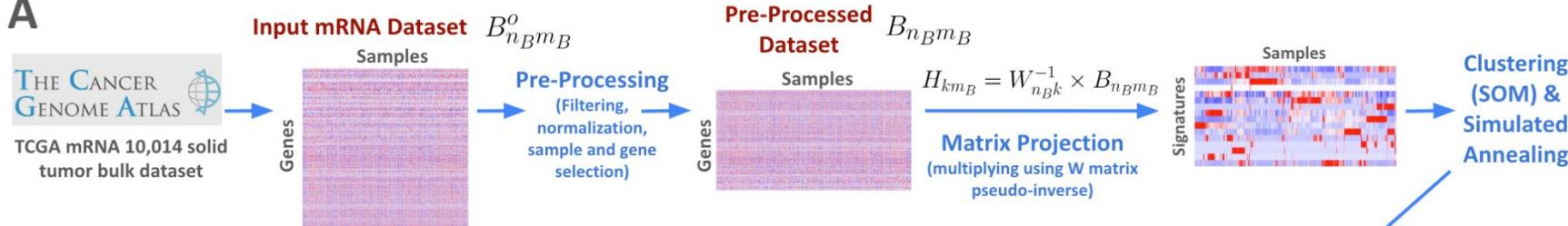
- bile_duct_cancer
- bladder_cancer
- bone_cancer
- brain_cancer
- breast_cancer
- cervical_cancer
- colon/colorectal_cancer
- endometrial/uterine_cancer
- esophageal_cancer
- fibroblast
- gastric_cancer
- head_and_neck_cancer
- kidney_cancer
- liposarcoma
- liver_cancer
- lung_cancer
- neuroblastoma
- ovarian_cancer
- pancreatic_cancer
- prostate_cancer
- rhabdoid
- sarcoma
- skin_cancer
- thyroid_cancer

Generating a Tumor States Model: Method Overview

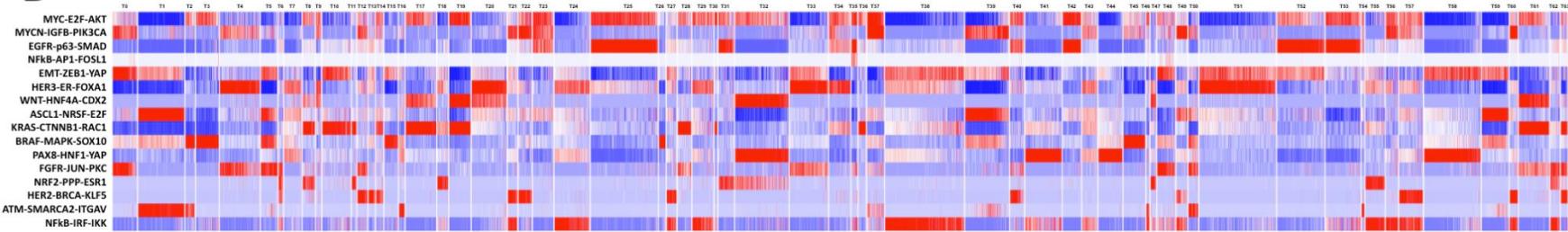
Generating a Tumor States Model

Projecting Dataset onto the Cellular State Model Signatures

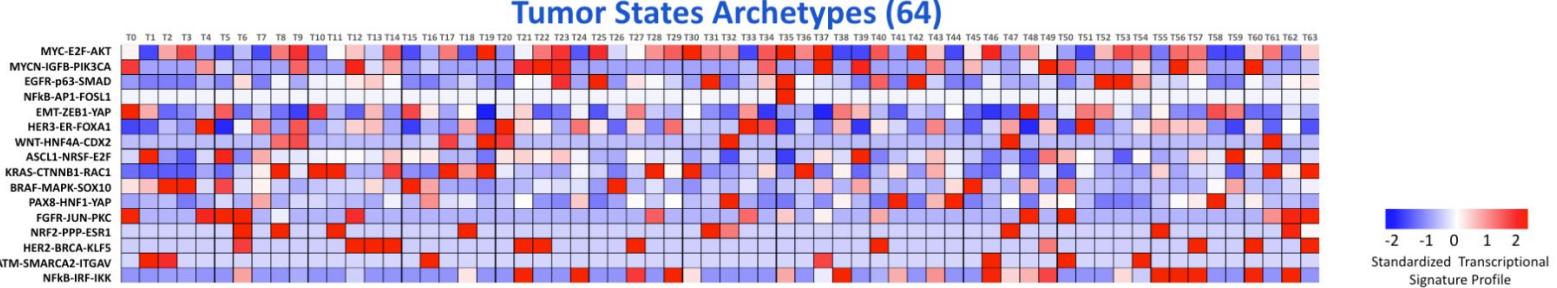
A



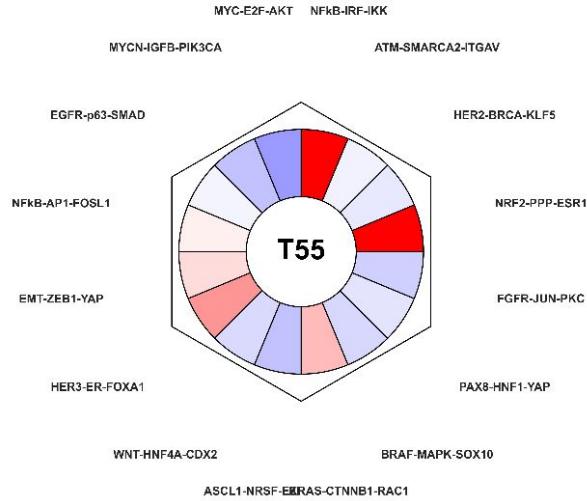
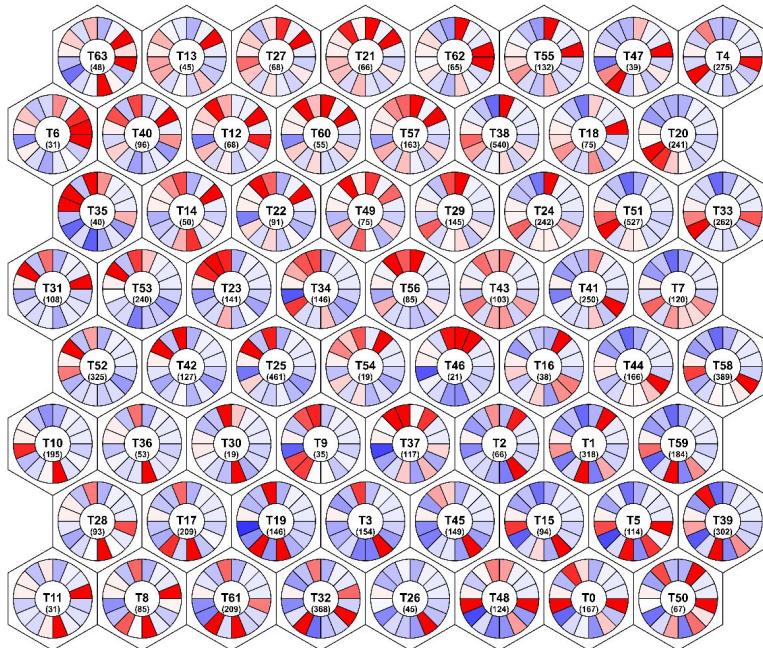
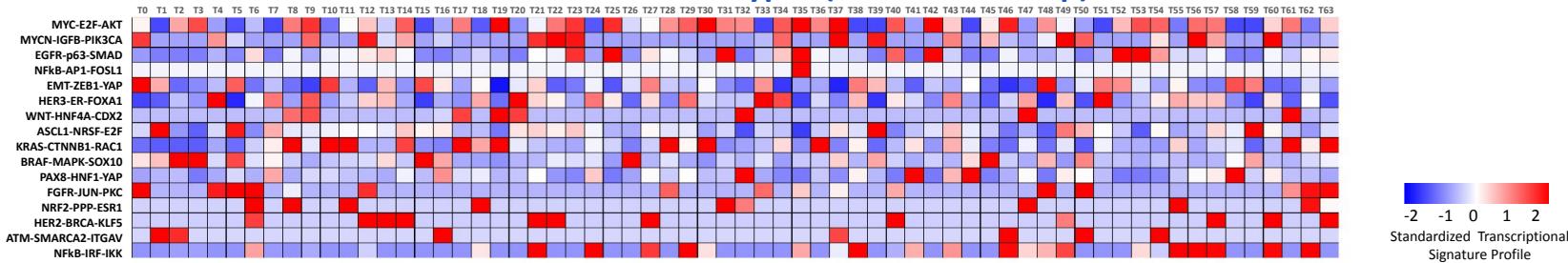
B



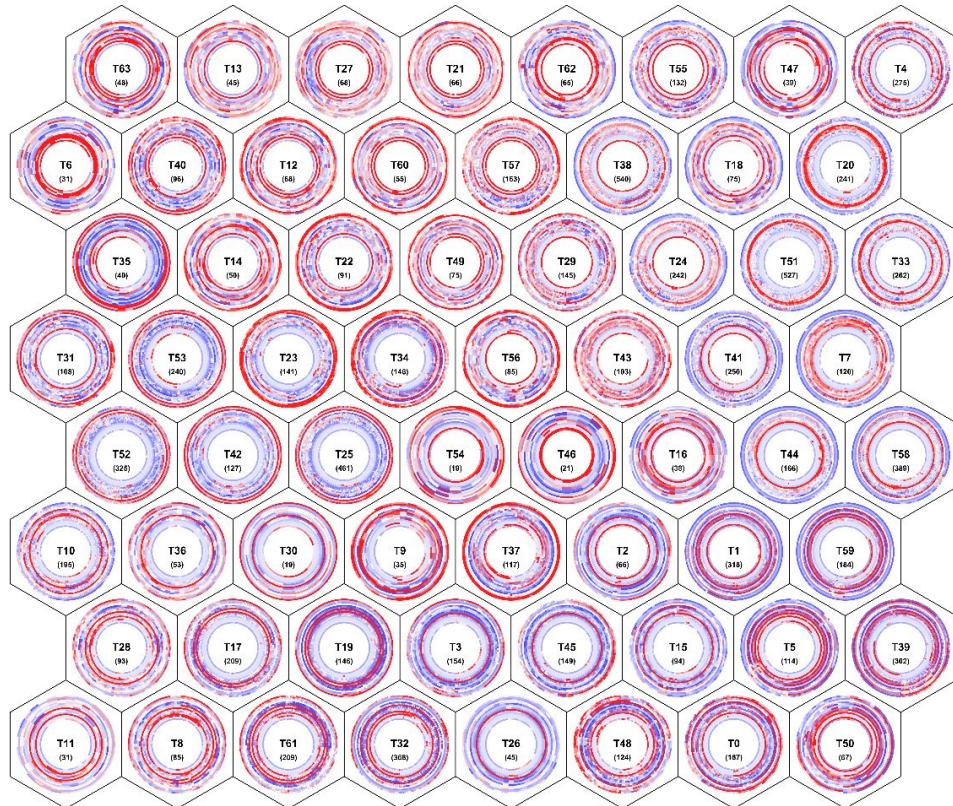
C



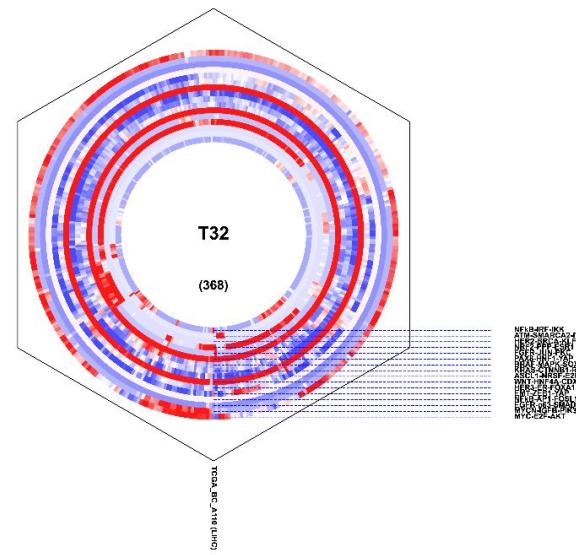
Tumor States Archetypes (Linear Heat Map)



Tumor Archetypes (TCGA PanCan)



