Tumor Somatic Immune Phenotype Prediction Driven by DNA Sequence and Deep Learning

Tianrui Qi, Jing He

Oncology Group
Therapeutic Area Genetics (TAG)
Regeneron Genetics Center (RGC)

Aug 16, 2024

Background

Significance:

- **What?** Describe the presence and activity of immune cells such as T cells and macrophages within the tumor tissue, i.e., immune response.
- **How?** Assessed using techniques like immunohistochemistry, flow cytometry, and genomic sequencing methods such as RNA sequencing to quantify immune cells in the tumor.
- **Why?** Predict a patient's response to immunotherapy; higher infiltration indicating increased sensitivity to treatment, i.e., may achieve better outcomes.

Opportunity 1: Data

- Stanford data (#patients = 8, #samples = 23, normal and tumor):
  Basal Cell Carcinoma pre and post anti-PD1 treatment that enhanced tumor immune infiltration.
- TCGA SKCM data (#patients = 470, #samples = 928, normal blood and tumor):
  Leukocyte Fraction from other omics data.
- RGC data (Germline WXS from blood)

Opportunity 2: Method

- DNABERT (2021): State-of-art performance on prediction of promoters, splice sites and transcription factor binding sites (human).
- DNABERT2 (2023): State-of-art performance across 28 distinct datasets across 7 tasks and 4 species (human, mouse, yeast, virus).

Species	Task	Num. Datasets	Num. Classes	Sequence Length
Human	Core Promoter Detection	3	2	70
	Transcription Factor Prediction	5	2	100
	Promoter Detection	3	2	300
	Splice Site Detection	1	3	400
Mouse	Transcription Factor Prediction	5	2	100
Yeast	Epigenetic Marks Prediction	10	2	500
Virus	Covid Variant Classification	1	9	1000

Table 1: Summarization of the Genome Understanding Evaluation (GUE) benchmark.

Background: More About Model

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

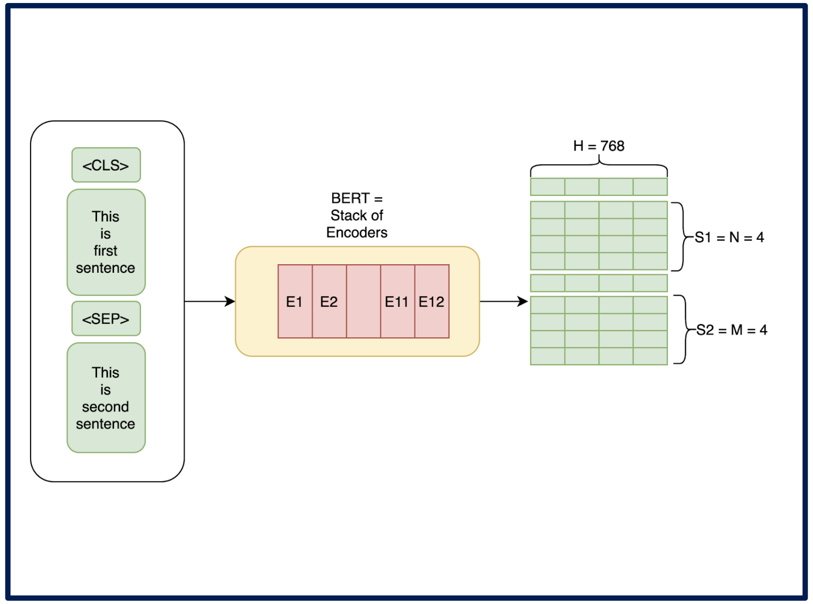
Genome analysis

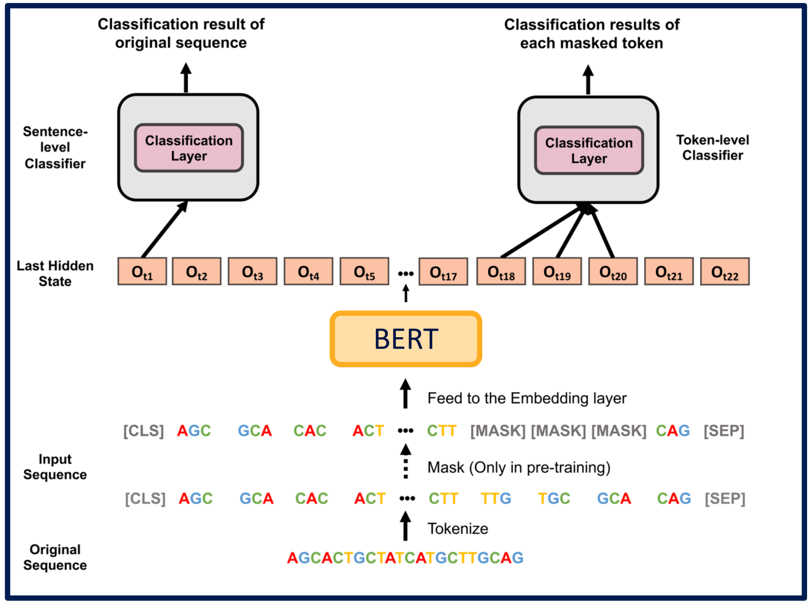
DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome

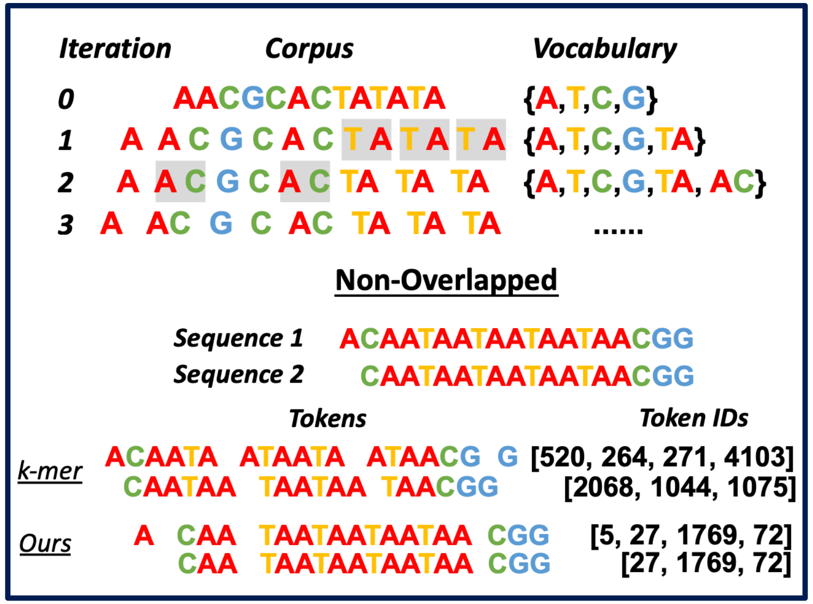
Yanrong Ji[1,†], Zhihan Zhou[2,†], Han Liu[2,*] and Ramana V. Davuluri[3,*]
[1]Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of

