

# **AI in Medicine (generative modeling)**

Nicholas Chia

Computational Biologist

ANL

# Cancer is bad...

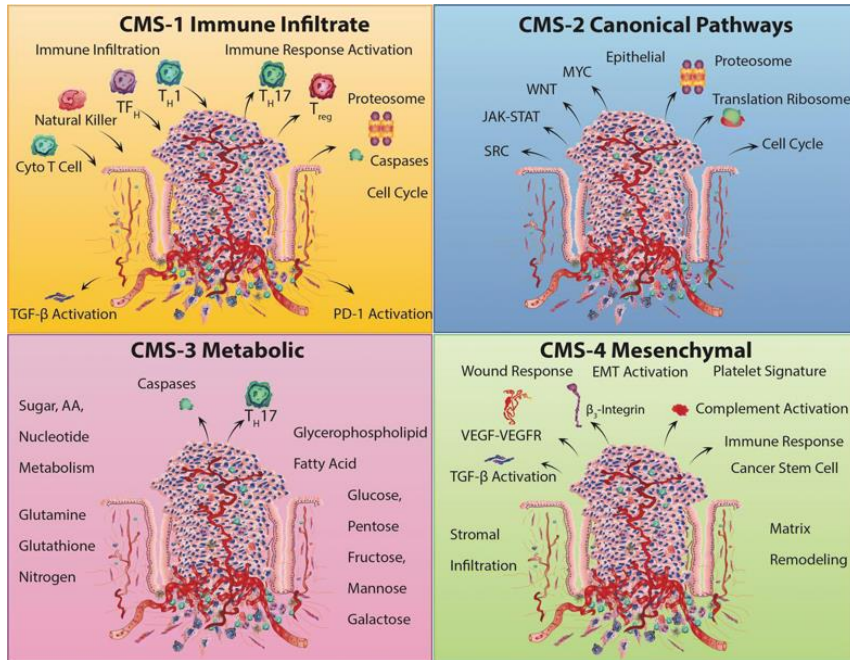
- **2 million** people get **diagnosed with cancer every year** in the United States.
- **42%** of those newly diagnosed cancers are estimated to be potentially **avoidable**.
- **Colorectal cancer**: Second most lethal cancer in U.S.
  - **>85% of tumors** are estimated to be **environmentally driven**, i.e., not arising from a family history or hereditary cause.



# Pattern Recognition & Cancer

## Classification

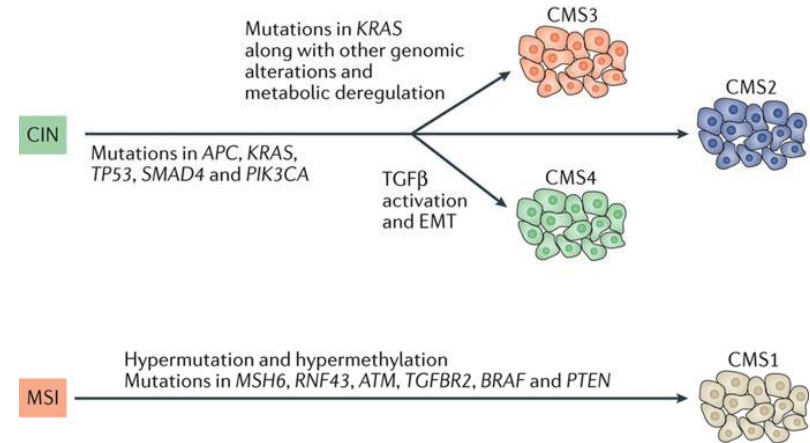
(Colorectal cancer subtype stratification)



Menter et.al (2019) CGR

## Prediction

(Correlative risk of genetic events)



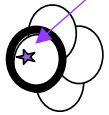
Guinney et.al (2015) Nature

We are thinking about it wrong.

Cancer is not a pattern recognition problem.

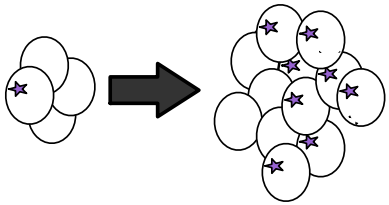
# Cancer is a game of cell evolution

Mutation 'A'