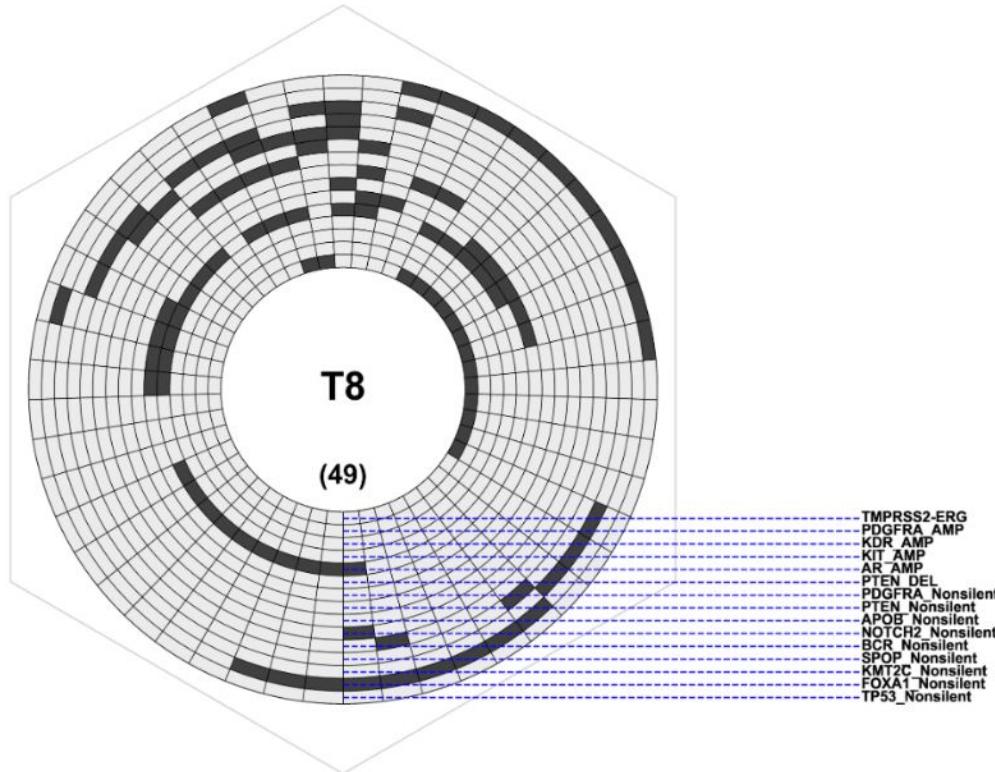
T32

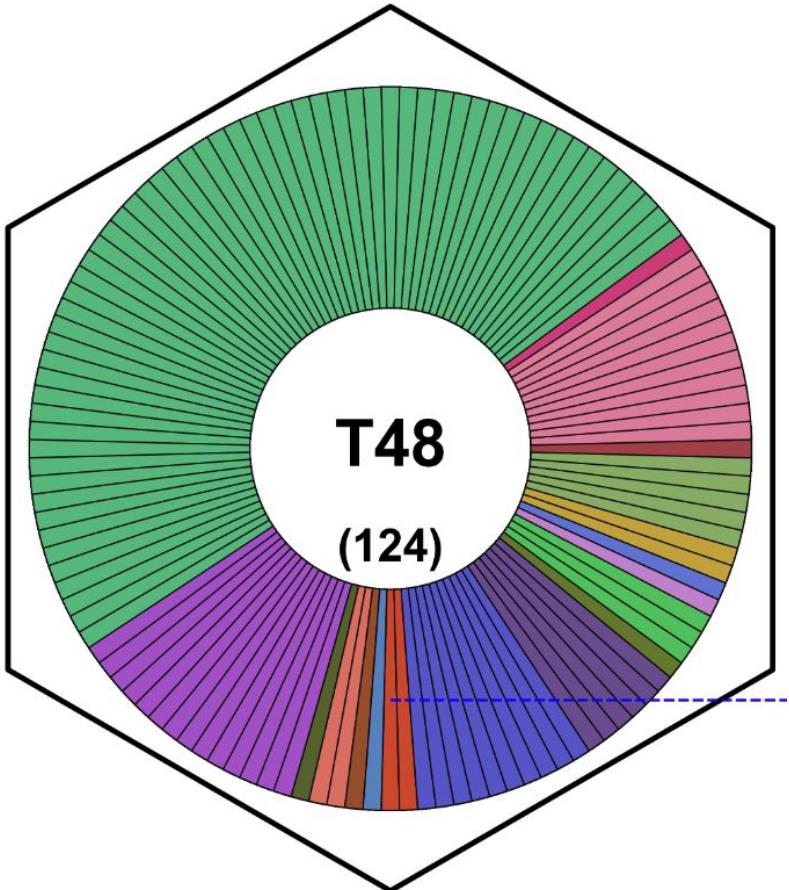


Some illustrative applications of the tumor states model

- I) Analyzing mutational and CNV landscapes of specific states
- II) Analyzing cancer type distributions
- III) Defining functional cancer taxonomies (e.g. breast cancer)
- IV) Projecting an external dataset onto the tumor states model
- V) Studying resistance transitions to cancer therapy in patients

Mutational Landscape of Tumor State T8

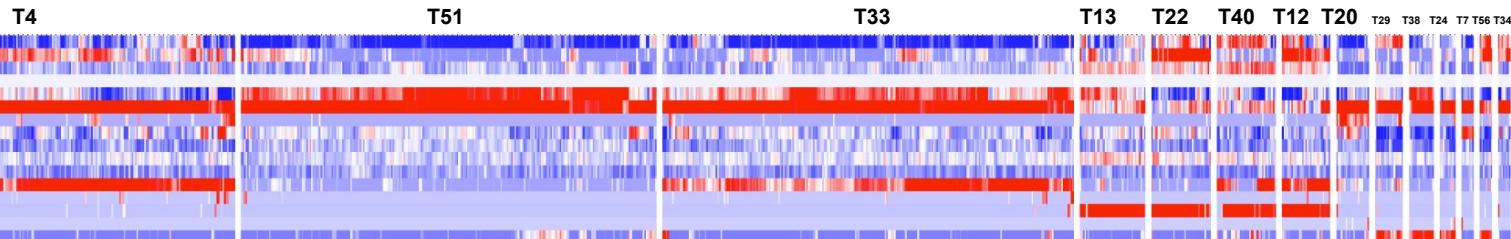




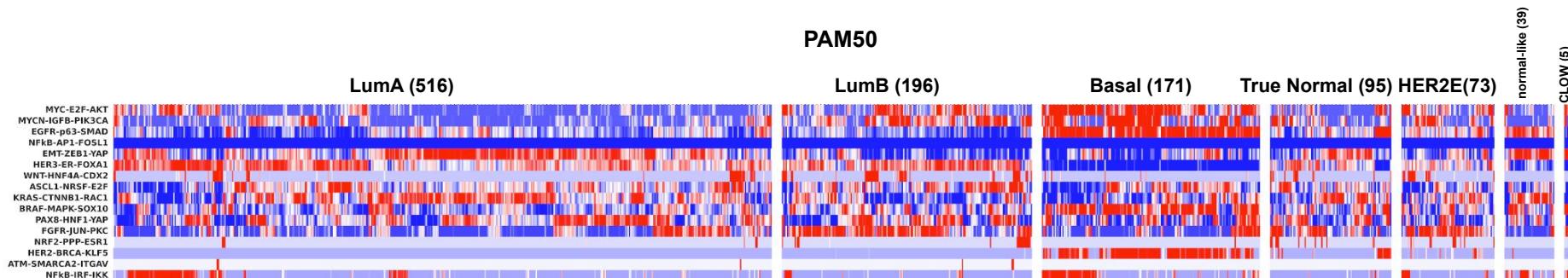
- ACC Adrenocortical carcinoma
- BLCA Bladder Urothelial Carcinoma
- BRCA Breast invasive carcinoma
- CESC Cervical endocervical adenocarcinoma
- CHOL Cholangiocarcinoma
- COAD Colon adenocarcinoma
- ESCA Esophageal carcinoma
- GBM Glioblastoma multiforme
- HNSC Head and Neck squamous cell carcinoma
- KICH Kidney Chromophobe
- KIRC Kidney renal clear cell carcinoma
- KIRP Kidney renal papillary cell carcinoma
- LGG Brain Lower Grade Glioma
- LIHC Liver hepatocellular carcinoma
- LUAD Lung adenocarcinoma
- LUSC Lung squamous cell carcinoma
- MESO Mesothelioma
- OV Ovarian serous cystadenocarcinoma
- PAAD Pancreatic adenocarcinoma
- PCPG Pheochromocytoma and Paraganglioma
- PRAD Prostate adenocarcinoma
- READ Rectum adenocarcinoma
- SARC Sarcoma
- SKCM Skin Cutaneous Melanoma
- STAD Stomach adenocarcinoma
- TGCT Testicular Germ Cell Tumors
- THCA Thyroid carcinoma
- THYM Thymoma
- UCEC Uterine Corpus Endometrial Carcinoma
- UCS Uterine Carcinosarcoma
- UVM Uveal Melanoma

The Model can Provide a Better (Functional) Taxonomy for Breast Cancer

CSA Tumor States



PAM50



Histological Subtypes

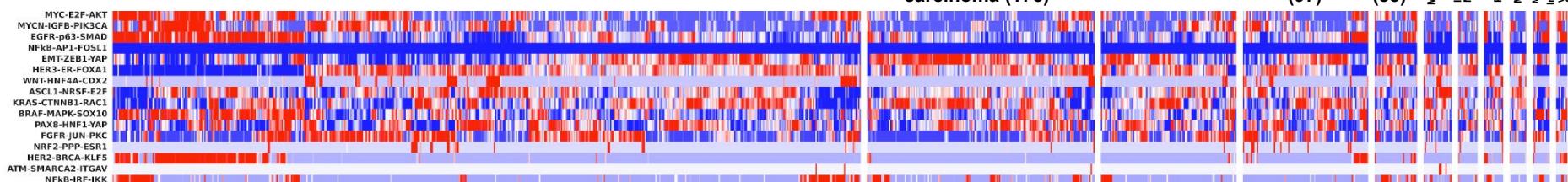
Invasive ductal carcinoma (579)

Invasive lobular carcinoma (176)

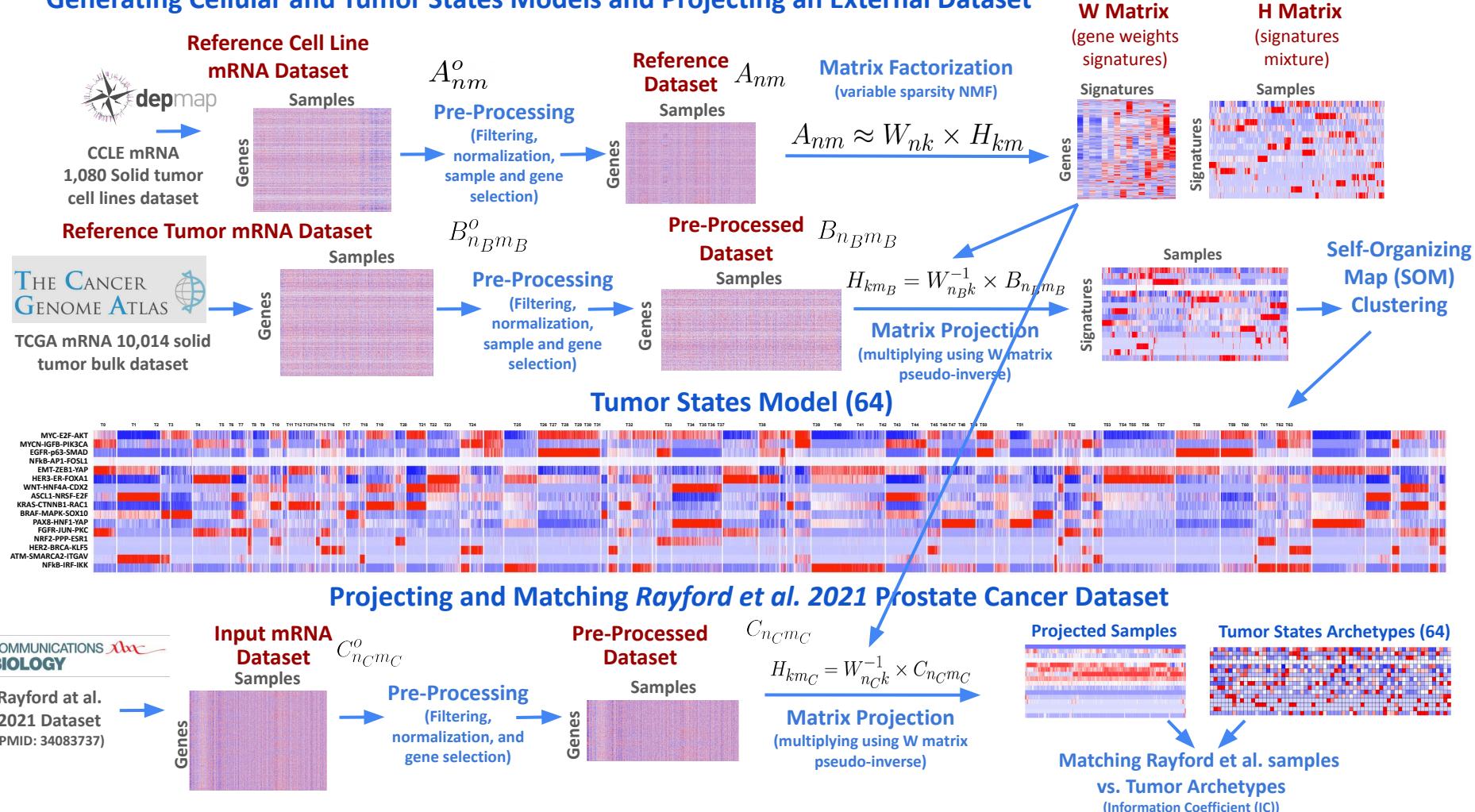
Mixed (105)

True Normal (97)
No Image (33)

Mucinous carcinoma (22)
Invasive Micropapillary carcinoma (15)
Papillary neoplasm (15)
Metaplastic carcinoma (13)
Mixed (13)
Metaplastic carcinoma with medullary features (9)