DNABERT-2: EFFICIENT FOUNDATION MODEL AND BENCHMARK FOR MULTI-SPECIES GENOME

Zhihan Zhou* Yanrong Ji* Weijian Li* Pratik Dutta† Ramana Davuluri† Han Liu*
Northwestern University* Stony Brook University†





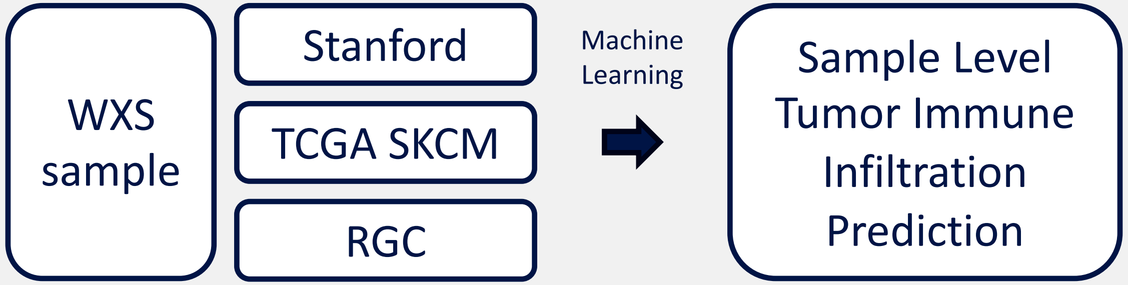


- Encoder-only models that utilize global context instead of decoder-only models like GPT that are unidirectionally.
- Useful when final predictions have high accuracy based off only an embedding / representation of the inputted sequence

- k-merization scheme for tokenizing the genome to extend BERT from human languae to DNA sequence

- Improve the tokenizing by compression algorithm Byte Pair Encoding (BPE) that merges frequent pairs of nucleotides instead a specific k-mer
- Biologically significant tokens

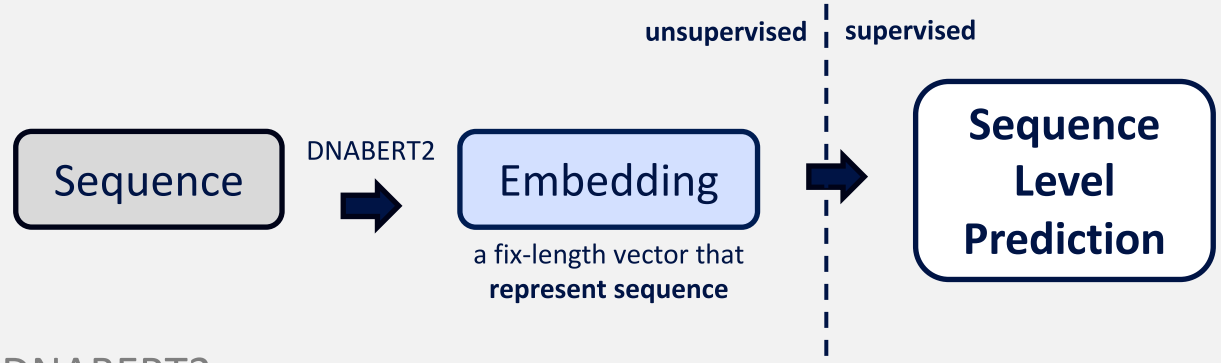
Goals



- Goal 1: Viability of predict tumor immune infiltration by DNA sequence on Stanford data
- Goal 2: Finetune the pipeline on TCGA SKCM data
- Goal 3: Predict on RGC data

Challenges: Requiring Context-specific Adaptation

Problem DNABERT2 Tries to Solve



Species	Task	Num. Datasets	Num. Classes	Sequence Length
Human	Core Promoter Detection	3	2	70
	Transcription Factor Prediction	5	2	100
	Promoter Detection	3	2	300
	Splice Site Detection	1	3	400
Mouse	Transcription Factor Prediction	5	2	100
Yeast	Epigenetic Marks Prediction	10	2	500
Virus	Covid Variant Classification	1	9	1000

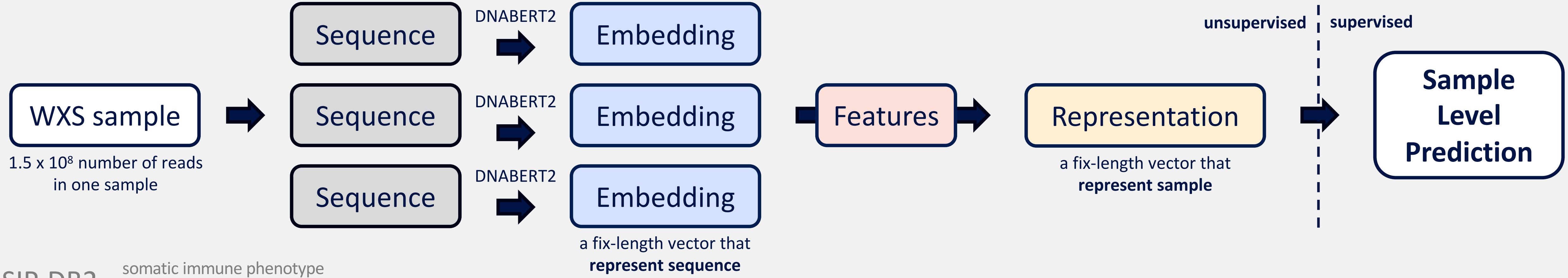
Table 1: Summarization of the Genome Understanding Evaluation (GUE) benchmark.