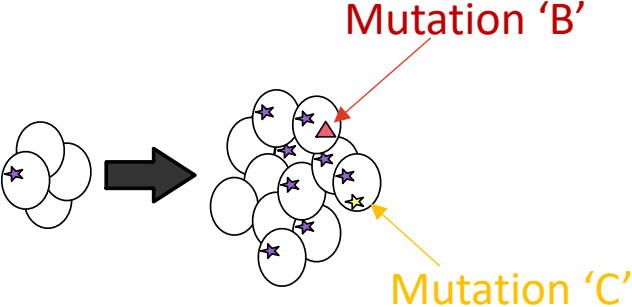


Founder tumor cell

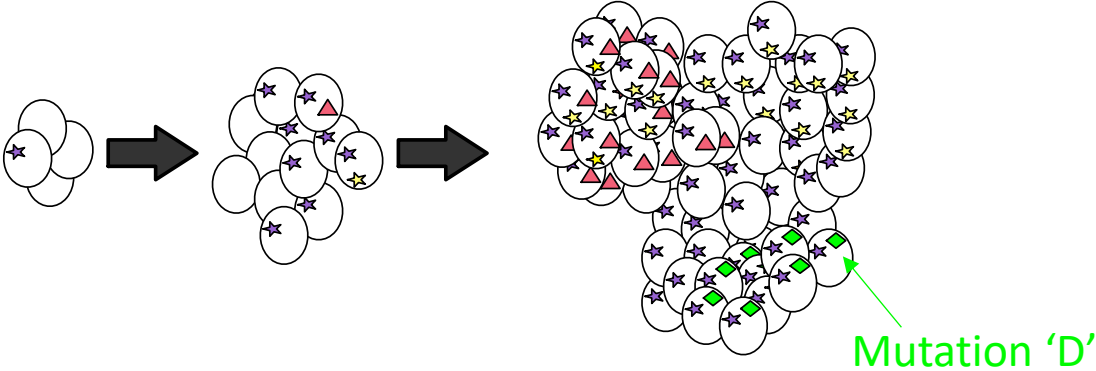
# Cancer is a game of cell evolution



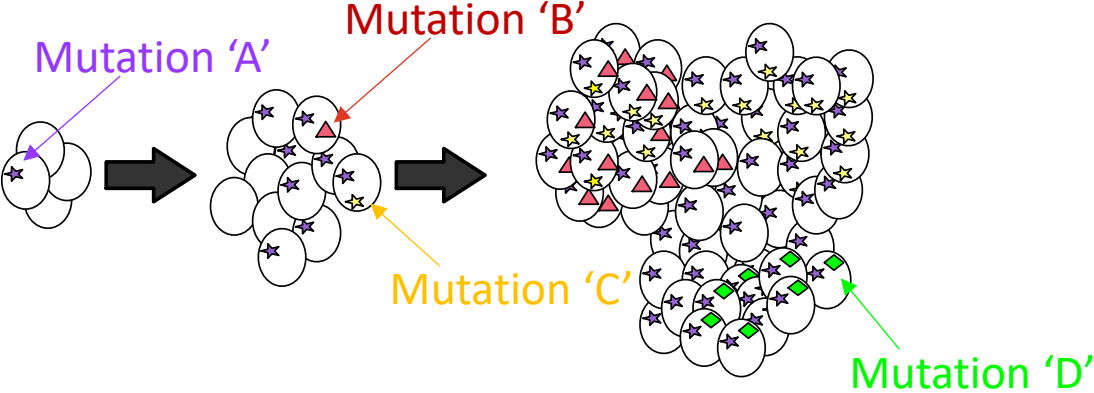
# Cancer is a game of cell evolution



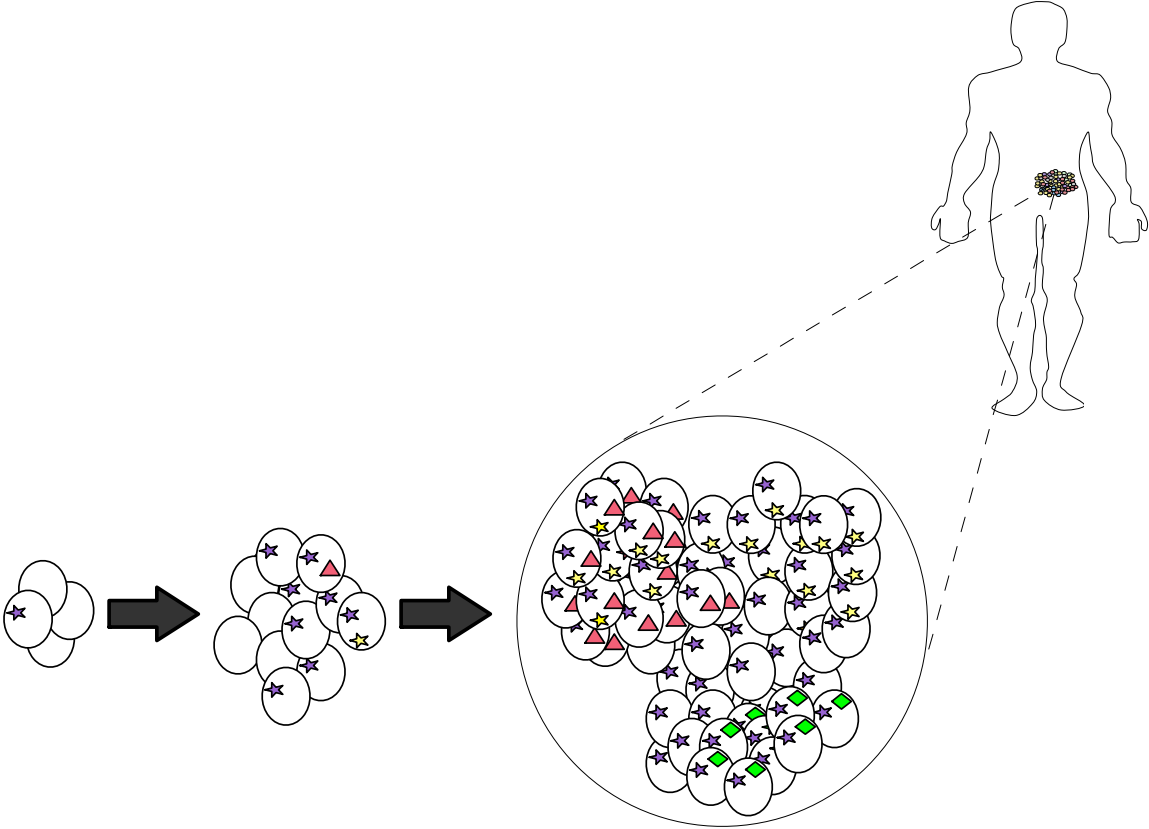
# Cancer is a game of cell evolution



# Cancer is a game of cell evolution

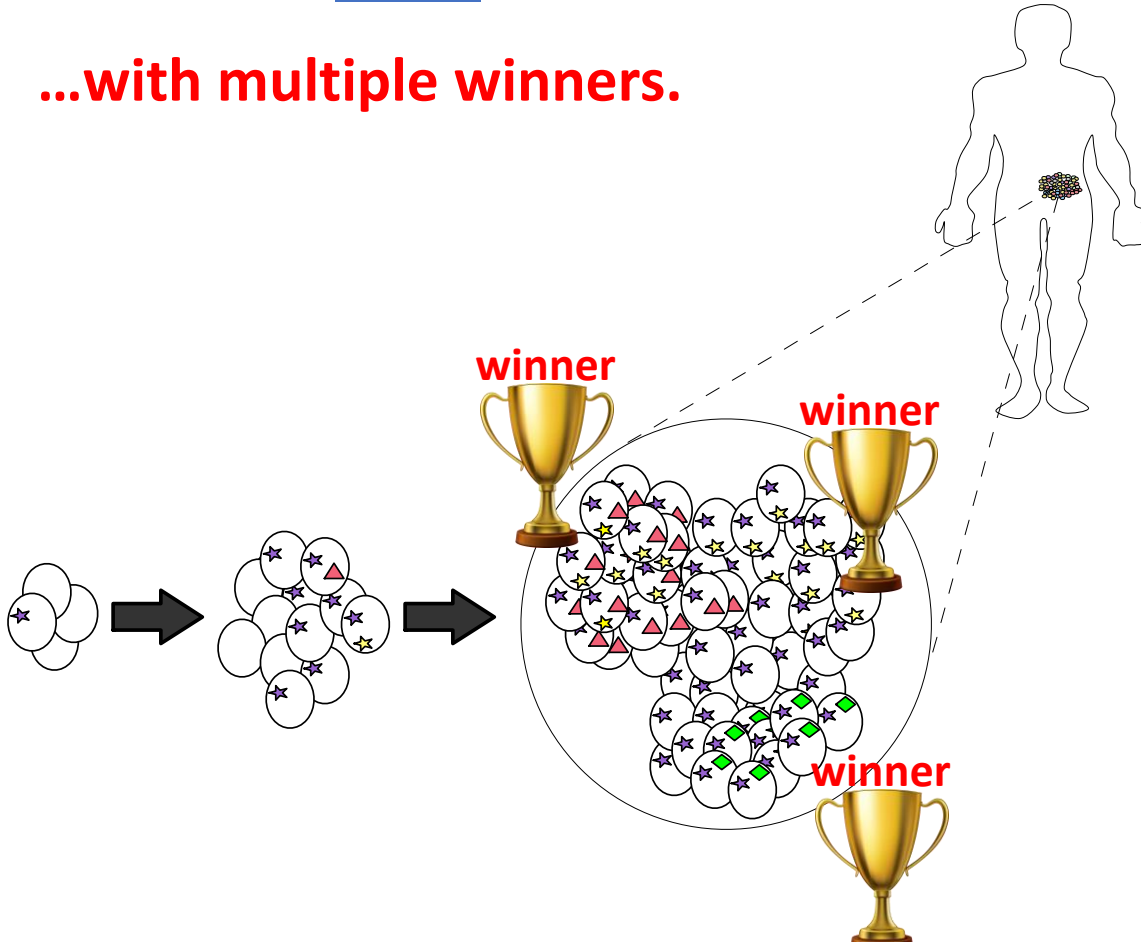


# Cancer is a game of cell evolution



# Cancer is a game of cell evolution

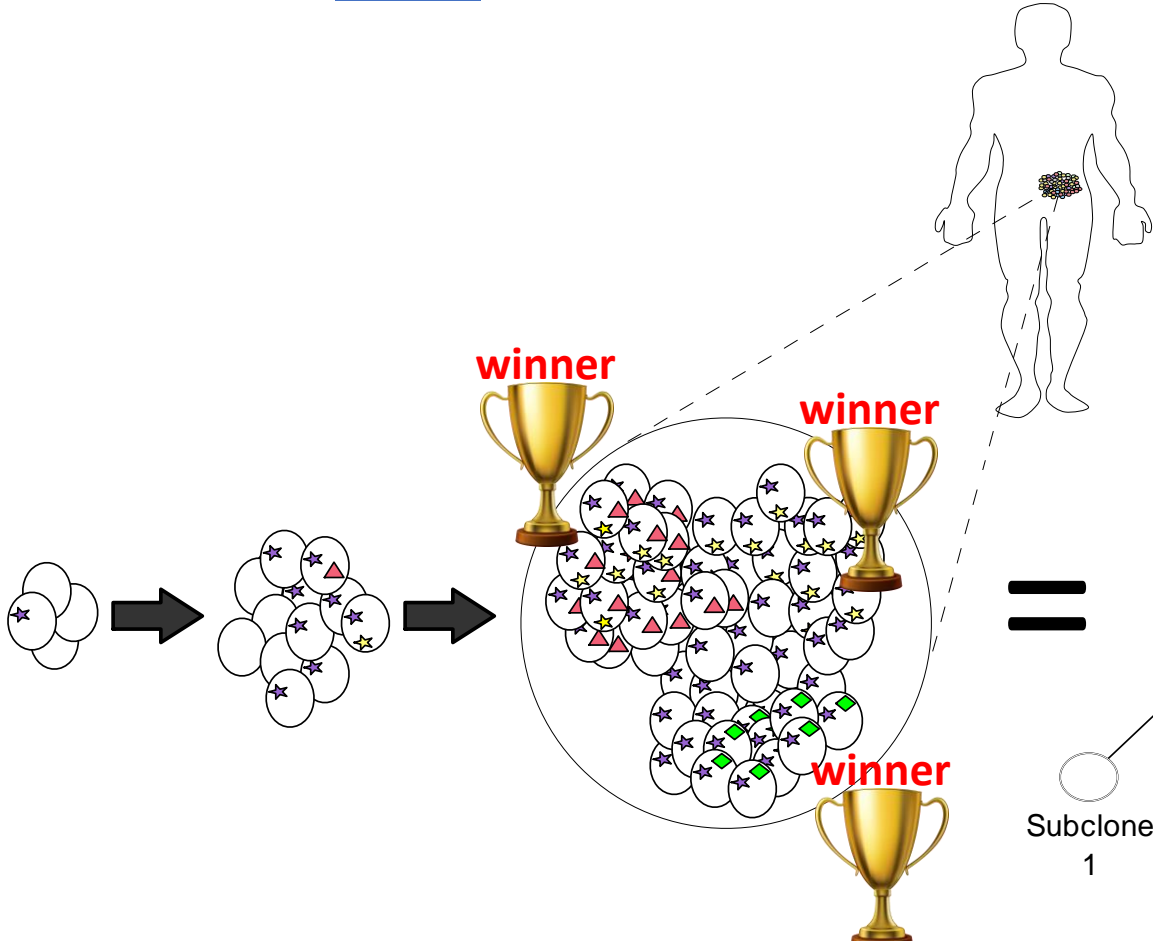