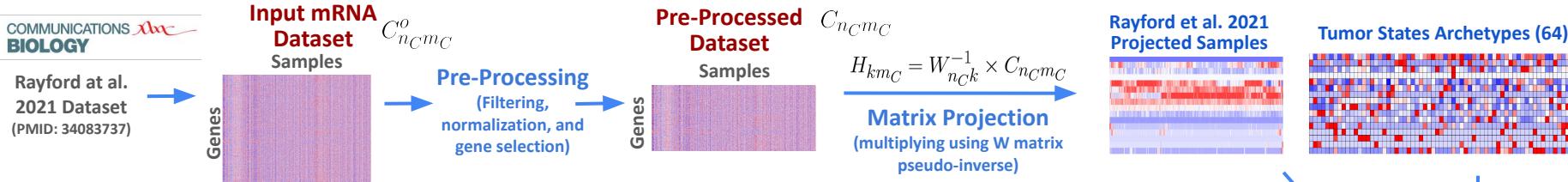


Generating Cellular and Tumor States Models and Projecting an External Dataset



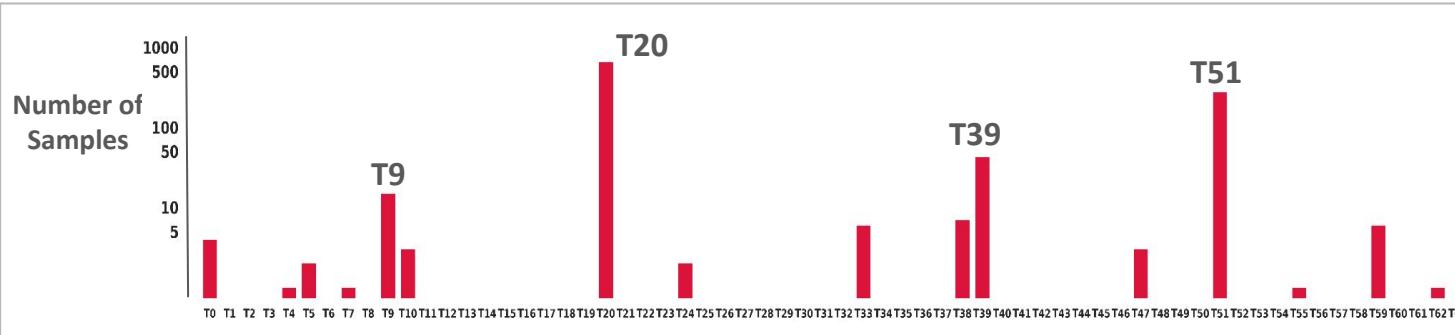
Projecting a New Dataset onto the Tumor States Model

A



C

Distribution of Projected and Matched Rayford et al. 2021 Samples Across Tumor States



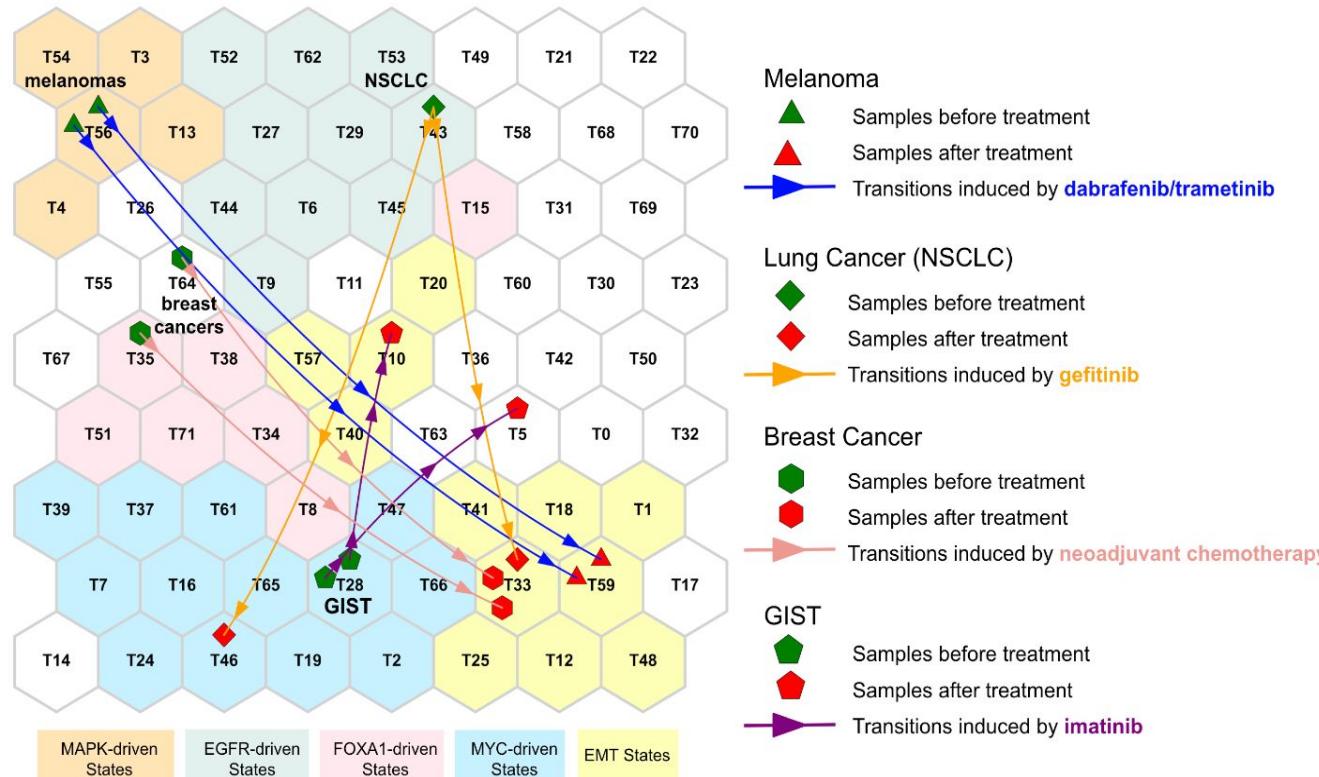
B

Matching Against Tumor States Model Archetypes (Information Coefficient (IC))

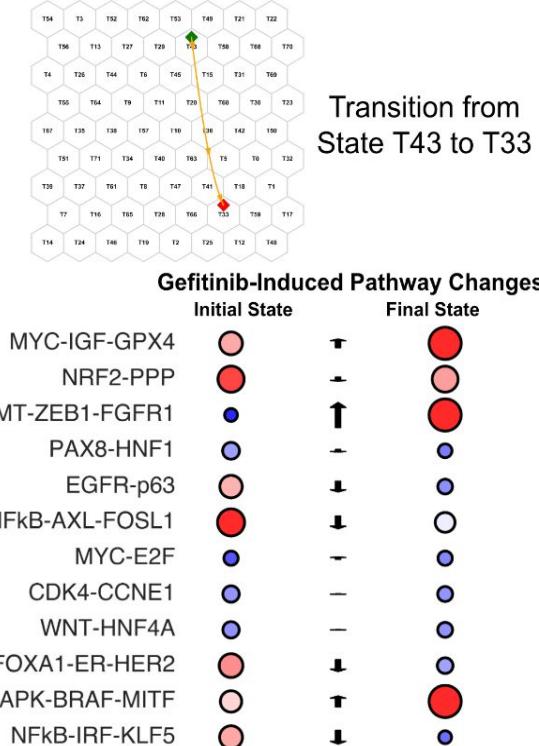
90% of samples match one of the tumor state model archetypes

Using a Tumor State Model to Study Resistance Transitions in Cancer Patients

A Examples of Tumor States Before/After Treatment and Corresponding Drug-Induced Transitions



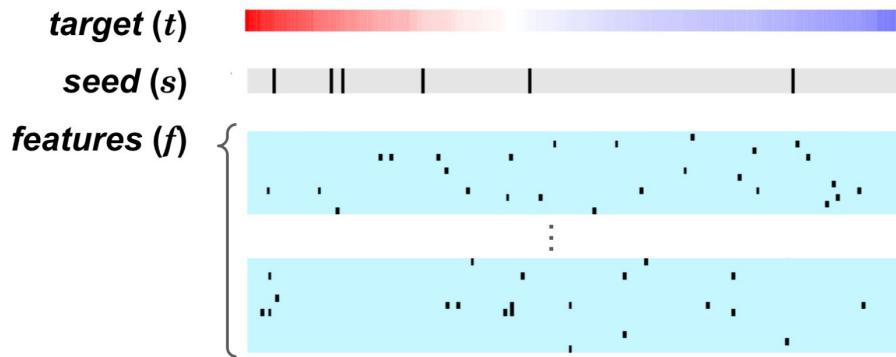
B Detailed Analysis of a Gefitinib-Resistance Transition in NSCLC Tumors



The REVEALER Algorithm: Overview and Applications

REVEALER identifies groups of **complementary genomic alterations that have significant association with a biological phenotype of interest**, such as pathway activation, gene dependency, or drug response

There are 3 inputs to REVEALER



The Problem: Given the target and the seed and a set of features, find those features that match the target but are complementary to the seed

nature biotechnology

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

nature > nature biotechnology > articles > article

Article | Published: 18 April 2016

Characterizing genomic alterations in cancer by complementary functional associations

Jong Wook Kim, Olga B Botvinnik, Omar Abudayyeh, Chet Birger, Joseph Rosenbluh, Yashaswi Shrestha, Mohamed F Abazeed, Peter S Hammerman, Daniel DiCarla, David J Konieczkowski, Cory M Johannessen, Arthur Liberzon, Amir Reza Alizad-Rahvar, Gabriela Alexe, Andrew Aguirre, Mahmoud Ghandi, Heidi Greulich, Francisca Vazquez, Barbara A Weir, Eilezer M Van Allen, Aviad Tsherniak, Diane D Shao, Travis J Zack, Michael Noble, Gad Getz, Rameen Beroukhim, Levi A Garraway, Masoud Ardakanian, Chiara Romualdi, Gabriele Sales, David A Barbie, Jesse S Boehm, William C Hahn, Jill P Mesirov & Pablo Tamayo ↗ — Show fewer authors

Nature Biotechnology 34, 539–546 (2016) | Cite this article

> bioRxiv [Preprint]. 2023 Nov 16:2023.11.14.567106. doi: 10.1101/2023.11.14.567106.

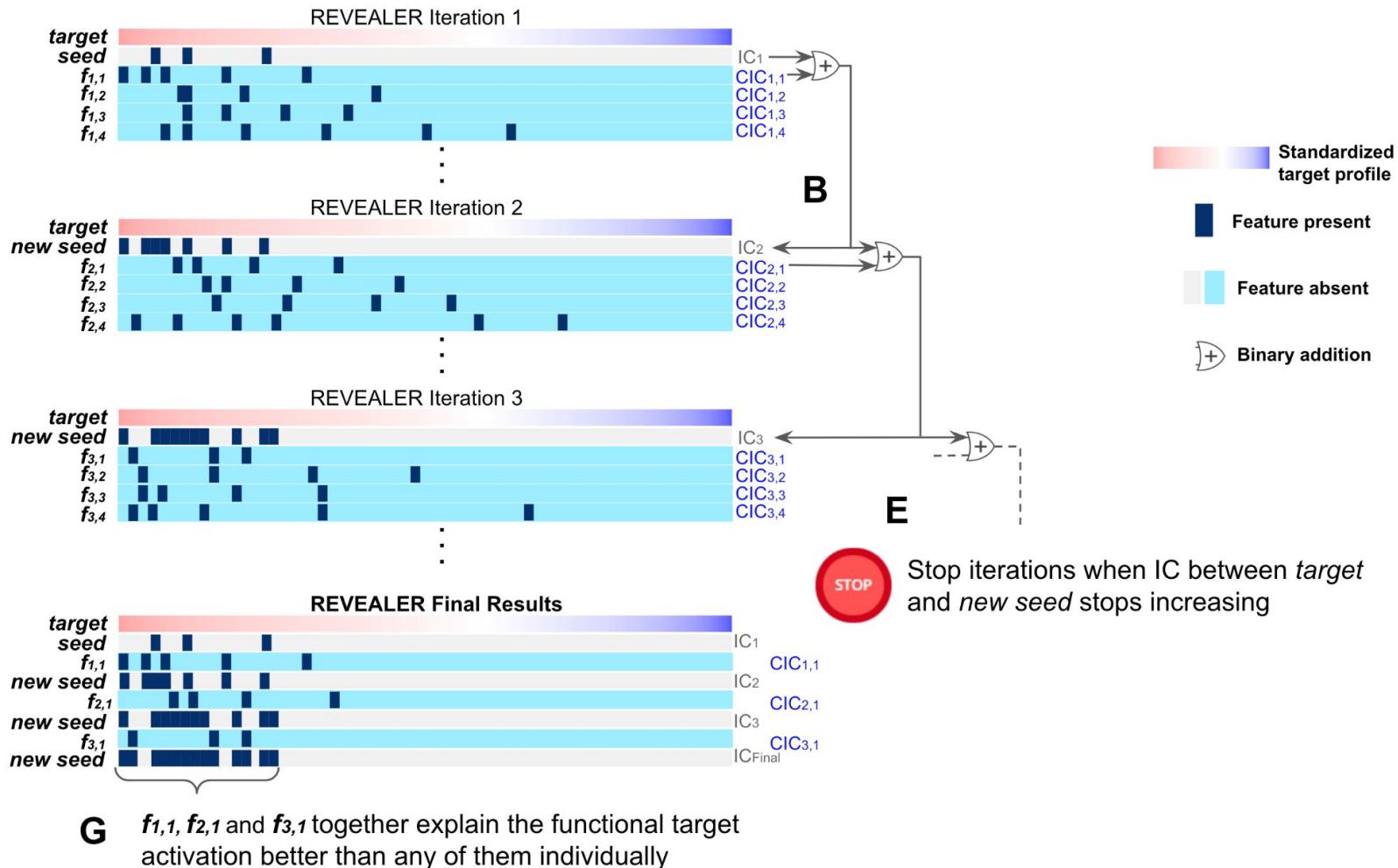
Deciphering the Functional Roles of Individual Cancer Alleles Across Comprehensive Cancer Genomic Studies