DNABERT2

Problem We Try to Solve

Gap between sequence and sample level prediction:
1. Number of reads vary by samples, chromosomes
2. Reads between samples not match, cannot compare directly

How we going to solve:
1. Feature select the most informative positions
2. Extract reads from these positions to represent the sample

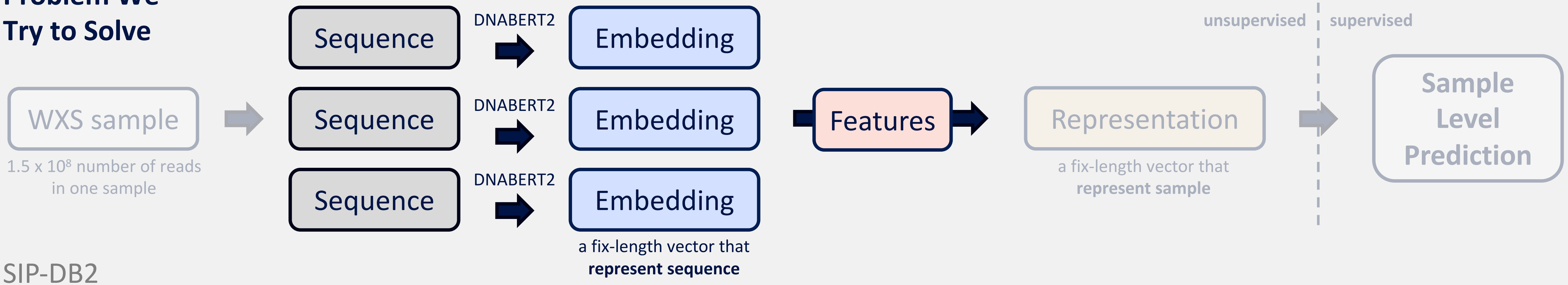


SIP-DB2 somatic immune phenotype prediction based on DNABERT2

RGC

Methods(v1): How to Represent a Given WXS Sample?

Problem We Try to Solve



SIP-DB2

Select Features

c: index of chromosome, 1-22 & X
N_c: num of reads for chromosome c, all samples
K_c: num of features(positions) for chromosome c, i.e., $K_c = 0.1 \times N_c$

	Sequence	Position
1	TCTTG…CACA	51,442
…	…	…
K_c	CTGGA…AAC	2,764,524
…	…	…
N_c	AAGCA…CTCAA	6,463,840

length vary by reads

DNABERT2

	Embedding	Position
1	-0.032, -0.121, …, 0.071	51,442
…	…	…
K_c	-0.137, -0.090, …, 0.107	2,764,524
…	…	…
N_c	0.022, -0.076, …, -0.085	6,463,840

fix-length vector of 768

max of cross reads distance

	Distance	Position
1	9.7	51,442
…	…	…
K_c	6.5	2,764,524
…	…	…
N_c	12.1	6,463,840

sort by distance

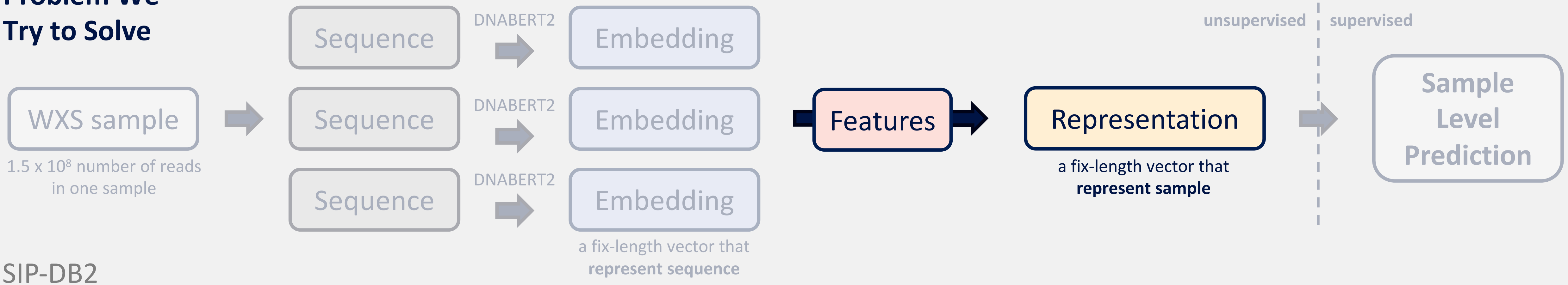
keep top K_c positions

	Distance	Position
1	12.1	6,463,840
	…	…
K_c	9.7	51,442

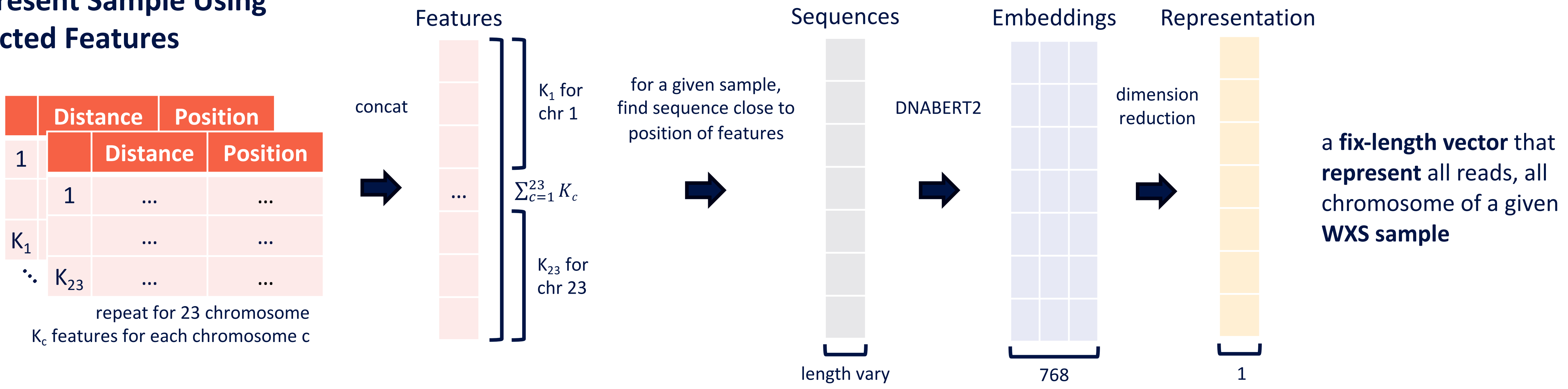
we have K_c number of most informative position of chromosome c

Methods(v1): How to Represent a Given WXS Sample?