...with multiple winners.



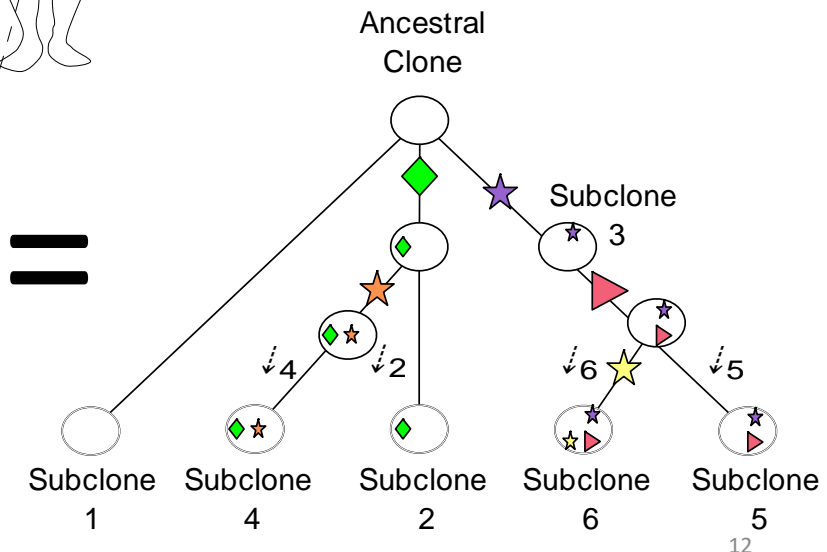
The **clones** found in a patient tumor have **already won** the game of cancer.

We want to learn what **actions & states** were most influential for each clone within a tumor and across patients.

# Cancer is a game of cell evolution



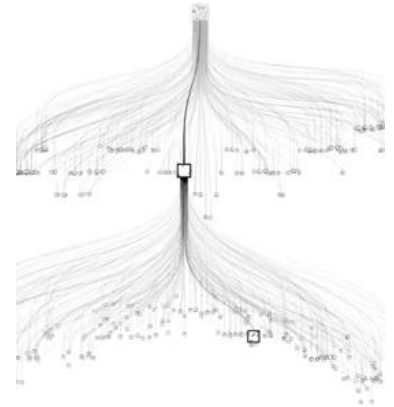
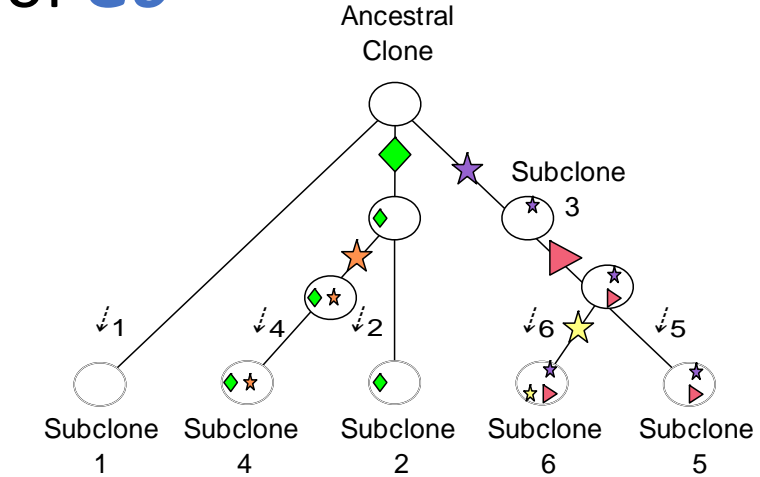
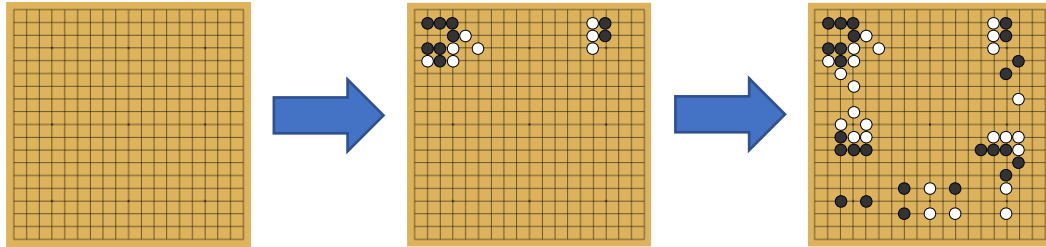
Tumor phylogenetic tree gives us the *mutational history* of each clone