Jiayan Yoshii Ma ^{1,2}, Stephanie Ting ^{1,2}, Bartholomew Tam ^{1,2}, Timothy Pham ^{1,2}, Michael Reich ³, Jill Mesirov ^{2,4}, Pablo Tamayo ^{1,2,4}, William Kim ^{1,2,4}

Affiliations + expand

PMID: 38014215 PMCID: PMC10680728 DOI: 10.1101/2023.11.14.567106

REVEALER Algorithm



REVEALER's Information-Theoretic Association Measures

2-way comparison $F(t, s/f)$

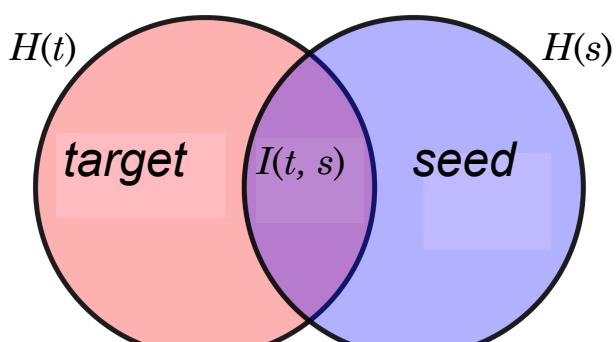
Mutual Information

$$I(t, s) = \int \int P(t, s) \log \frac{P(t, s)}{P(t)P(s)} dt ds$$

$$I(t, s) = \sum_{i=1}^{n_t} \sum_{s=0}^1 P(t_i, s) \log \frac{P(t_i, s)}{P(t_i)P(s)}$$

Information Coefficient (IC)

$$IC(t, s) = \text{sign}(\rho) \sqrt{1 - \exp(-2I(t, s))}$$



3-way comparison $G(t, f, s)$

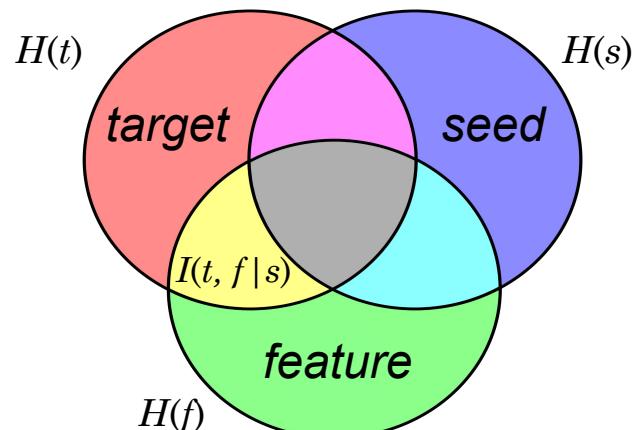
Conditional Mutual Information

$$I(t, f|s) = \int \int \int P(t, f, s) \log \frac{P(t, f, s)P(s)}{P(t, s)P(f, s)} dt df ds$$

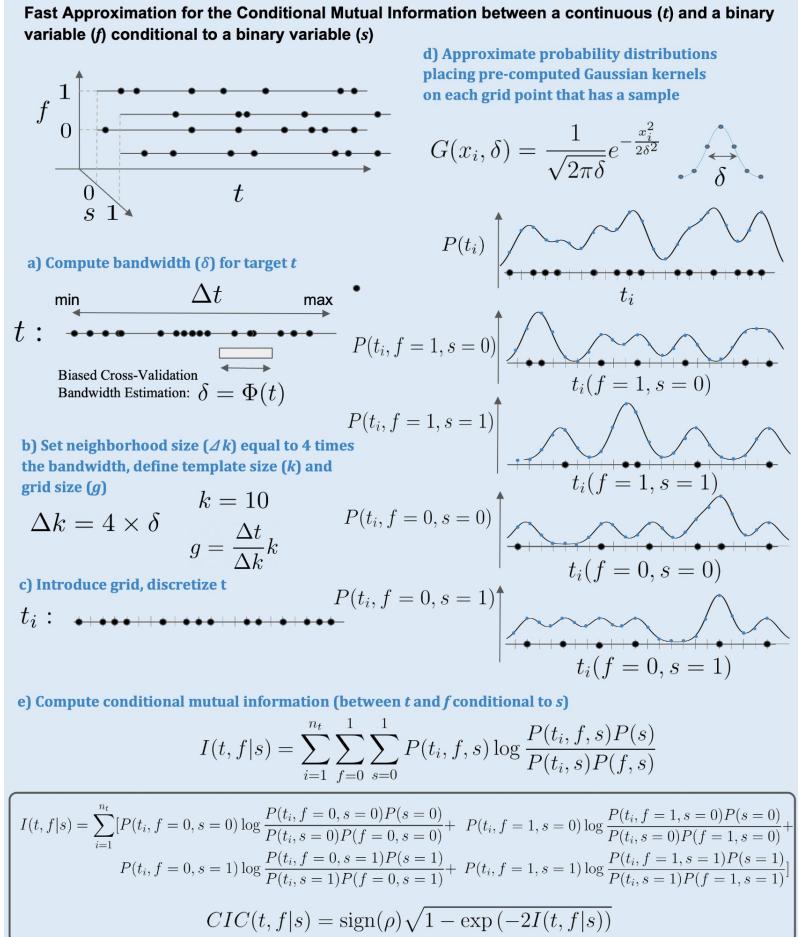
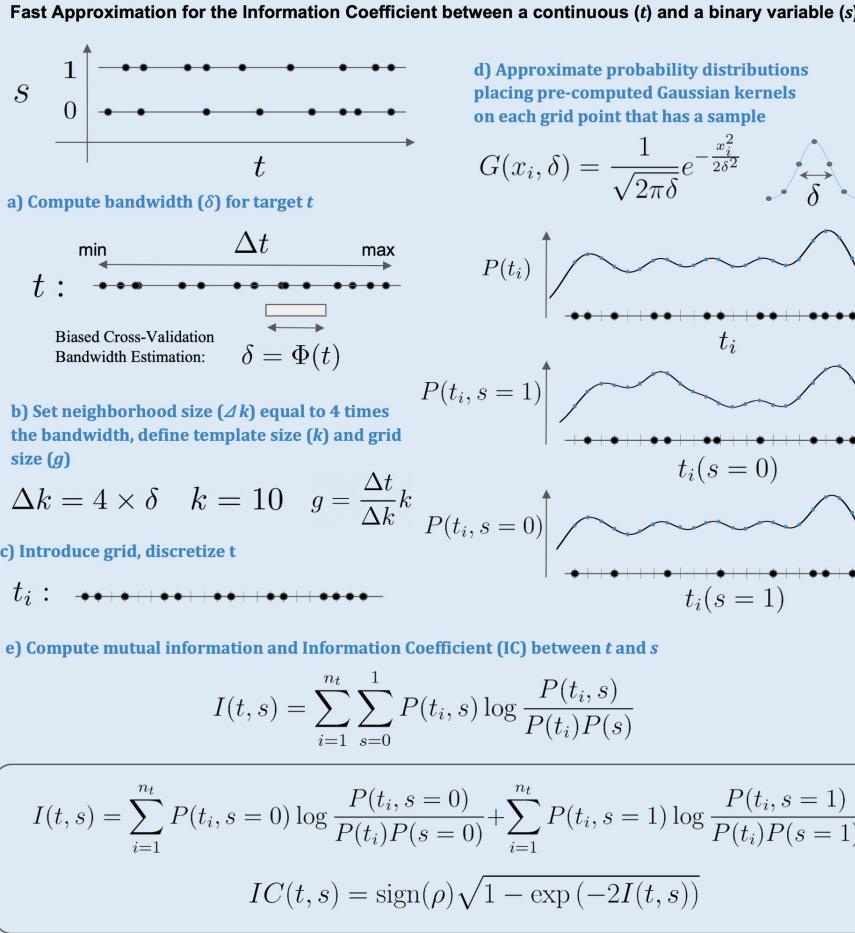
$$I(t, f|s) = \sum_{i=1}^{n_t} \sum_{f=0}^1 \sum_{s=0}^1 P(t_i, f, s) \log \frac{P(t_i, f, s)P(s)}{P(t_i, s)P(f, s)}$$

Conditional Information Coefficient (CIC)

$$CIC(t, f|s) = \text{sign}(\rho) \sqrt{1 - \exp(-2I(t, f|s))}$$



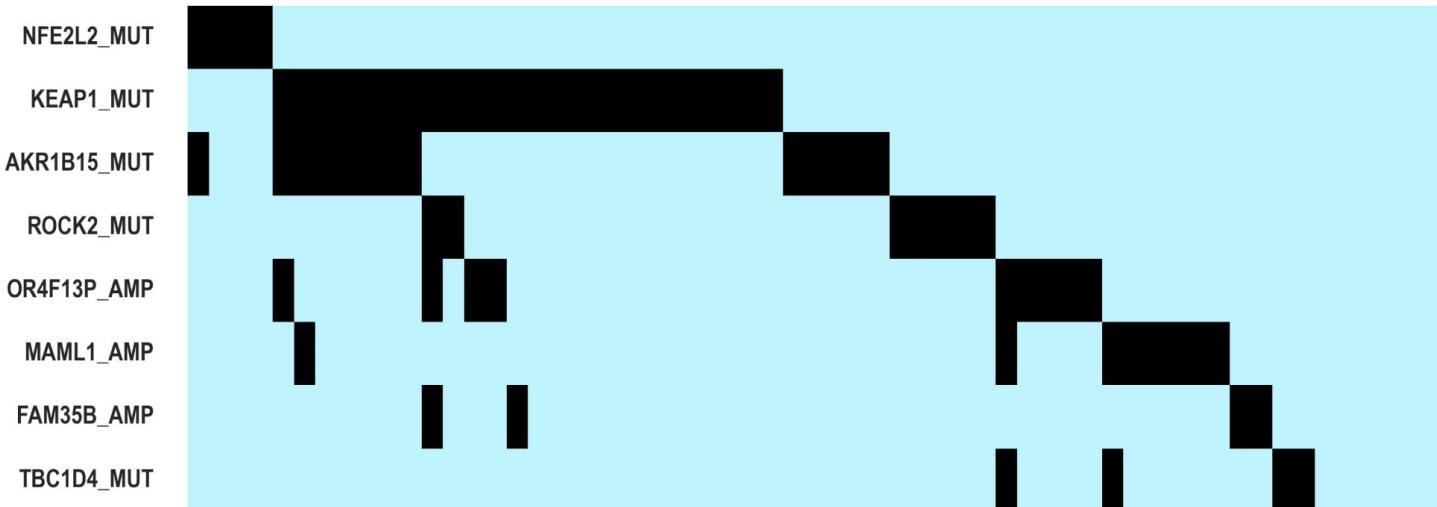
eREVEALER: A new more efficient implementation of the algorithm in Python/Cython



NRF2 Example: REVEALER Results



Top $\frac{1}{3}$ of samples with higher target profiles

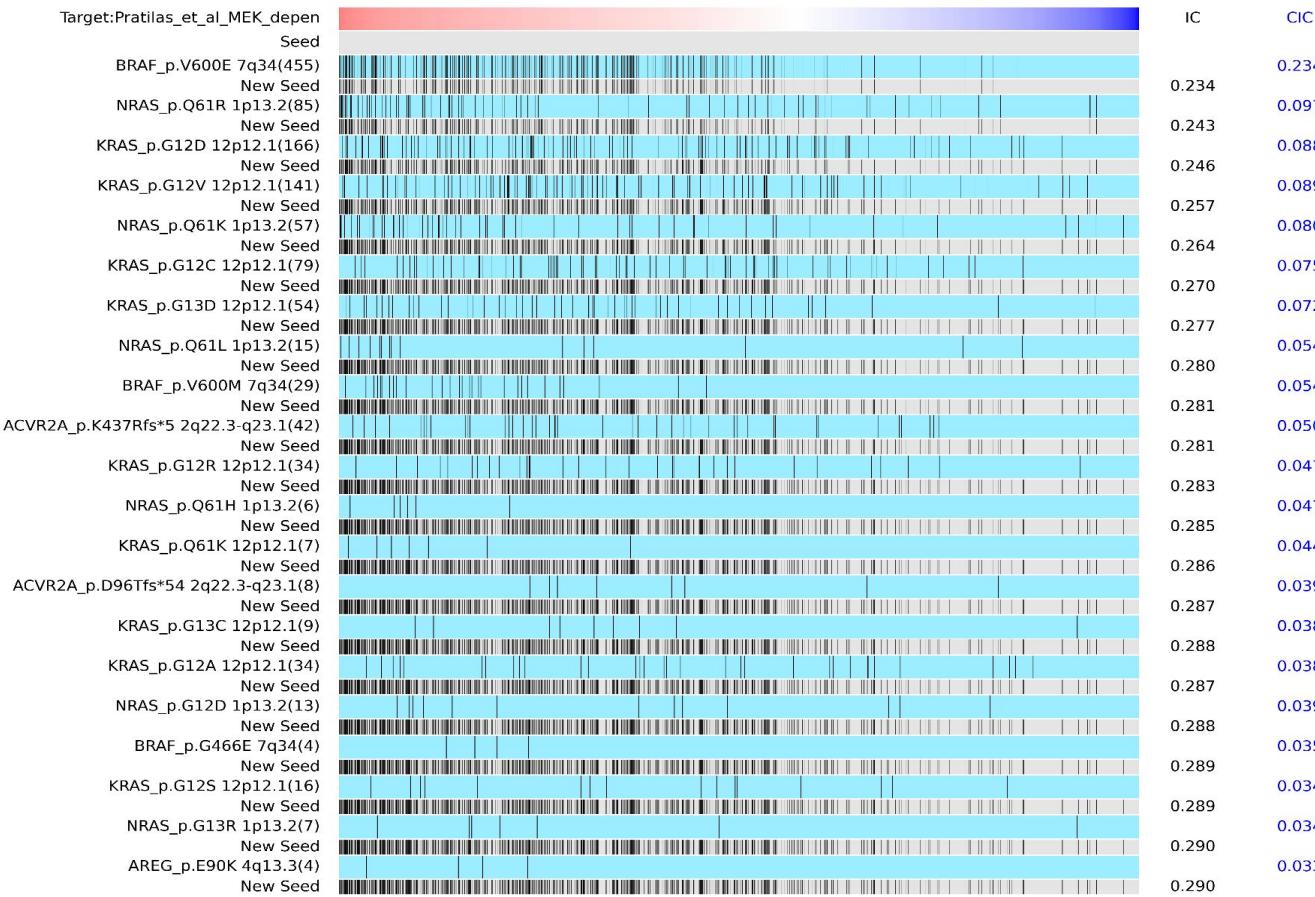


Do the REVEALER Findings Make Sense?