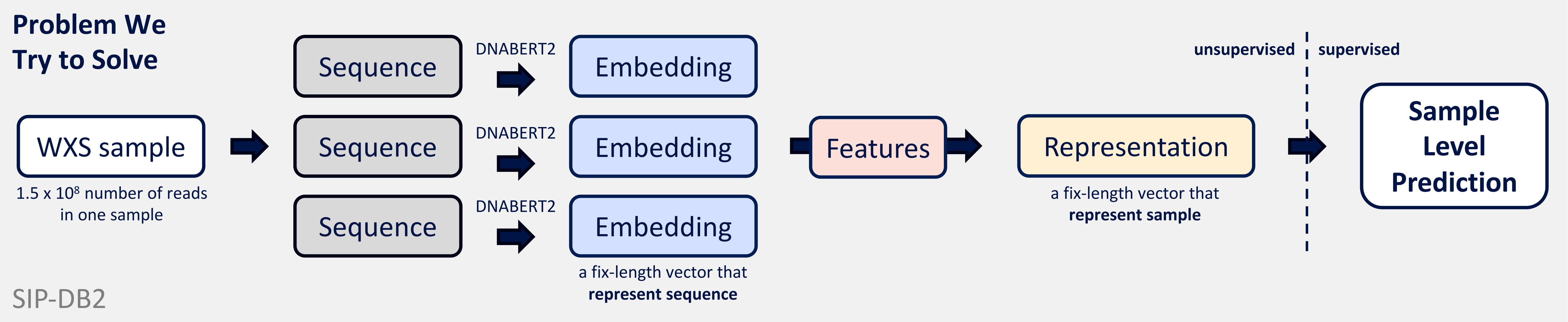
Problem We Try to Solve



Represent Sample Using Selected Features



Methods(v1): Profiling of Each Steps



Filter Sequence using SNPs

Resource

- CPU: 1 core / process
- Memory: 45GB / process; depend on length (bp) of WXS, which is fixed
- Disk: 0.2GB / 1M reads; 30GB / sample with 150M reads

Speed

- 30K reads / second; 1.5 hour / sample with 150M reads

Sequence to Embedding

Resource

- CPU: 1 core / process
- GPU (T4): at least 2GB / process; depend on batch size
- Memory: 3GB / 1M reads; depend on number of reads and save to disk frequency
- Disk: 3GB / 1M reads; 4.5GB / sample with 1.5M reads after SNPs filter

Speed

- 500 reads / second; 1 hour / sample with 1.5M reads after SNPs filter

Feature Selection

Resource

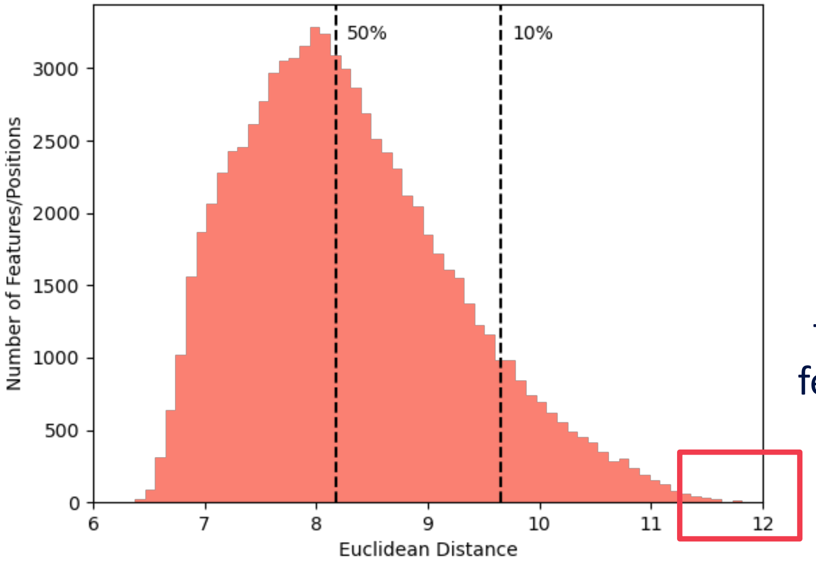
- CPU: 5 core / process
- Memory: 2GB / sample pair where each sample with 1.5M reads after SNPs filter
- Disk: 0.5MB / 1M reads; 300MB if we use 400 samples with 1.5M reads each

Speed

- 1 second / sample pair with 1.5M reads each after SNPs filter; O(sample^2)

Results: Biological Interpretation of the Top Features

Distribution of features' distance of chromosome 6
Feature Selection by method(v1)
Train on Stanford (N=23)