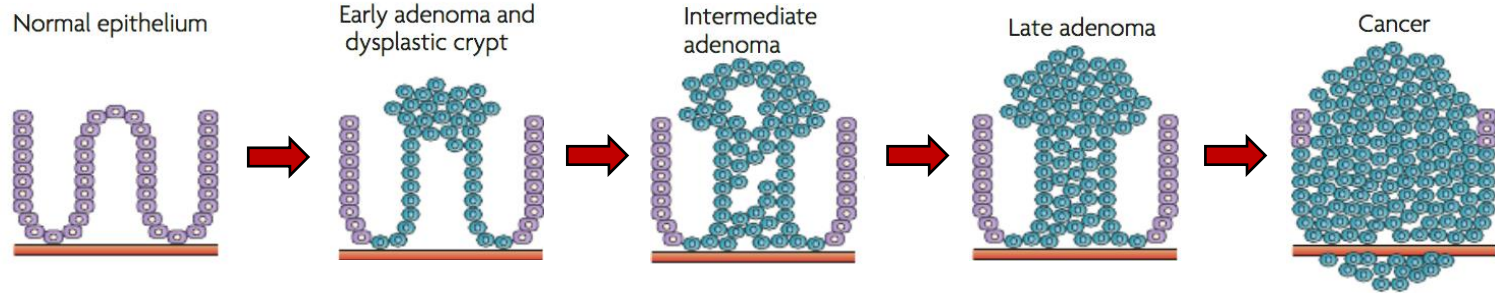
# Cancer is a more complex version of *Go*

**Similarities:** Combinatorial problem with many possible moves, board configurations, and strategies to win.

**Differences:** *state-space*, *action-space* and # *players* in cancer are evolving concurrently!



# COLORECTAL CANCER EVOLUTION

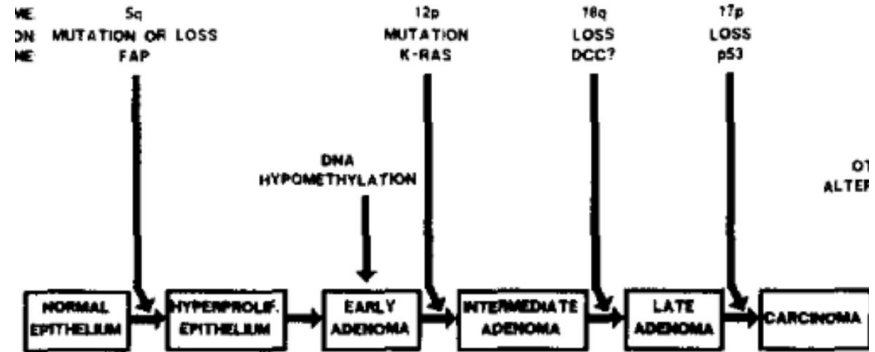


Cell, Vol. 61, 759-767, June 1, 1990, Copyright © 1990 by Cell Press

## A Genetic Model for Colorectal Tumorigenesis

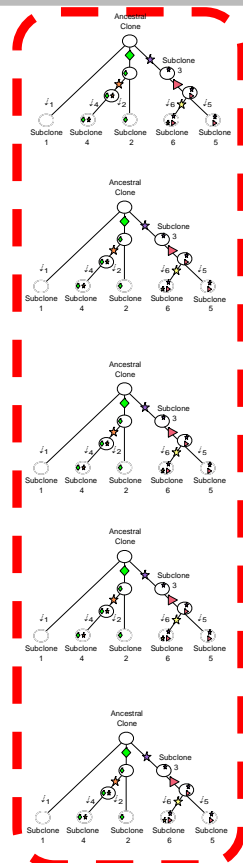
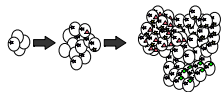
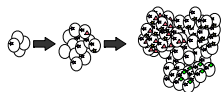
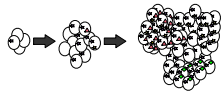
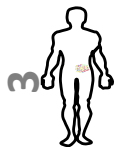
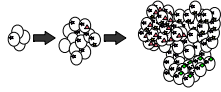
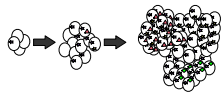
Eric R. Fearon and Bert Vogelstein

June 1, 1990, Copyright © 1990 by Cell Press

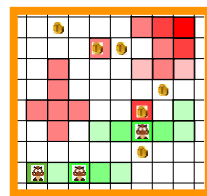


# INFERRING THE REWARD FUNCTIONS OF CANCER

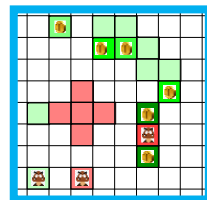
Patient 1  
Patient 2  
Patient 3  
Patient 4  
Patient 5



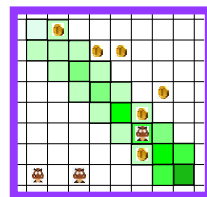
Set of expert demonstrations



Reward function  $R_1$



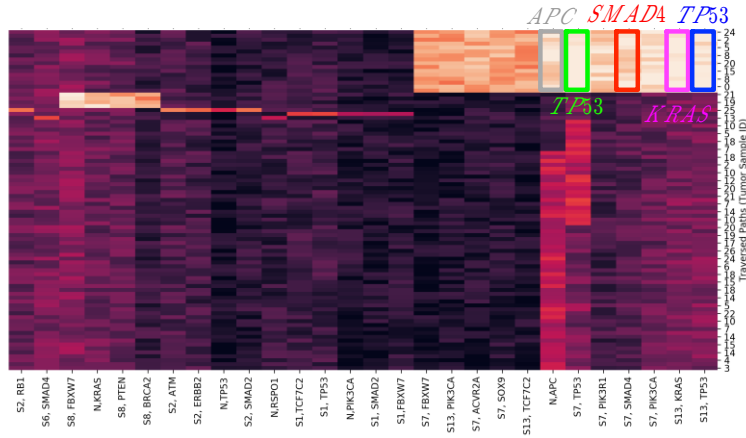
Reward function  $R_2$



Reward function  $R_3$

# PUR-IRL LEARNS LANDMARK MODEL

A



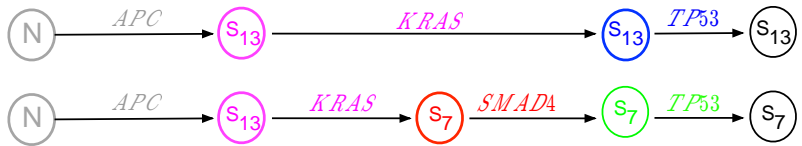
June 1, 1990, Copyright © 1990 by Cell Press