- **KEAP1_MUT:** Cancer-derived mutations in KEAP1 impair NRF2 degradation and are frequent in Squamous Cell Lung Carcinoma (*Hast et al. Cancer Res 2014 Feb 1;74(3):808-17*).
- **AKR1B15_MUT:** is a member of the aldo-keto reductase family and among all the human genes regulated by the Keap1/Nrf2 pathway, they are consistently the most overexpressed in response to Nrf2 activators (*Penning Chem Res Toxicol. 2017 Jan 17; 30(1): 162–176*).
- **OR4F13P_AMP (15q26):** This CNA represents amplification of chr15q22/26 containing NOX5 (NADPH oxidase 5). NOX5 alpha and beta isoforms have been shown to increase the production of extracellular superoxide and H₂O₂, and in turn to increase the activity of NRF2. In *Kim et al. 2016* we experimentally validated this finding (*Kim et al. Nat Biotechnol. 2016 May;34(5):539-46.*)
- **MAML1_AMP (5q35):** This region contains SQSTM1. Its gene product p62 activates NRF2 and promotes resistance to redox stress (*Li et al. Cancer Cell. 2013 Dec 9; 24(6): 738–750*).
- **FAM35B_AMP (10q11):** This region contains MAPK8, the JNK1 protein, encoded by MAPK8, up-regulates the expression of NRF2 (*Namani et al. Aging (Albany NY). 2019 Dec 31; 11(24): 12600–12623.*)
- **TBC1D4_MUT:** TBC1D4 is part of the insulin signaling pathway, and is possible that NRF2 affect TBC1D4 by modulating the cellular redox state and influencing insulin sensitivity (*Tessneer et al. Endocrinology 2014 Sep;155(9):3315-28.*)
- **ROCK2_MUT:** Not obvious at first sight. To be analyzed in more depth.

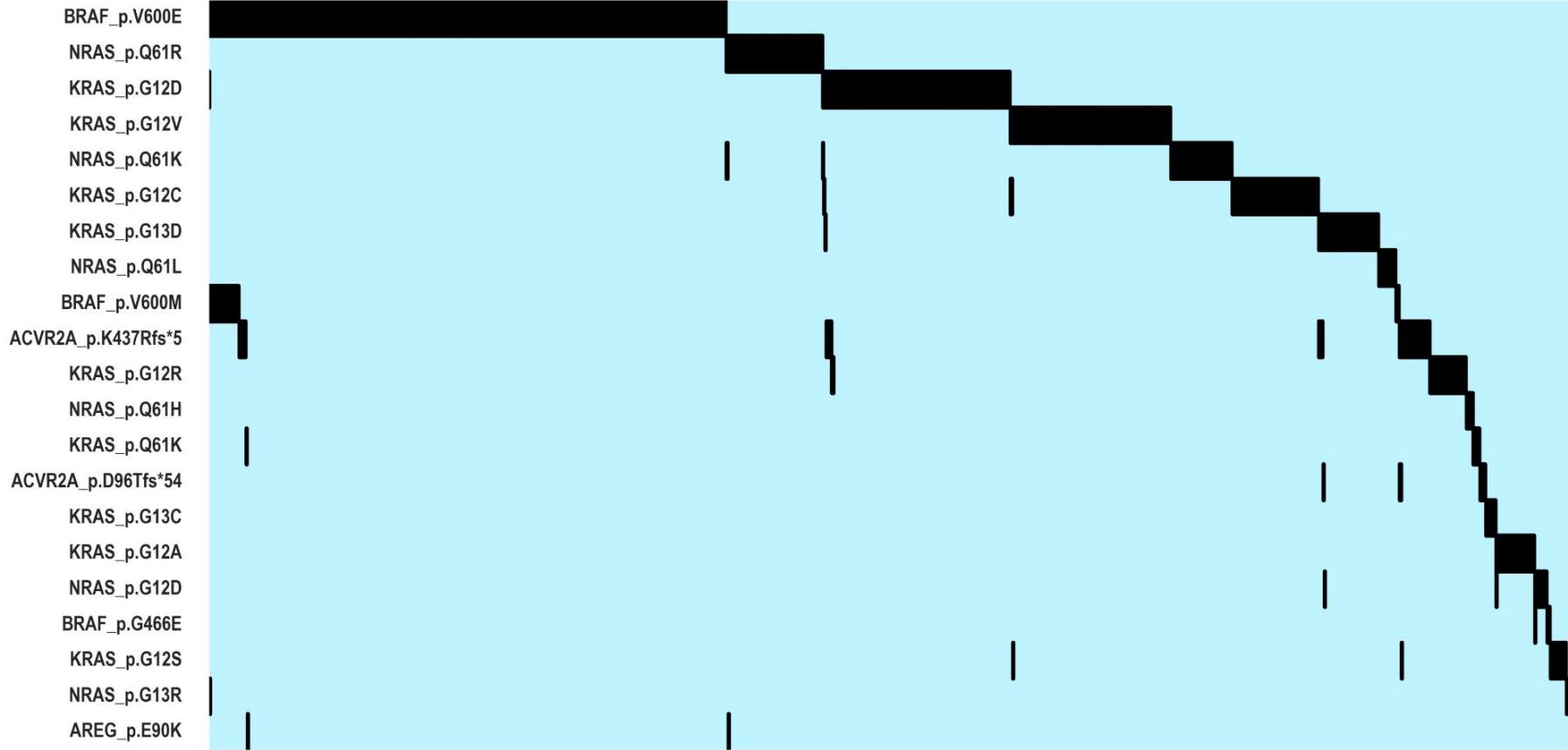
Individual alleles associated with a MEK inhibitor sensitivity signature

TCGA Pan-Cancer dataset (10,956 samples, 32 tissues)



The long long tail of genomic alterations that activate MAPK

Samples (1202 14.66%) with at least one mutation in top 25.0% of samples with highest target



The *Celsus* approach to predict drug response



UCSD Center for Cancer Target Discovery and Development

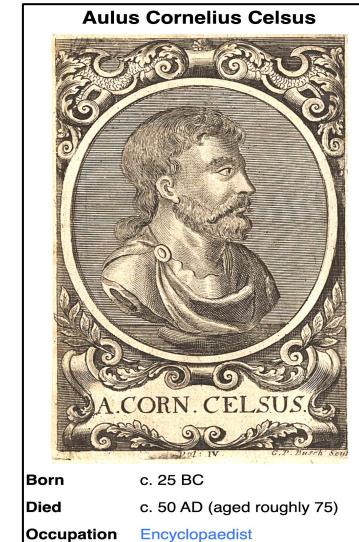


The *Celsius* Project

A Prototype Machine Learning/AI System to
Predict Novel Therapeutic Approaches and
Individual Patient Response to Cancer Drugs

*I am of opinion that the **Art of Medicine** ought to be **rational**, but to draw instruction from **evident causes**, all obscure ones being rejected from the practice of the Art, although not from the practitioner's study.*

—Aulus Cornelius Celsus (c. 25 BC – c. 50 AD)



Machine Learning/AI Inferential System

Generation/Training

A

Genomic Profiles
and CSA features

States
Membership

Transcriptional
Signatures

Gene Sets

Mutations and CNA
(single genes and signatures)

Epigenetics
(e.g. Methylation)

Protein Expression

Response/Resistance
Phenotypes
for many Drugs

B Information-Theoretic Feature Selection

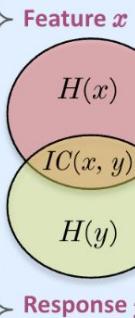
Information
Coefficient

Informative
Feature Set
(IFS)

$H(x)$

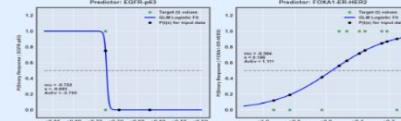
$H(y)$

$IC(x, y)$



C Generating Weight of Evidence (WoE) Models

Conditional Probabilities



Bayesian Posterior log odds

$$Ev(y|x) = \log \frac{P(y=1|x)}{P(y=0|x)}$$

$$\frac{P(y=1|x)}{P(y=0|x)} = \frac{P(y=1)}{P(y=0)} \cdot \frac{P(x|y=1)}{P(x|y=0)}$$

Independent
Samples
Genomic Profiles
and CSA features

Applying/Querying

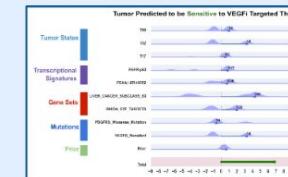
E Large Language Model (LLM)

Evidence
Table(s) Sample(s) IFS

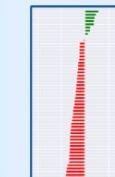


D Bayesian Nomograms

per sample/drug



per sample
across drugs



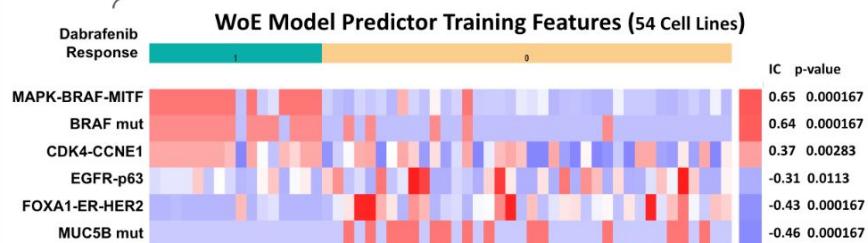
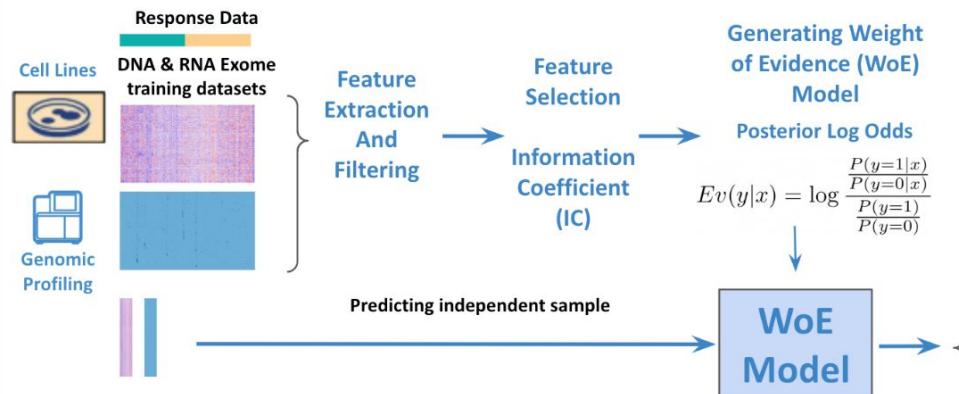
per drug
across samples