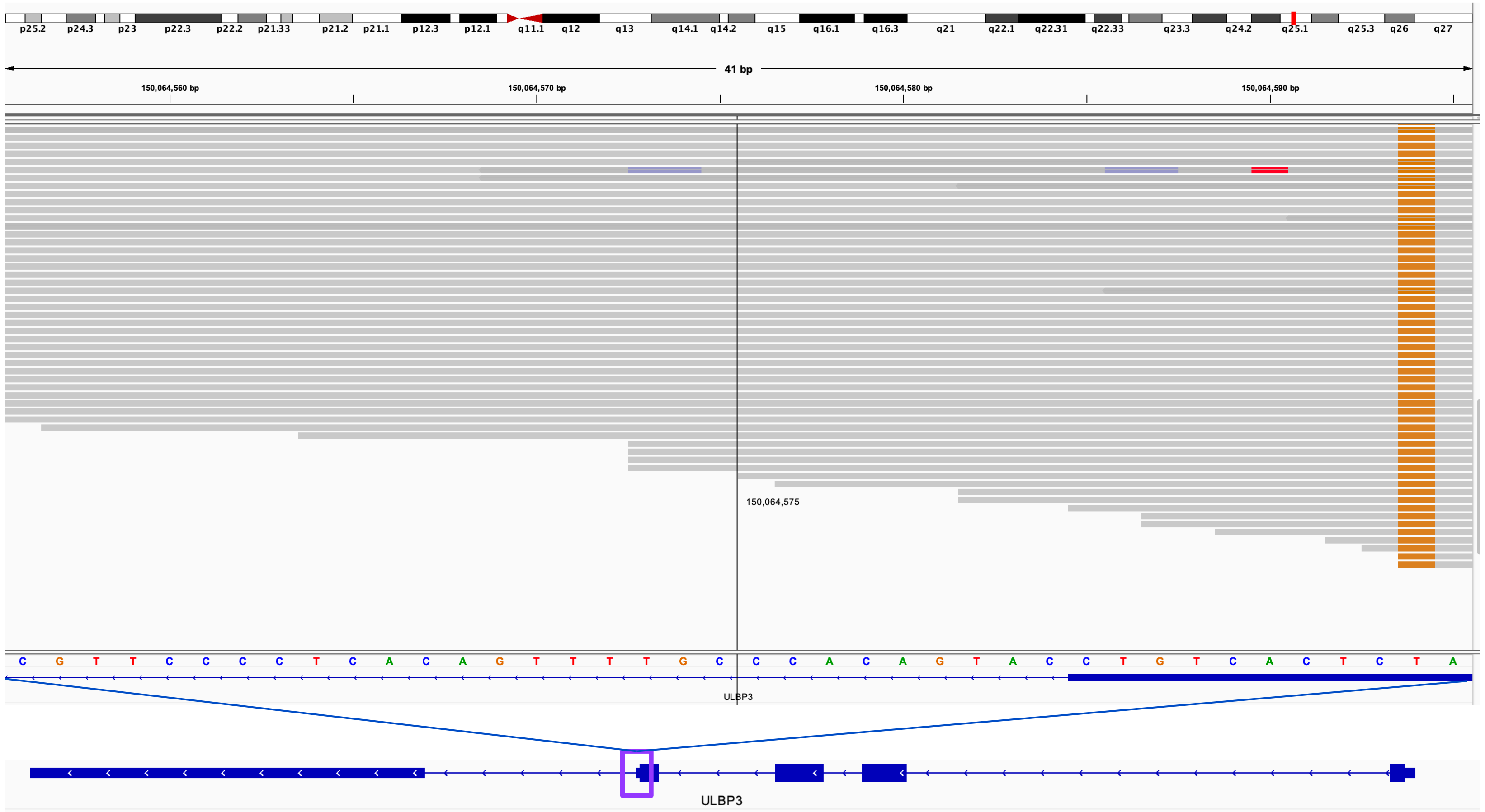
top 10 features

Distance can use as score to indicate how important this feature/position is

Chr	Position	Annotation
6	31148372	HLA-C
6	**150064576**	ULBP3: NKG2D ligand, ULBP3 are expressed on immature and mature dendritic cells and plays an important role in the cytotoxic effect of NK cells against iDC.
6	31271878	HLA-C
6	**90147296**	BACH2: Enables sequence-specific double-stranded DNA binding activity. Involved in primary adaptive immune response involving T cells and B cells.
6	31271900	HLA-C
6	31222832	nearest gene HCG27 and HLA-C. LncRNA HCG27 Promotes Glucose Uptake Ability of HUVECs by MiR-378a-3p/MAPK1 Pathway
6	31586832	LST1, leukocyte specific transcript 1
6	150025312	RAET1L, RAET1L belongs to the RAET1 family of major histocompatibility complex (MHC) class I-related genes, which are located within a 180-kb cluster on chromosome 6q24.2-q25.3.
6	31271884	HLA-C
6	32052816	TNXB, tenascin XB, his protein plays an important role in organizing and maintaining the structure of tissues that support the body's muscles, joints, organs, and skin (connective tissues).

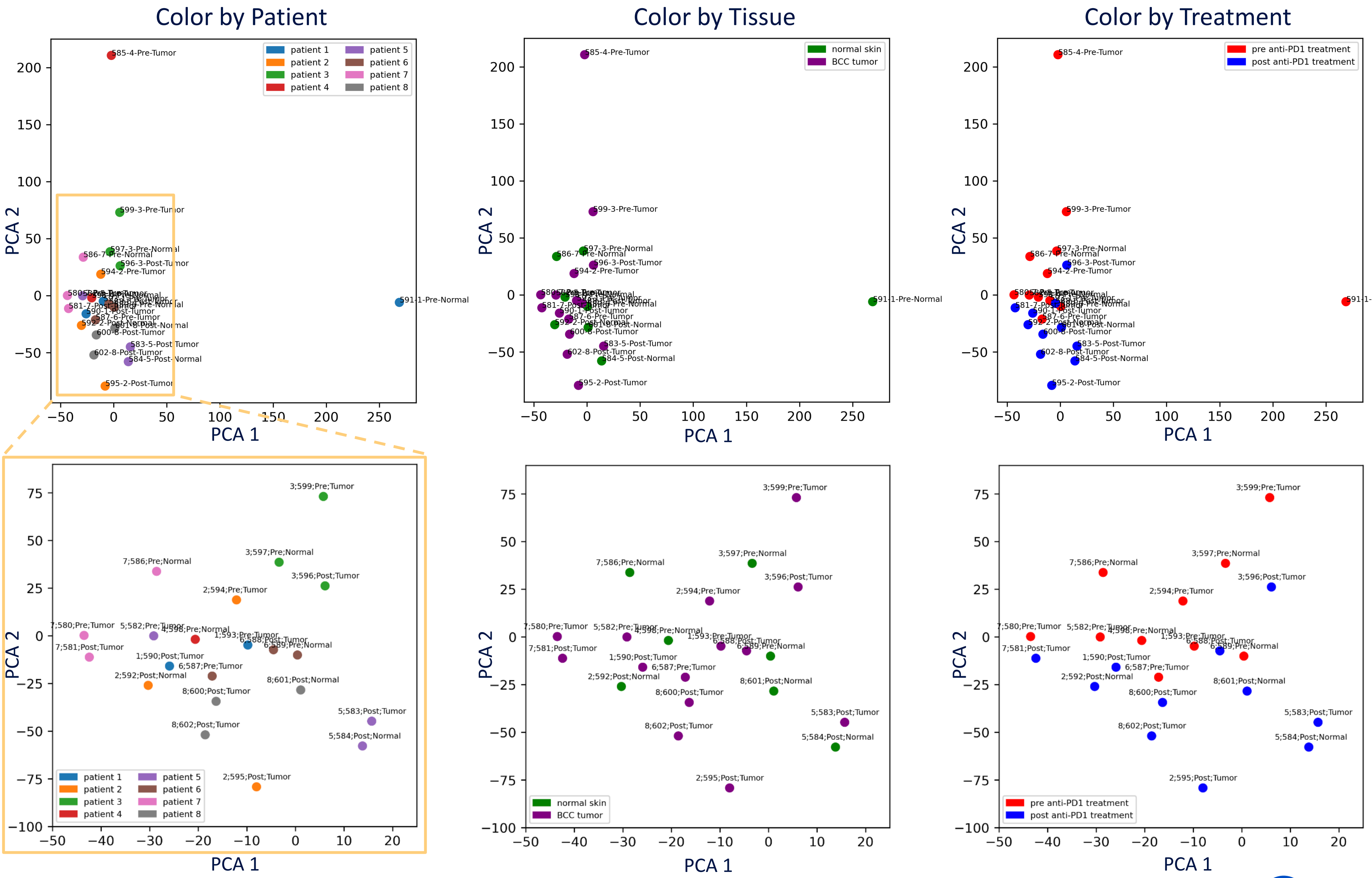
Results: Biological Interpretation of the Top Features

6	150064576	ULBP3: NKG2D ligand, ULBP3 are expressed on immature and mature dendritic cells and plays an important role in the cytotoxic effect of NK cells against iDC.