Cell, Vol. 61, 759-767, June 1, 1990, Copyright © 1990 by Cell Press

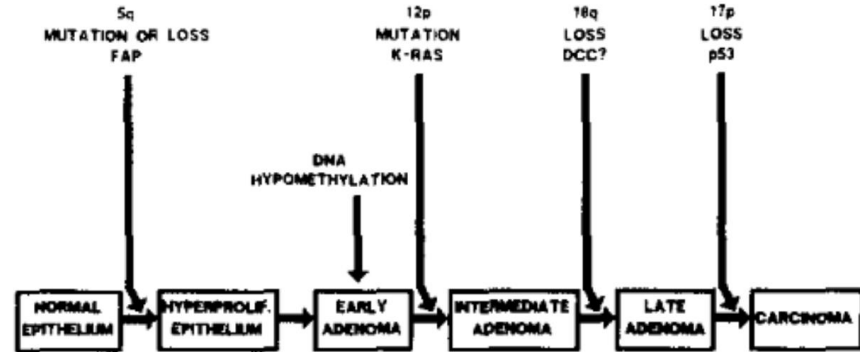
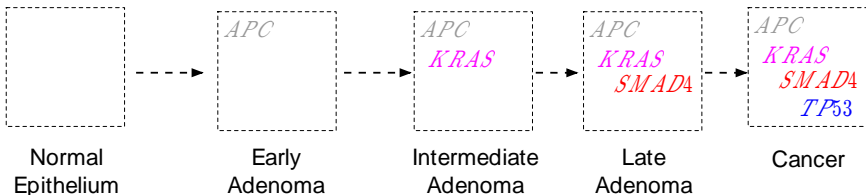
## A Genetic Model for Colorectal Tumorigenesis

Eric R. Fearon and Bert Vogelstein

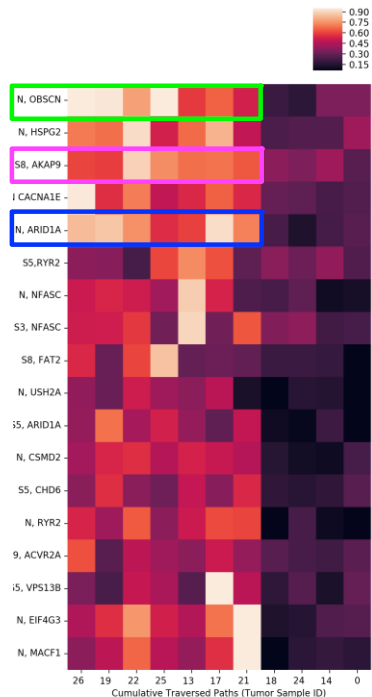
B



C



# PUR-IRL UNCOVERS CRITICAL EARLY MUTATIONS WITHOUT HUMAN SUPERVISION



## Mutation analysis of adenomas and carcinomas of the colon:

### Early and late drivers

Roger K. Wolff<sup>1</sup> | Michael D. Hoffmann<sup>1</sup>  
Lori C. Sakoda<sup>2</sup> | Wade S. Samowitz<sup>3</sup>

Published online 2016 Apr 12. doi: [10.1016/j.bbdis.2016.03.012](https://doi.org/10.1016/j.bbdis.2016.03.012)

AKAP-9 promotes colorectal cancer development by regulating interacting protein 4

Zhi-Yan Hu,<sup>1,2,3</sup> Yan-Ping Liu,<sup>1,2,3</sup> Lin-Ying Xie,<sup>1,2,3</sup> Xiao-Yan Wang,<sup>1,2,3</sup> Fang Yang,<sup>1,2,3</sup> Zu-Guo Li<sup>1,2,3</sup>

Journal of Clinical Oncology

An American Society of Clinical Oncology Journal

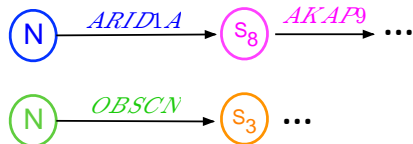
The characteristics of ARID1A mutations in colorectal cancer.

Amir Mehrvarz Sarshekeh, Jonathan M. Loree, Ganiraju C. Manyam, Allan Andresson Lima

We used whole exome sequence data from triplet samples (colon carcinoma, colon adenoma, and normal tissue) from 18 individuals to assess gene mutation rates. Of the 2,204 genes that were mutated, APC, TTN, TP53, KRAS, **OBSCN**, SOX9, PCDH17, SIGLEC10, MYH6, and BRD9 were consistent with genes being an **early driver of carcinogenesis**, in that they were mutated in multiple adenomas and multiple carcinoma.

...we have demonstrated that **AKAP-9** plays a **critical role** in the proliferation, migration and invasion of CRC in vitro as well as the tumorigenesis in vivo. Importantly, we found that AKAP-9 interacts with cdc42 interacting protein 4 (CIP4) and modulates its expression in CRC cells. Moreover, AKAP-9 appears to mediate TGF- $\beta$ -induced epithelial-mesenchymal transition (EMT) via CIP4. Collectively, our study has provided a novel mechanism by which **AKAP-9** regulates CRC tumorigenesis and metastasis.

**ARID1A** mutations were more common in **early stages**. ... This is the largest study evaluating ARID1A mutations in CRC. The majority of mutations appear to be truncating and clonal, suggesting that they have functional significance. ARID1A-mutated tumors demonstrate enrichment of wild-type TP53 but they are more likely to have MSI-H, PIK3CA and BRAF mutations.