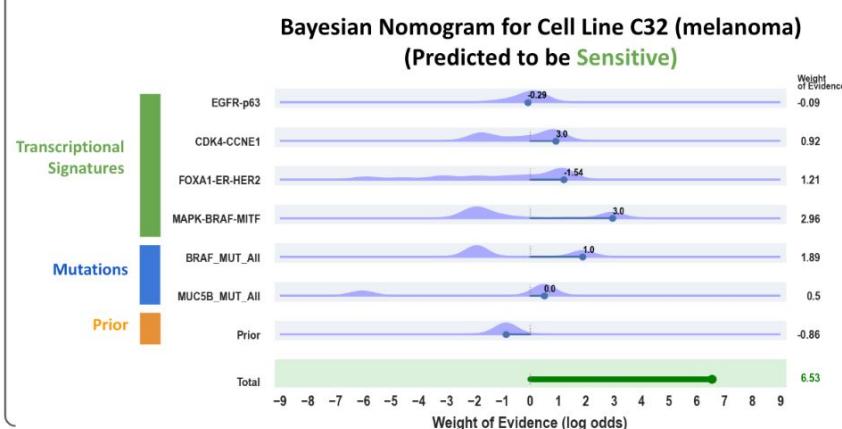
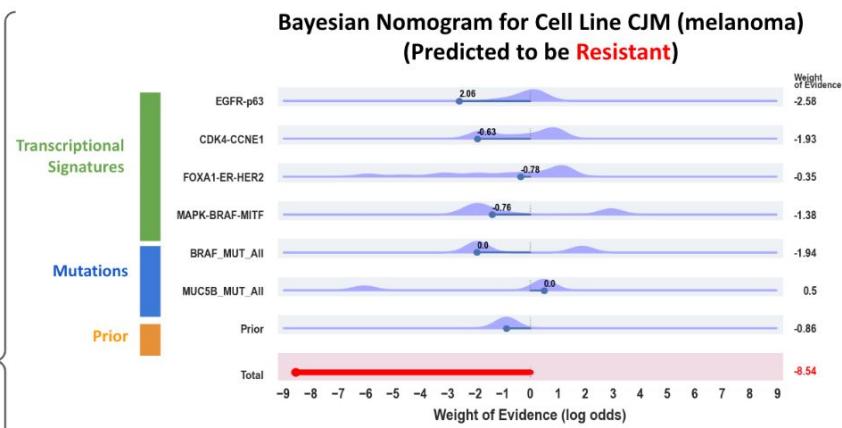
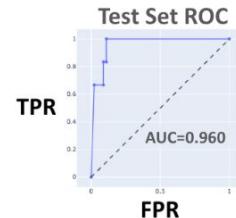
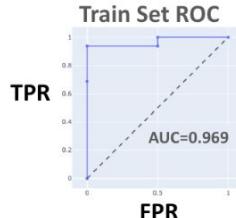
F User/LLM Interaction

- ↔ Summing Evidence
- ↔ Specific Questions
- ↔ In-Depth Follow-Up
- ↔ What-If Scenarios
- ↔ Analysis of Evidence Patterns

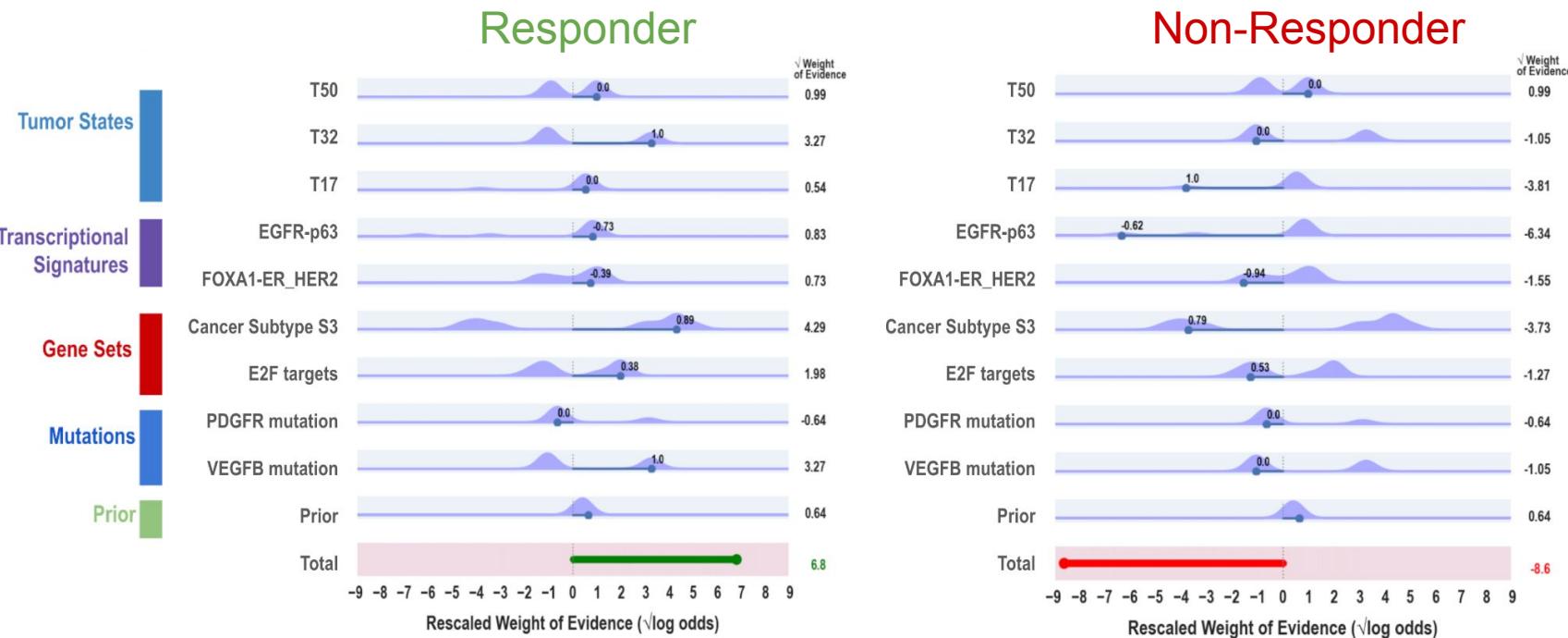
Celsus Bayesian Weight of Evidence Model (WoE) to predict Response to BRAF Inhibitor Dabrafenib



Predictive Performance



Another example: multi-omic model that predicts response to a targeted therapy in patient samples



Use of Large Language Models (LLM) to Analyze Patterns of Evidence in the Bayesian Model

Analysis of Evidence Patterns

This attached table contains as rows a group of genomic features and as columns the amount of evidence, i.e. the posterior log odds of response to the drug for each of 9 tumor samples. If you analyze the patterns of evidence across the 9 tumor samples, can you see interesting configurations or patterns that you may want to highlight?

Example_for_LLM_2.txt
TEXT - 680 BYTE

Observations and Patterns

1. T32: High Variability

- Samples 2, 3, and 4 show high values (10.67) compared to other samples, which are consistently low (-1.10).
- This suggests these three samples might have a distinct genomic profile related to T32.

2. HOSHIDA_LIVER_CANCER_SUBCLASS_S3: High Positive and Negative Extremes

- Sample 1 has a very high value (35.64), while samples 8 and 9 have very low values (-69.92 and -20.71, respectively).
- This feature exhibits significant variability, indicating diverse responses among the samples.

3. EGFR-p63: Strong Negative Outliers

- Sample 6 shows a very negative value (-40.24) and sample 8 also has a significantly negative value (-11.86), while the other samples are relatively stable around 0.69.
- These outliers suggest a potential strong inhibitory effect in these samples.

4. FOXA1-ER-HER2: Mixed Responses

- This feature shows a range from positive to negative values across samples, with sample 7 having the highest positive value (1.82) and sample 6 showing a notable negative value (-2.40).
- The variation indicates differing impacts of this feature on the tumor samples.

5. Consistent Low or High Values Across Features:

- PDGFRB_Missense_Mutation:** Mostly consistent with -0.41 except for samples 2 and 4 (9.87).
- VEGFB_Nonsilent:** Mostly consistent with 10.67 or -1.10, showing a binary pattern.
- These features might represent binary-like genomic variations in the samples.

6. T50: Mixed Responses Similar to FOXA1-ER-HER2:

- Samples 2, 3, 4, and 6 show positive values (0.98), while other samples have negative values (-0.81).
- Indicates potential differential response to the drug among the samples.

Specific Questions

Which feature provides the strongest evidence for the positive drug's response

④ GPT-4o Poe

To determine which feature provides the strongest evidence for a positive drug response, we need to identify the feature with the highest positive value in the "Evidence" column.

Here is the relevant part of the table:

Feature	Evidence
T50	-0.81
T32	-1.10
T17	0.29
EGFR-p63	0.69
FOXA1-ER-HER2	1.30
HOSHIDA_LIVER_CANCER_SUBCLASS_S3	35.64
ISHIDA_E2F_TARGETS	0.46
PDGFRB_Missense_Mutation	-0.41
VEGFB_Nonsilent	10.67
prior	0.41

From the table, the feature **HOSHIDA_LIVER_CANCER_SUBCLASS_S3** has the highest positive evidence with a value of 35.64.

Therefore, the feature **HOSHIDA_LIVER_CANCER_SUBCLASS_S3** provides the strongest evidence for a positive response to the drug

What-If Scenarios

Suppose that you had an identical tumor sample to the one in the table but that it had the **VEGFB_Nonsilent** feature equal to 0 (wild-type) instead of 1 (mutant) with a corresponding amount of evidence equal to -1.099. In that scenario would that sample be predicted to respond to the drug?

④ GPT-4o Poe

To determine whether the modified tumor sample would be predicted to respond to the drug we need to adjust the evidence provided by the **VEGFB_Nonsilent** feature and then recalculate the total amount of evidence.

Here's the updated table with the modified feature:

Feature	Value	Evidence
T50	1	-0.81
T32	0	-1.10
T17	0	0.29
EGFR-p63	-0.725	0.69
FOXA1-ER-HER2	-0.247	1.30
HOSHIDA_LIVER_CANCER_SUBCLASS_S3	1129	35.64
ISHIDA_E2F_TARGETS	9190	0.46
PDGFRB_Missense_Mutation	0	-0.41
VEGFB_Nonsilent	0	-1.099
prior	0.6	0.41

Now let's sum the adjusted evidence:

apache ⌂ Copy
$$-0.81 + (-1.10) + 0.29 + 0.69 + 1.30 + 35.64 + 0.46 + (-0.41) + (-1.099) + 0.41$$

Calculating step by step:

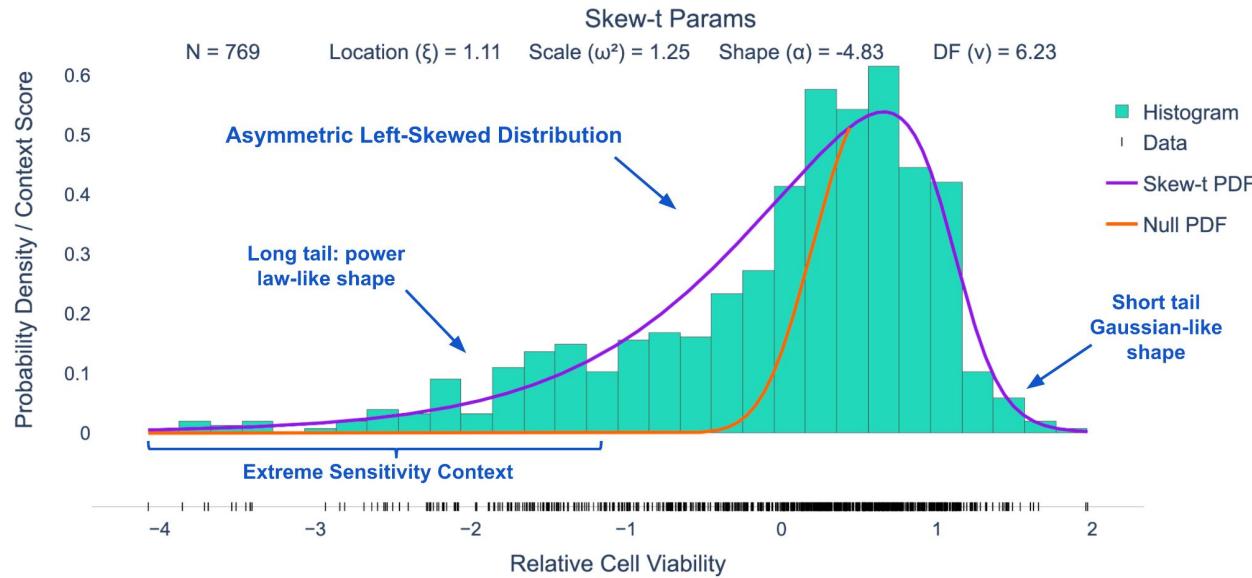
$$\begin{aligned} 1. -0.81 + (-1.10) &= -1.91 \\ 2. -1.91 + 0.29 &= -1.62 \\ 3. -1.62 + 0.69 &= -0.93 \\ 4. -0.93 + 1.30 &= 0.37 \\ 5. 0.37 + 35.64 &= 36.01 \\ 6. 36.01 + 0.46 &= 36.47 \\ 7. 36.47 + (-0.41) &= 36.06 \\ 8. 36.06 + (-1.099) &= 34.961 \\ 9. 34.961 + 0.41 &= 35.371 \end{aligned}$$

The total sum of the evidence is 35.371.

Since the total sum is still greater than zero, the modified tumor sample would still be predicted to respond to the drug

Predicting Response and Analysing the Sensitivity Landscape of Olaparib

Modeling Profile of Drug Sensitivity with a Skew-t Distribution



Student's t -density

$$t(x, y) = \frac{\Gamma((y+1)/2)}{\sqrt{y\pi} \Gamma(y/2)} (1 + x^2/y)^{-(y+1)/2}$$

Student's t -distribution

$$T(x, y) = \int_{-\infty}^x t(u, y) du = 1 - \frac{1}{2} I_{\phi}(y/2, 1/2)$$

I_{ϕ} is the Incomplete Beta Function, $\phi = \frac{y}{x^2 + y}$.