Results: Anti-PD1 Treated Basal Cell Carcinoma Patients from Stanford

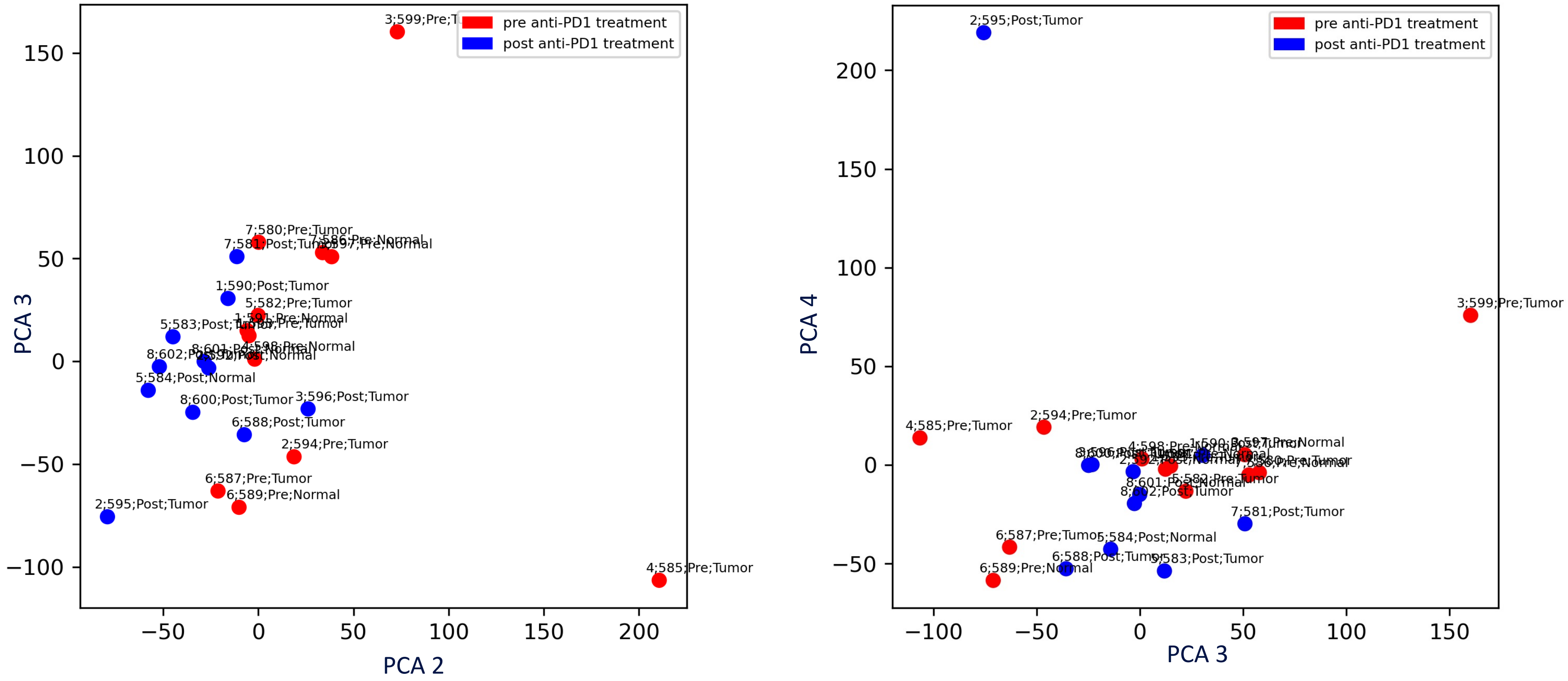
- Features select from 23 chromosome of all Stanford data samples
- Projection of embedding to PCA
- Visualization of metadata on PCA (unsupervised)



Results: Anti-PD1 Treated Basal Cell Carcinoma Patients from Stanford

- Features select from 23 chromosome of all Stanford data samples
- Projection of embedding to PCA
- Visualization of metadata on PCA (unsupervised)

Color by Treatment



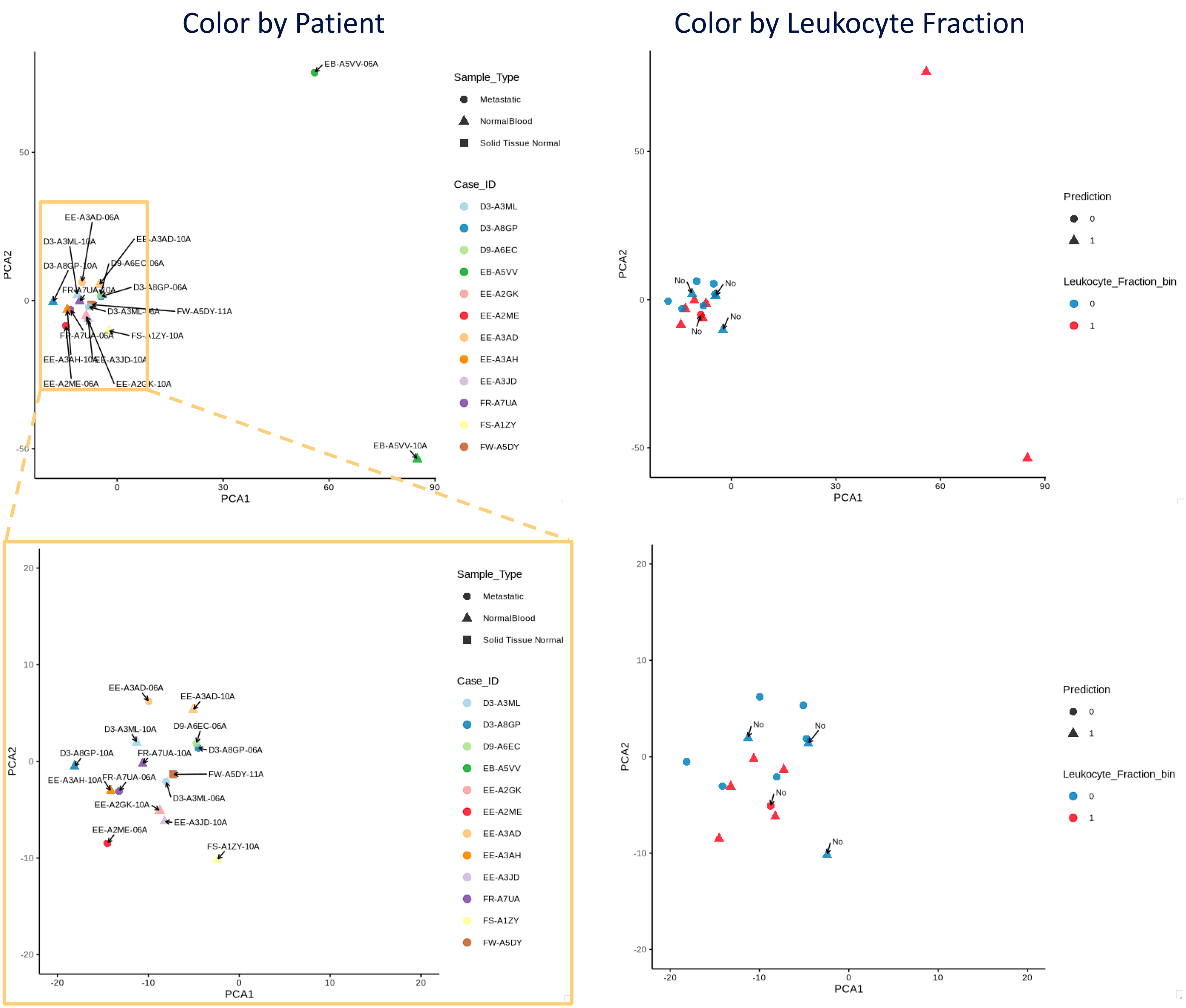
Results: Anti-PD1 Treated Basal Cell Carcinoma Patients from Stanford Supplemental for Stanford Metadata