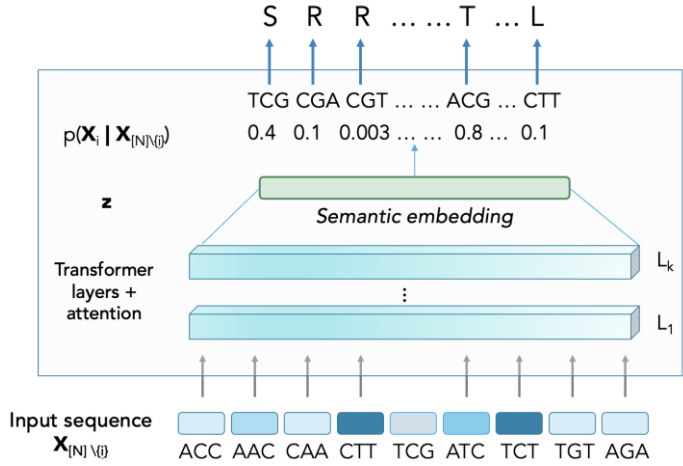


# LLMS for BioSequences

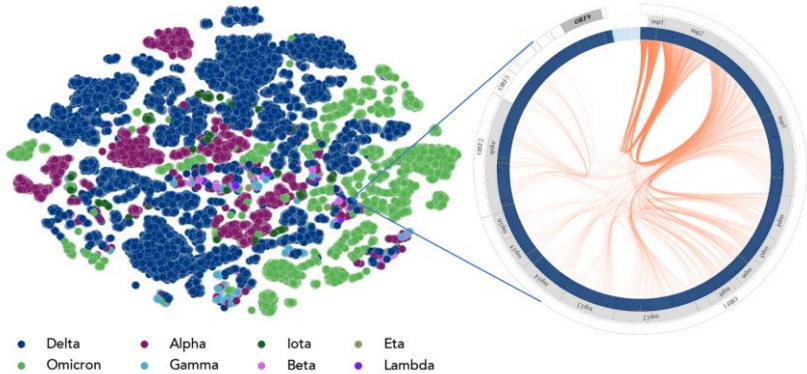
Model Name (Year)	# Params	Tasks   Performance
LOGO (2023)	1M	Encodes variant/site as input   Promotor regions   Promotor-Enhancer interaction   Variant effect prediction
scBERT (2022)	3M	Marker-agnostic cell subtype classification   F1-score = 0.691(SOTA 0.659) Identification of novel cell types   Accuracy = 0.329 (SOTA 0.174)
ProteinBERT(2022)	16M	Uses GO Annotations as input Remote homology   Accuracy 0.22 Protein secondary structure prediction   Accuracy 0.74 Protein stability   Accuracy 0.76
TAPE Transformer(2019)	38M	Remote homology   Accuracy 0.21 Contact residue   Accuracy 0.36 Protein secondary structure prediction   Accuracy 0.73 Protein stability   Accuracy 0.73
INTERACT (2022)	40M	mQTL prediction for schizophrenia   AUROC 0.86. Polygenic risk prediction   R <sup>2</sup> improvement 5-27%
DNABERT (2020)	110M	Promotor site prediction   F1 0.965 (SOTA 0.917) TF binding site prediction   F1 SOTA, AUPRC 0.1495 Splice site prediction   F1 0.91 (SOTA 0.83) Functional variants   F1 0.84
Epigenetic BERT (2021)	110M	TF binding site prediction   AUPRC 0.5405
ENFORMER (2020)	244M	200kb prediction window (vs Expecto/Basenji2 20kb) Gene expression prediction   Correlation: 0.85 Non-coding variant effect prediction   SOTA Variant effect prediction (eQTLs)   <AUROC>: 0.747 Cell-specific enhancer-gene pair recall (CRISPRi)   SOTA
ESM1b (2021)	670M	Single amino variant effect   Acc: 53% 3D Structure prediction Predict amino acid properties Remote homology   AUC: 0.77 Family identification   Acc: 71% Contact residue   Acc: 49%
ProGen (2023)	1.2B	Design functional proteins   Lysozyme expression/activity 72%/73% "Twilight zone" functionality   88%/31%
ProtTrans/ProtT5-XL-BFD (2021)	3B	NGS taxa classification   Accuracy: 98.2% Binding residues   Acc: 39% Subcellular localization   Accuracy: 86% Single amino variant effect prediction   Acc: 53% 3D structure prediction
ESM2 (2022)	15B	3D Structure prediction   angstrom-level Metagenomic structural evolutionary relationships
GenSLM (2023)	25B	Predict future COVID-19 variants   Accuracy: Within 1 mutation Annotation of function for unknown genes Design Functional proteins (untested) Hierarchical DNA-protein model

# Genome-scale Language Models (GenSLMs)

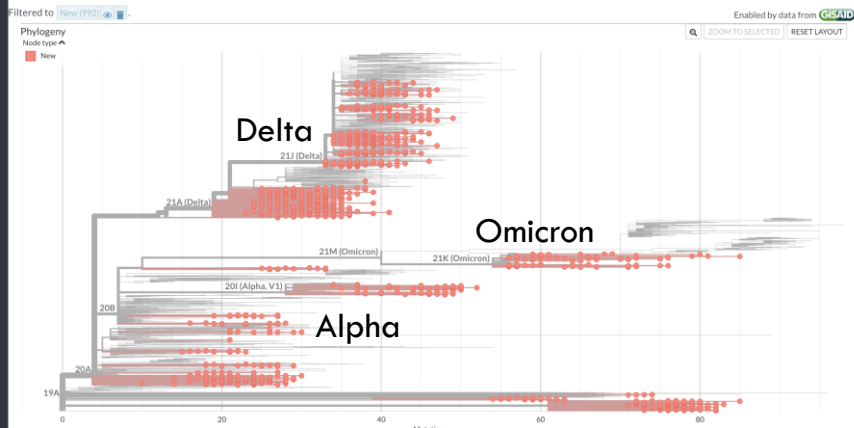
## 1. Pre-train LLMs on ~100M DNA sequences