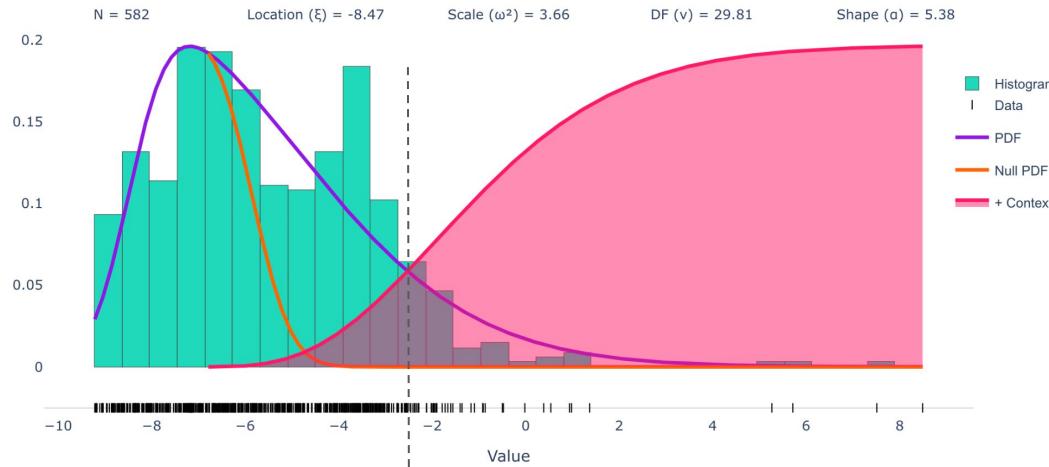
Skew-t Distribution

$$S_T(z; \xi, \omega^2, \alpha, \nu) = \frac{2}{\omega} t(z, \nu) T(\alpha z \sqrt{\frac{\nu+1}{\nu+z^2}}, \nu+1)$$

Skew-t Analysis of the Olaparib Sensitivity Profile in Cell Lines

(sensitivity is on the right side)

olaparib (PARP1;PARP2)



Threshold
of
sensitivity

Olaparib response

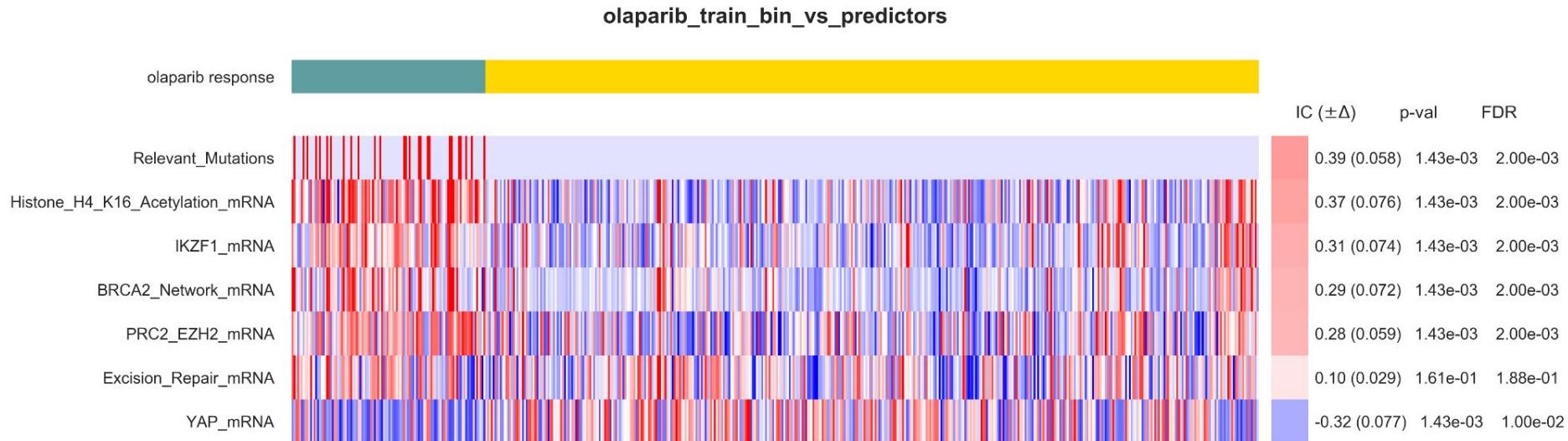


Many allelic mutations align with olaparib response but they are weak and infrequent predictors

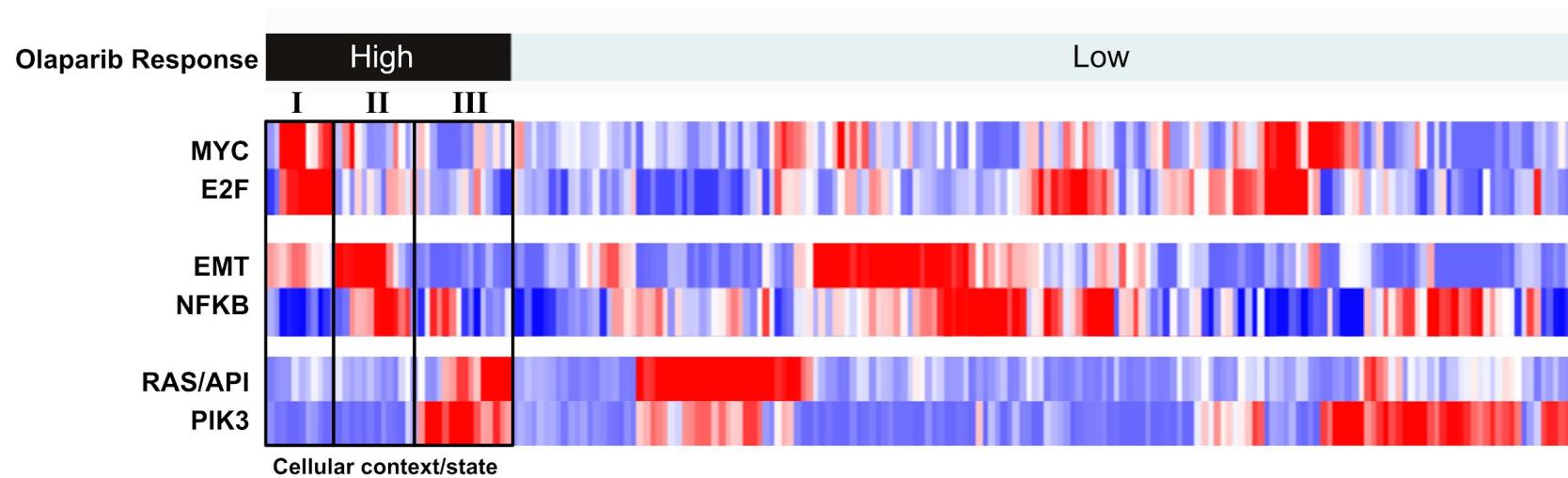
These 50 are combined into one *Relevant Mutations* feature



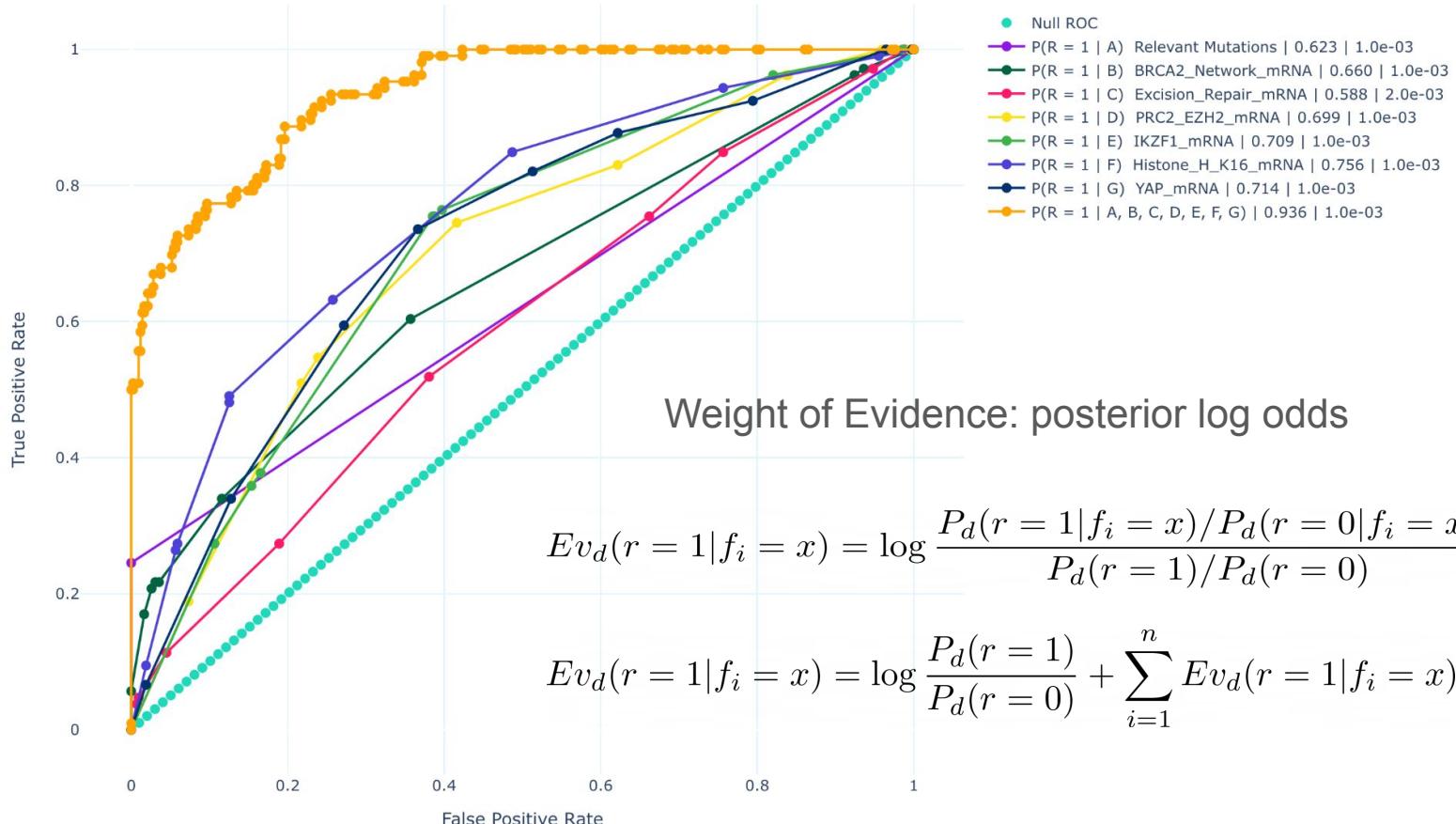
These are Genomic Features Motivated by our Biological Understanding of the Drug Mechanism of Action and also Supported by the Data



Cancer cell lines profiles of transcriptional components representing major oncogenic pathways suggest there are 3 cellular context/states (I, II, III) that dominate PARPi sensitivity.



Increasing predictive capability for olaparib response as more genomic features provide their complementary weight of evidence to the Bayesian model



Supplementary Slides

vsNMF Details

Matrix factorization and generation of transcriptional signatures. The resulting reference dataset (A_{nm} , $n = 3,958$ genes, $m = 1,080$ samples) is then *factorized* into a product of two lower-rank matrices,

$$A_{nm} \approx W_{nk} \times H_{km}, \quad (1)$$

using a new *variable-sparsity* NMF (vsNMF) iterative gradient descent algorithm (manuscript in preparation), minimizing the Euclidean error $E(W, H) = \sum_{ij}(A_{ij} - (W \times H)_{ij})^2$, where the standard W and H matrix updates in the direction of the negative gradient,⁴²

$$\begin{aligned} W_{t+1} &= W_t - \mu_W (W_t H_t - A) H_t^T & H_{t+1} &= H_t - \mu_H W_t^T (W_t H_t - A) \\ \mu_W &= \frac{W_t}{W_t H_t H_t^T} & \mu_H &= \frac{H_t}{W_t^T W_t H_t} \end{aligned} \quad , \quad (2)$$

are followed with an additional projection step for the H matrix onto a sparsity-constrained space,⁴³ in such a way that the rows of H obey preset values of a sparsity function $S(x)$. For an H row vector x of n elements, this function measures how much the ratio between L_1 (manhattan) and L_2 (euclidean) norms deviates from its dense limit (\sqrt{n}),^{44,45}

$$S(x) = \frac{\sqrt{n} - \frac{L_1(x)}{L_2(x)}}{\sqrt{n} - 1} \quad L_1(x) = \sum_{i=1} |x_i| \quad L_2(x) = \sqrt{\sum_{i=1} x_i^2} \quad . \quad (3)$$

One remarkable property of NMF is its ability to produce a *dimensionality-reduced* representation of an input dataset that encodes a significant amount of data relationships using only a few positive components, exposing common patterns, and providing a working summary of the data.^{1,42,46,47} The vsNMF algorithm we described above extends the approach of Hoyer⁴³ and allows us to control the degree of sparseness and the corresponding balance between local vs. global structure in the NMF factorization. This increases the capability of NMF to capture biologically relevant localized patterns in the dataset (e.g. cellular states). In this way, it is advantageous to define a power-law distribution or sparsities, $S_k = a - b/k^\alpha$ for H and leaving W unconstrained, so that the H matrix structure has more emphasis on local structure than the original NMF algorithm, which produces a roughly uniform distribution of H -sparsities. In the end, the W_{nk} matrix contains the weight of each vsNMF *transcriptional signature* and the H_{km} matrix corresponds to a “mixture” of signatures that approximates each A_{nm} column.