Patient	Tumor Type	Treatment	Ongoing Vismodegib treatment	Prior treatment	Response	Best % change	pre site	days pre treatment	post site	scRNA days post treatment	Adaptive pre site	Adaptive days pre treatment	Adaptive post site	Adaptive days post treatment	PBMC Adaptive days pre treatment	PBMC Adaptive days post treatment	Exome pre site	Exome days pre treatment	Exome post site	Exome days post treatment
su001	BCC	Pembrolizumab	+	Vismodegib	Yes	-81	L arm	-239, -78	L arm	83	L arm	-78, -5	L arm	83, 146	-239	104	L arm	-78	L arm	146
su002	BCC	Pembrolizumab	-	Vismodegib	Yes	0*	Nose	0	Nose	62	Nose	-602	NA	NA	NA	117	Nose	0	Nose	62
su003	BCC	Pembrolizumab	-	Vismodegib	Yes	-100	R arm	-243	R chest	16, 121	NA	NA	R chest	15	-243	16	R arm	-244	R arm	155
su004	BCC	Cemiplimab	-	-	Yes	-25	Knee	1	Knee	38	NA	NA	NA	NA	NA	NA	Knee	0	NA	NA
su005	BCC	Pembrolizumab	-	Vismodegib	No	5	L ear	0	L ear	105	L ear	-28	L ear	42	0	42	L ear	0	L ear	105
su006	BCC	Pembrolizumab	+	Vismodegib	No	-11	R neck	0	R neck	21	R neck	-280	R neck	140	-84	140	R neck	-79	R neck	21
su007	BCC	Pembrolizumab	-	Vismodegib	No	10	L cheek	-3	L cheek	60	L cheek	-3	L cheek	60	-1855	39	L cheek	-3	L cheek	60
su008	BCC	Pembrolizumab	+	Vismodegib	No	0	R arm	-91	R arm	43	R arm	-43	R arm	42	-98	43	NA	NA	R arm	43

Results: Immune Prediction of Untreated Melanoma Patient from TCGA

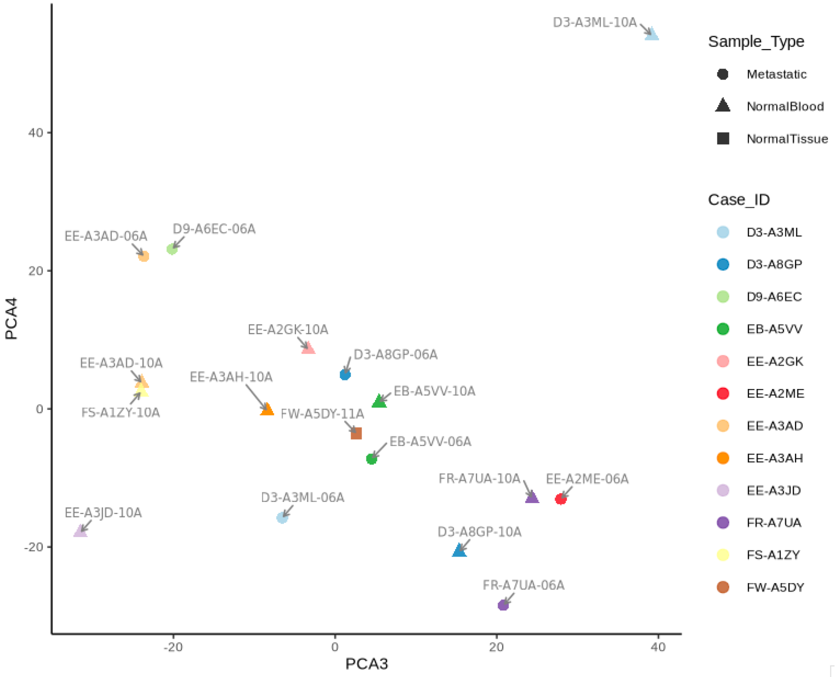
- Features select from 23 chromosome of all Stanford data samples
- Projection of embedding to PCA
- Visualization of metadata on PCA (unsupervised)
- 13/17 samples predicted correctly