## 2. Fine-tuning on SARS-CoV-2 reveals evolutionary patterns



## 3. Generate new sequences that look like extant and future SARS-CoV-2 genomes (shown as red dots in the phylogeny)



Slide courtesy of:  
Alexander Brace, Kyle Hippe, Arvind Ramanathan

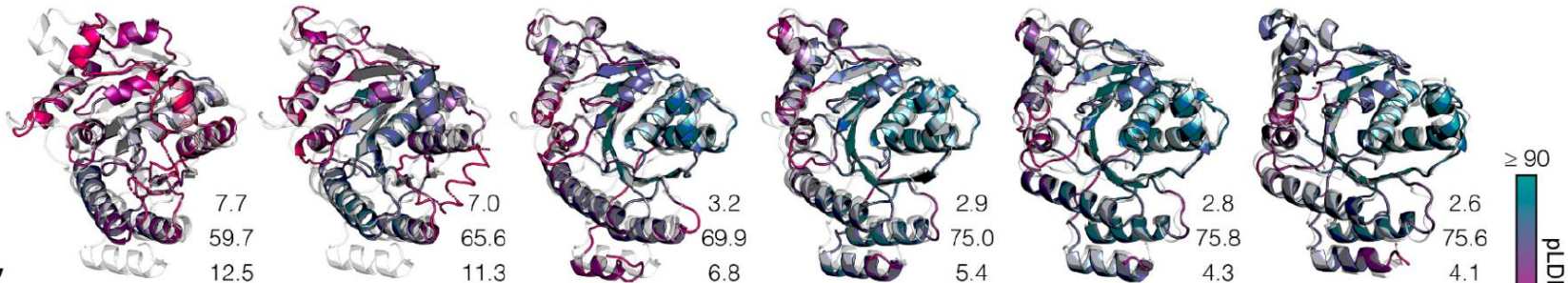
# Protein Models

## ESM

### F

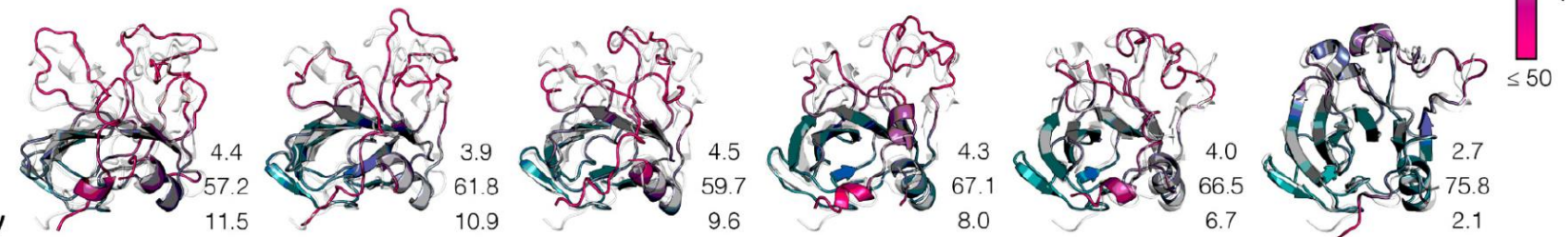
Target:  
7QQA

**RMSD**  
**pIDDT**  
**Perplexity**



Target:  
T1056

**RMSD**  
**pIDDT**  
**Perplexity**



ESM-2 (8M)

ESM-2 (35M)

ESM-2 (150M)

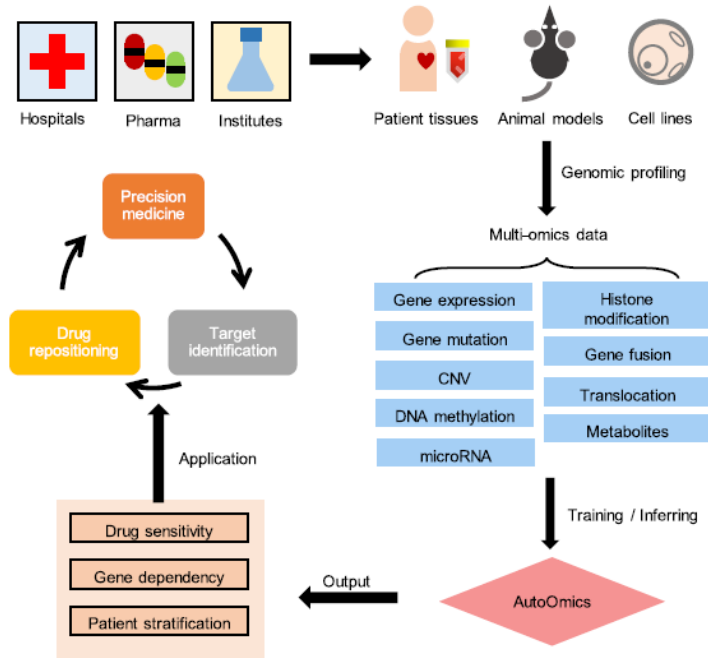
ESM-2 (650M)

ESM-2 (3B)

ESM-2 (15B)