Signature Annotation

Signature annotation and short-naming. The svNMF decomposition is a data-driven unbiased procedure; however, the transcriptional signatures are most useful when they can be interpreted biologically. This is achieved by annotating and characterizing them by matching their sample profiles (rows of H) against single-sample GSEA⁹ profiles (Fig. 1C), derived from the reference dataset, using gene sets from the *Molecular Signatures Database* (MSigDB,^{8,48}). The matching estimates the *mutual information* $I(h_i, g_j)$ between each row of H (h_i) and the enrichment scores of each gene set (g_j).

$$I(h_i, g_j) = \sum_{h_i} \sum_{g_j} P(h_i, g_j) \frac{P(h_i, g_j)}{P(h_i)P(g_j)} \quad IC(h_i, g_j) = \text{sign}(\rho(h_i, g_j)) \sqrt{1 - e^{-2I(h_i, g_j)}} \quad . \quad (4)$$

The mutual information is rescaled to lie in the [-1, 1] interval to define the *Information Coefficient* (IC) that measures dependency between variables and can capture linear and non-linear dependencies.^{5,49} The sign of the standard correlation coefficient $\rho(h_i, g_j)$ is used to provide *directionality*. The gene sets most associated with each *transcriptional signature* profile provide comprehensive annotations and become part of the CSA model. These annotations are also used to give biological short-name identifiers to the signatures (Fig 1D, Fig 2B).

|

Defining Cellular and Tumor States

Defining cellular states (clustering). The H matrix profiles from the vsNMF decomposition are clustered using a *Self-Organizing Map* algorithm (SOM,^{10,50}) to define putative *states*, namely, groups of samples with similar *signature* amplitudes (Fig. 1E, Fig. 2A). The SOM algorithm also defines a 2D map that organizes the collection of states onto a 2D hexagonal layout. The n_s states are further *refined* by removing samples that do not share enough similarity with their state centroid. Finally, the states and archetypes are re-arranged on a hexagonal layout (Fig. 1E, Fig. 2D) and to maximize the similarity (*IC*) of neighboring states on the layout using a *Simulated Annealing* algorithm (SA^{51,52}). For each state a corresponding *archetype* is computed using *median values* of the *state* members' *signature* amplitudes (Fig. 2C). The *archetypes* can be used as *barcodes* and visualized using linear or circular heatmaps.

Defining tumor states. To define tumor states we select datasets representing bulk tumor mRNA and preprocess and merge them in the same way as the *reference dataset* described above. Then the resulting tumor dataset, B , is projected onto the space of the transcriptional signatures using a *non-negative least squares solver*⁵³ for each column of B ,

$$H^B = [h_1, h_2, \dots, h_{m_B}] : \arg \min_{h_i} \|W_{n_B k} \times h_i - b_i\|^2 \text{ for } h_i \geq 0 \quad (5)$$

where n_B is the set of genes in common between B and the *reference* dataset A . Once projected the sample profiles H^B are clustered using the SOM and Simulated Annealing algorithms described above. In the case of the *CSA Model 1.0*, the result is a 72-state tumor state model with a corresponding hexagonal layout and archetypes (Fig 1G-H) generated using the solid tumor pan-cancer TCGA mRNA dataset.⁵⁴ As the cellular and tumor states share the same transcriptional signatures, they can be associated with each other, e.g., by using single-cell data and pseudo-bulk representations of tumors. This provides a many-to-many mapping between the cellular and tumor states as they are observed in tumor samples. Another approach to mapping them makes use of a *mixture model* by mapping one model (tumor) states vs. the other model (cellular).

Characterizing Cellular and Tumor States

Cellular and tumor state characterization. The cellular and tumor states can be further studied, annotated, and characterized, and provide additional information to augment a CSA model. This can be accomplished, for example, by comparing the genomic features of the samples that are members of each state vs. the rest (Fig 1I-J, Fig 2E). Multi-omic properties, such as differential gene, protein, and pathway expression, or the enrichment of specific gene sets, or specific genomic abnormalities, e.g., mutations, copy number, etc. can be very informative in terms of delineating the relevant makeup of the state and help explain the natural history of each cancer as a distinct biological entity. Another approach to providing a deeper annotation of the transcriptional signatures and the CSA states is to apply the REVEALER^{5,6} algorithm to delineate their most relevant genetic alterations landscapes. This type of powerful association analysis can provide a link between a functional view of cancer,^{4,37,55,56} such as that represented in a CSA model, and more traditional ways to model cancers based on a structural characterization of the cancer genome, driver vs. passenger mutations, etc.

References

1. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4164–4169 (2004).
2. Hoshida, Y., Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Subclass mapping: identifying common subtypes in independent disease data sets. *PLoS One* **2**, e1195 (2007).
3. Tamayo, P. *et al.* Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 5959–5964 (2007).
4. Kim, J. W. *et al.* Decomposing Oncogenic Transcriptional Signatures to Generate Maps of Divergent Cellular States. *Cell Syst* **5**, 105–118.e9 (2017).
5. Kim, J. W. *et al.* Characterizing genomic alterations in cancer by complementary functional associations. *Nat. Biotechnol.* **34**, 539–546 (2016).
6. Ma, J. Y. *et al.* Deciphering the Functional Roles of Individual Cancer Alleles Across Comprehensive Cancer Genomic Studies. *bioRxiv* (2023) doi:10.1101/2023.11.14.567106.
7. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
8. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
9. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
10. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 2907–2912 (1999).
11. Kasabov, N. *et al.* Integrating Local and Personalised Modelling with Global Ontology Knowledge Bases for Biomedical and Bioinformatics Decision Support. In *Computational Intelligence in Biomedicine and Bioinformatics: Current Trends and Applications* (eds. Smolinski, T. G., Milanova, M. G. & Hassani, A.-E.) 93–116 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008).
12. Wu, Y., Tamayo, P. & Zhang, K. Visualizing and Interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding. *Cell Syst* **7**, 656–666.e4 (2018).
13. Hanford, A. R. *et al.* DiSCoVERing Innovative Therapies for Rare Tumors: Combining Genetically Accurate Disease Models with In Silico Analysis to Identify Novel Therapeutic Targets. *Clin. Cancer Res.* **22**, 3903–3914 (2016).
14. Banerjee, S. *et al.* KITlow Cells Mediate Imatinib Resistance in Gastrointestinal Stromal Tumor. *Mol. Cancer Ther.* **20**, 2035–2048 (2021).
15. Cañadas, I. *et al.* Tumor innate immunity primed by specific interferon-stimulated endogenous retroviruses. *Nat. Med.* **24**, 1143–1150 (2018).
16. Konieczkowski, D. J. *et al.* A melanoma cell state distinction influences sensitivity to MAPK pathway inhibitors. *Cancer Discov.* **4**, 816–827 (2014).
17. Tamayo, P. *et al.* Predicting relapse in patients with medulloblastoma by integrating evidence from clinical and genomic features. *J. Clin. Oncol.* **29**, 1415–1423 (2011).
18. Rusert, J. M. *et al.* Functional Precision Medicine Identifies New Therapeutic Candidates for Medulloblastoma. *Cancer Res.* **80**, 5393–5407 (2020).
19. Deichaite, I. *et al.* Differential regulation of TNF α and IL-6 expression contributes to immune evasion in prostate cancer. *J. Transl. Med.* **20**, 527 (2022).
20. Huang, F. W. *et al.* Exome Sequencing of African-American Prostate Cancer Reveals Loss-of-Function Mutations. *Cancer Discov.* **7**, 973–983 (2017).
21. Feng, X. *et al.* A Platform of Synthetic Lethal Gene Interaction Networks Reveals that the GNAQ Uveal Melanoma Oncogene Controls the Hippo Pathway through FAK. *Cancer Cell* **35**, 457–472.e5 (2019).
22. Ren, S. *et al.* HPV E2, E4, E5 drive alternative carcinogenic pathways in HPV positive cancers. *Oncogene* **39**, 6327–6339 (2020).
23. Liu, C. *et al.* Cannabinoids Promote Progression of HPV-Positive Head and Neck Squamous Cell Carcinoma via p38 MAPK Activation. *Clin. Cancer Res.* **26**, 2693–2703 (2020).
24. Ando, M. *et al.* Chromatin dysregulation and DNA methylation at transcription start sites associated with transcriptional repression in cancers. *Nat. Commun.* **10**, 2188 (2019).
25. Panuganti, B. A., Carico, C., Jeyarajan, H., Flagg, M. & Tamayo, P. Transcriptional subtypes of glottic cancer characterized by differential activation of canonical oncogenic programming. *Head Neck* **45**, 2851–2861 (2023).

26. Guo, T. *et al.* Clinical and genomic characterization of chemoradiation-resistant HPV-positive oropharyngeal squamous cell carcinoma. *Front. Oncol.* **14**, 1336577 (2024).
27. Wu, X. *et al.* Metformin Inhibits Progression of Head and Neck Squamous Cell Carcinoma by Acting Directly on Carcinoma-Initiating Cells. *Cancer Res.* **79**, 4360–4370 (2019).
28. Flynn, S. M. *et al.* The multikinase inhibitor RXDX-105 is effective against neuroblastoma and. *Oncotarget* **10**, 6323–6333 (2019).
29. Reilly, B. M. *et al.* DNA Methylation Identifies Genetically and Prognostically Distinct Subtypes of MDS. *Blood* vol. 132 106–106 Preprint at <https://doi.org/10.1182/blood-2018-99-111535> (2018).
30. Stewart, M. L. *et al.* KRAS Genomic Status Predicts the Sensitivity of Ovarian Cancer Cells to Decitabine. *Cancer Res.* **75**, 2897–2906 (2015).
31. Cheung, H. W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 12372–12377 (2011).
32. Lee, S., Jun, J., Kim, W. J., Tamayo, P. & Howell, S. B. WNT Signaling Driven by R-spondin 1 and LGR6 in High-grade Serous Ovarian Cancer. *Anticancer Res.* **40**, 6017–6028 (2020).
33. Ren, Y. *et al.* Targeted tumor-penetrating siRNA nanocomplexes for credentialing the ovarian cancer oncogene ID4. *Sci. Transl. Med.* **4**, 147ra112 (2012).
34. Verhaak, R. G. W. *et al.* Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J. Clin. Invest.* **123**, 517–525 (2013).
35. Raeder, M. B. *et al.* Integrated genomic analysis of the 8q24 amplification in endometrial cancers identifies ATAD2 as essential to MYC-dependent cancers. *PLoS One* **8**, e54873 (2013).
36. Viswanathan, V. S. *et al.* Dependency of a therapy-resistant state of cancer cells on a lipid peroxidase pathway. *Nature* **547**, 453–457 (2017).
37. Hahn, W. C. *et al.* An expanded universe of cancer targets. *Cell* **184**, 1142–1155 (2021).
38. Pham, T. V. *et al.* Role of ultraviolet mutational signature versus tumor mutation burden in predicting response to immunotherapy. *Mol. Oncol.* **14**, 1680–1694 (2020).
39. Joshi, S. *et al.* Macrophage Syk-PI3K γ Inhibits Antitumor Immunity: SRX3207, a Novel Dual Syk-PI3K Inhibitory Chemotype Relieves Tumor Immunosuppression. *Mol. Cancer Ther.* **19**, 755–764 (2020).
40. Boichard, A. *et al.* APOBEC-related mutagenesis and neo-peptide hydrophobicity: implications for response to immunotherapy. *Oncoimmunology* **8**, 1550341 (2019).
41. Yard, B. D. *et al.* A genetic basis for the variation in the vulnerability of cancer to DNA damage. *Nat. Commun.* **7**, 11428 (2016).
42. Lee, D. D. & Seung, H. S. Algorithms for Non-negative Matrix Factorization. in *Advances in Neural Information Processing Systems 13* (eds. Leen, T. K., Dietterich, T. G. & Tresp, V.) 556–562 (MIT Press, 2001).
43. Hoyer. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*
44. Rickard, N. H. A. Comparing Measures of Sparsity. *IEEE Transactions on Information Theory* **55**, 4723–4741 (2009).
45. Penghang Yin, Ernie Esser, and Jack Xin. Ratio and difference of l1 and l2 norms and sparse representation with coherent dictionaries. *Communications in Information and Systems* **14**, 87–109 (2014).
46. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
47. Devarajan, K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput. Biol.* **4**, e1000029 (2008).
48. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
49. Linfoot, E. H. An informational measure of correlation. *Information and Control* **1**, 85–89 (1957).
50. Kohonen, T. Self-Organizing Maps. *Springer Series in Information Sciences* Preprint at <https://doi.org/10.1007/978-3-642-97966-8> (1997).
51. Kirkpatrick, S., Gelatt, C. D., Jr & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
52. van Laarhoven, P. J. & Aarts, E. H. *Simulated Annealing: Theory and Applications*. (Springer Science & Business Media, 2013).
53. Bultheel, A. & Cools, R. *The Birth of Numerical Analysis*. (World Scientific, 2010).
54. Wang, Z., Jensen, M. A. & Zenklusen, J. C. A Practical Guide to The Cancer Genome Atlas (TCGA). *Methods Mol. Biol.* **1418**, 111–141 (2016).
55. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564–576.e16 (2017).
56. Howard, T. P. *et al.* Functional Genomic Characterization of Cancer Genomes. *Cold Spring Harb Symp Quant Biol* **81**, 237–246 (2016).