# “AutoOmics: New multimodal approach for multi-omics research”

<https://www.sciencedirect.com/science/article/pii/S266731852100012X>

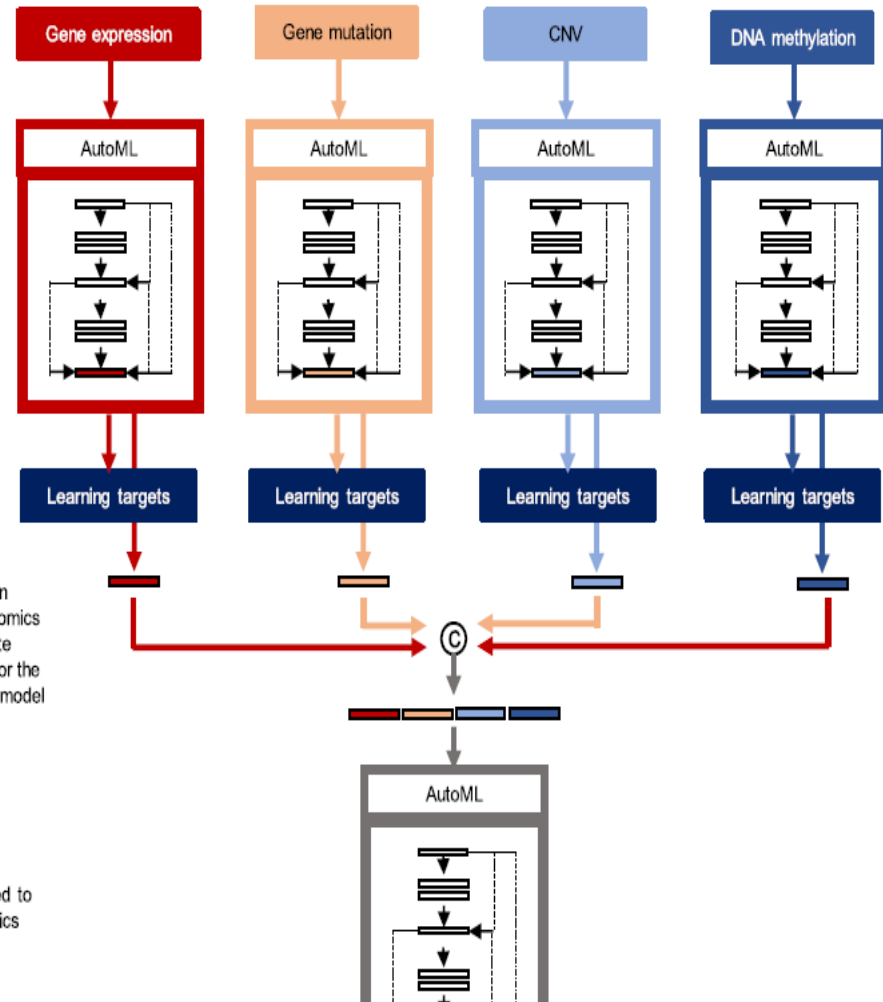


**Step 1:** Omics data

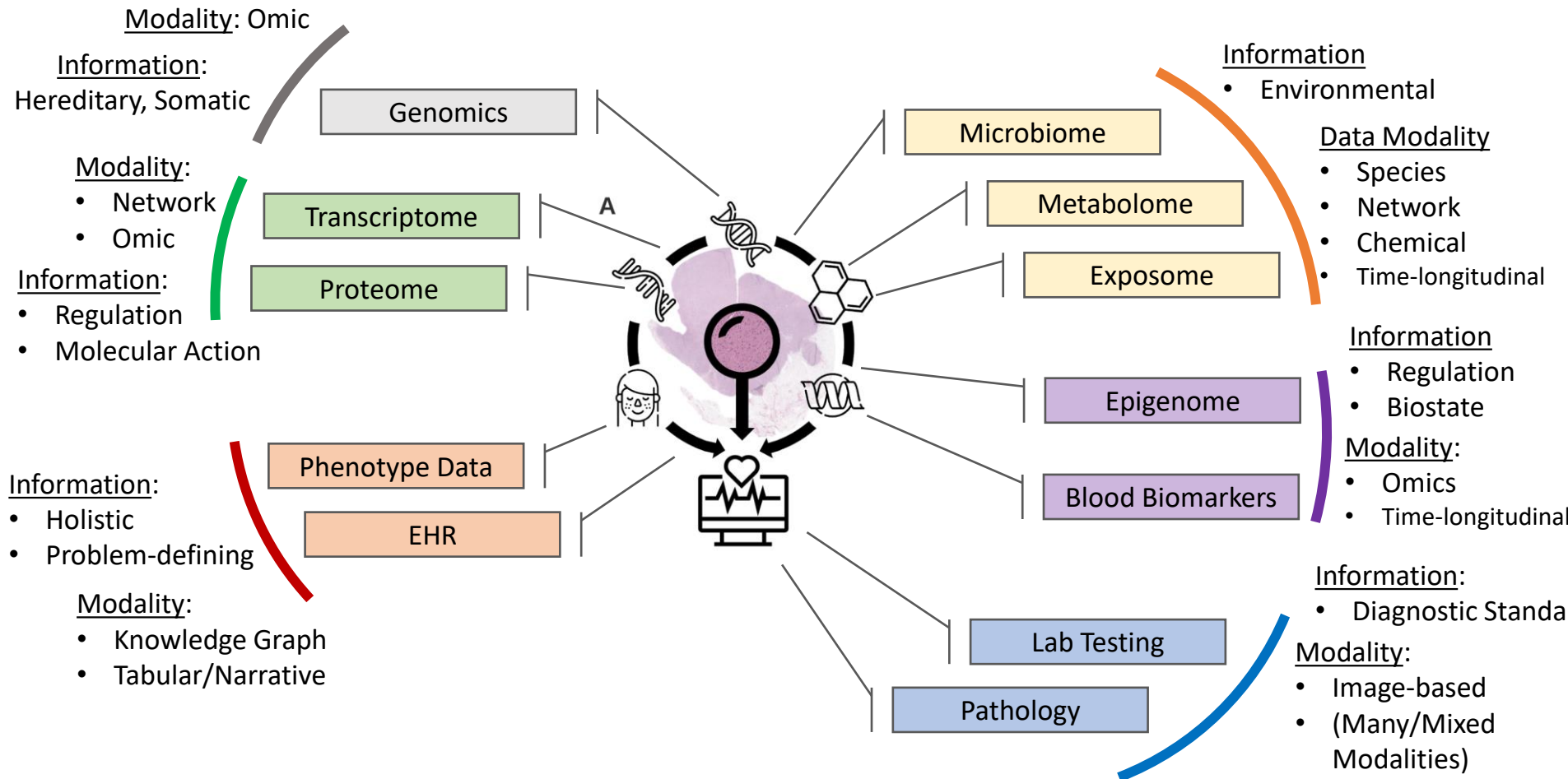
**Step 2:** AutoML is used to train the single-omics DNN models separately

**Step 3:** The last hidden layers from the single-omics models are concatenate together as the input for the final multi-omics DNN model

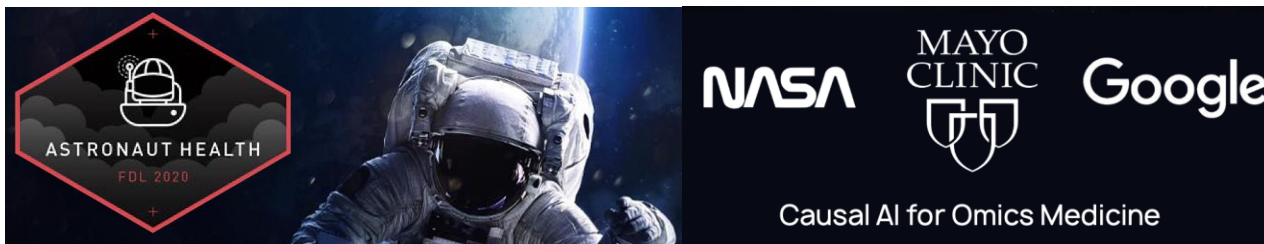
**Step 4:** AutoML is used to train the final multi-omics DNN model



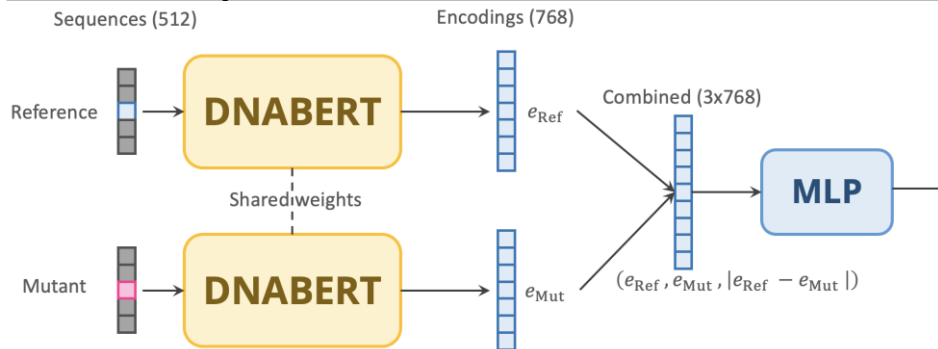
# Multi-modal Learning for Multiple Omics and Clinical Data Modalities



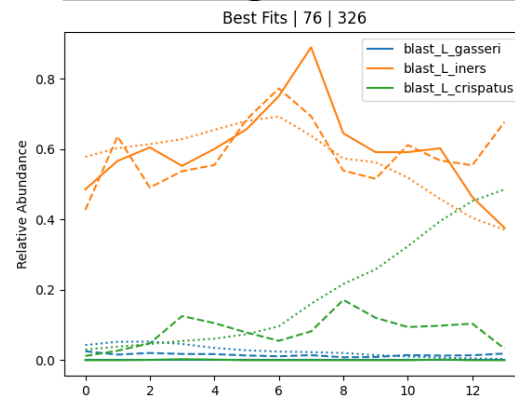
# CANCER EVOLUTION: ONGOING WORK



## Feature Representation – “Model-free” Action Space



## Time Longitudinal Modeling

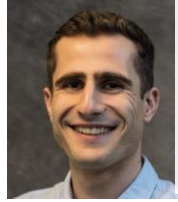




**Heidi  
Nelson**



**John  
Kalantari  
(Yrikka  
founder)**



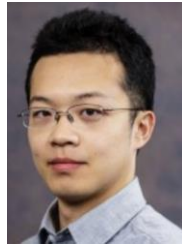
**Kia  
Khezeli  
(Yrikka  
founder)**



**Lisa  
Boardman**



**Mohammed  
El-Kebir**



**Chuanyi  
Zhang (now  
Regeneron)**

Funding, with Thanks

- DeWitt and Curtiss Family Foundation
- NCI R01-CA179243
- Frontier Development Lab
- Center for Individualized Med
- Department of Energy

**Rick  
Stevens**



**Tom  
Brettin**



**Azton  
Wells**



**Brian  
Hsu**



**Arvind  
Ramanathan**



**Alex  
Brace**



**Kyle  
Hippe**



**Gautham  
Dharuman**