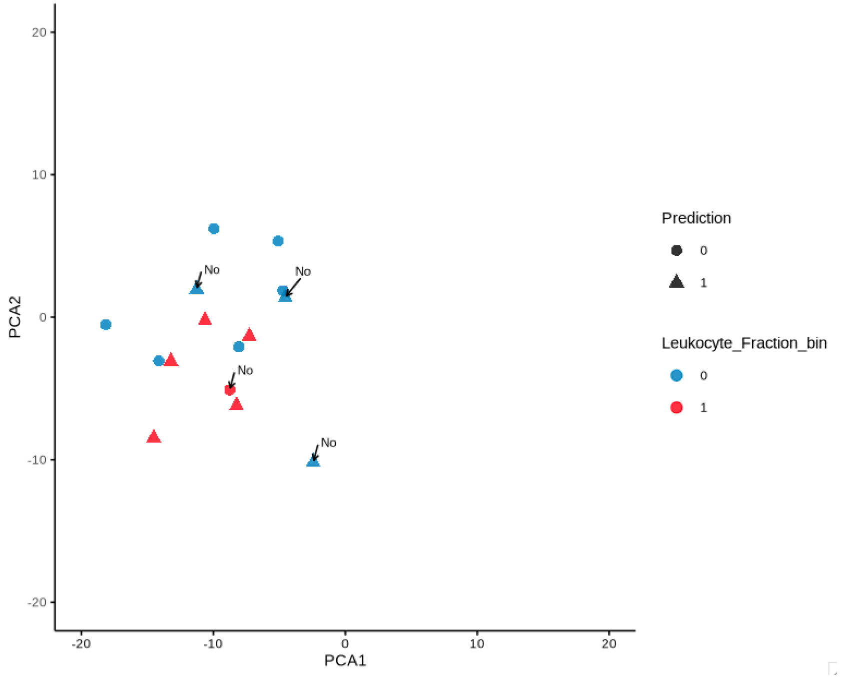
		Ground Truth of Leukocyte Fraction	
		< 0.5	> 0.5
Prediction of Leukocyte Fraction	< 0.5	6	3
	> 0.5	1	7



Results: Immune Prediction of Untreated Melanoma Patient from TCGA

- Features select from 23 chromosome of all Stanford data samples
- Projection of embedding to PCA
- Visualization of metadata on PCA (unsupervised)
- 13/17 samples predicted correctly





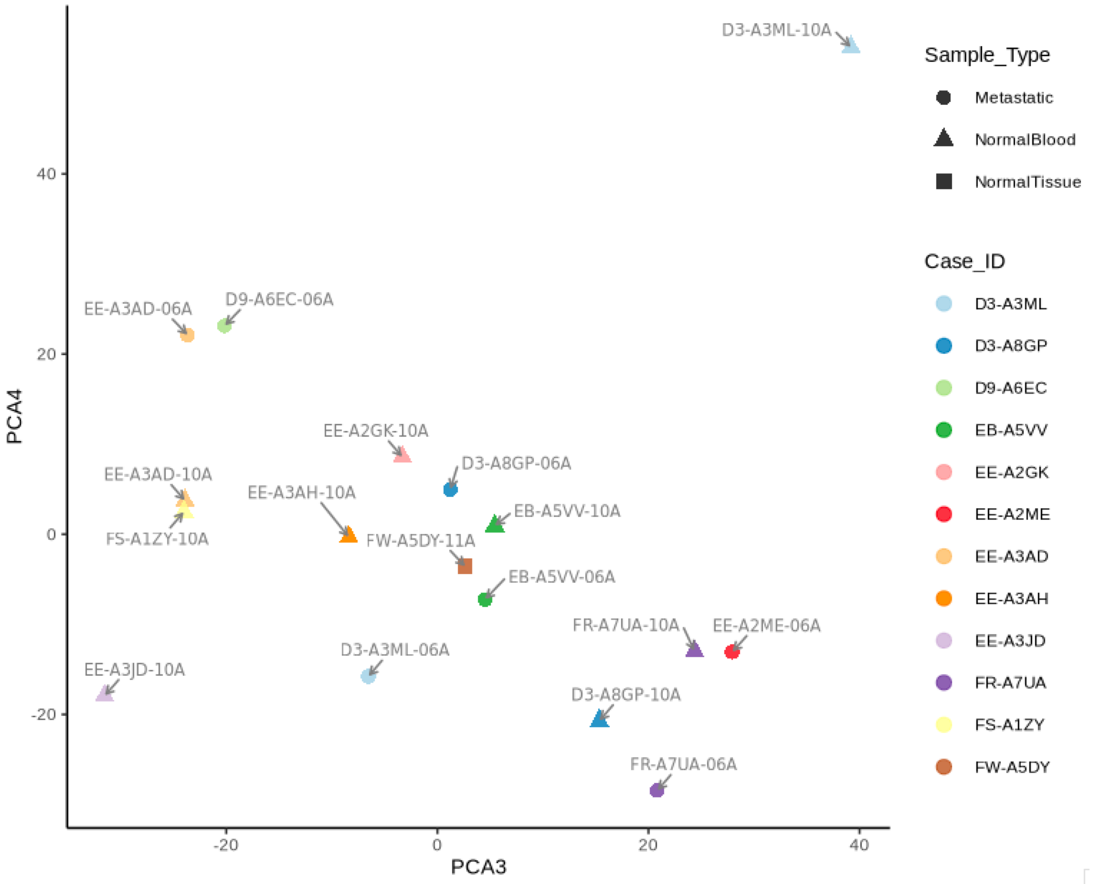
RGC

Results: Immune Prediction of Untreated Melanoma Patient from TCGA Supplemental for TCGA Metadata

TCGA Participant Barcode	TCGA Subtype	DL prediction	DL predict correct	Leukocyte Fraction	Th1 Cells	Th2 Cells	T Cells CD8	Eosinophils
TCGA-EE-A2GK	SKCM.Triple_WT	Low	No	0.90	-769.06	-647.24	0.08	0.00
TCGA-EB-A5VV	NA	High	Yes	0.90	-667.82	-1124.27	0.12	0.00
TCGA-EE-A2ME	SKCM.BRAF_Hotspot_Mı	High	No	0.80	368.91	896.15	0.28	0.00
TCGA-FR-A7UA	NA	High	Yes	0.78	1236.60	279.26	0.41	0.00
TCGA-FW-A5DY	NA	High	Yes	0.76	-79.90	-633.45	0.10	0.00
TCGA-EE-A3JD	SKCM.NF1_Any_Mutants	High	Yes	0.72	798.37	763.81	0.16	0.00
TCGA-D3-A8GP	NA	High	No	0.05	-1108.04	233.51	0.27	0.00
TCGA-D9-A6EC	NA	Low	Yes	0.04	-176.10	336.34	0.04	0.00
TCGA-EE-A3AD	SKCM.BRAF_Hotspot_Mı	Low	Yes	0.04	-567.50	735.12	0.00	0.00
TCGA-EE-A3AH	SKCM.BRAF_Hotspot_Mı	Low	Yes	0.04	-1060.93	389.18	0.08	0.01
TCGA-D3-A3ML	SKCM.RAS_Hotspot_Mut	High	No	0.03	-71.67	677.88	0.20	0.00
TCGA-FS-A1ZY	SKCM.RAS_Hotspot_Mut	High	No	0.02	-1010.42	402.17	0.13	0.00

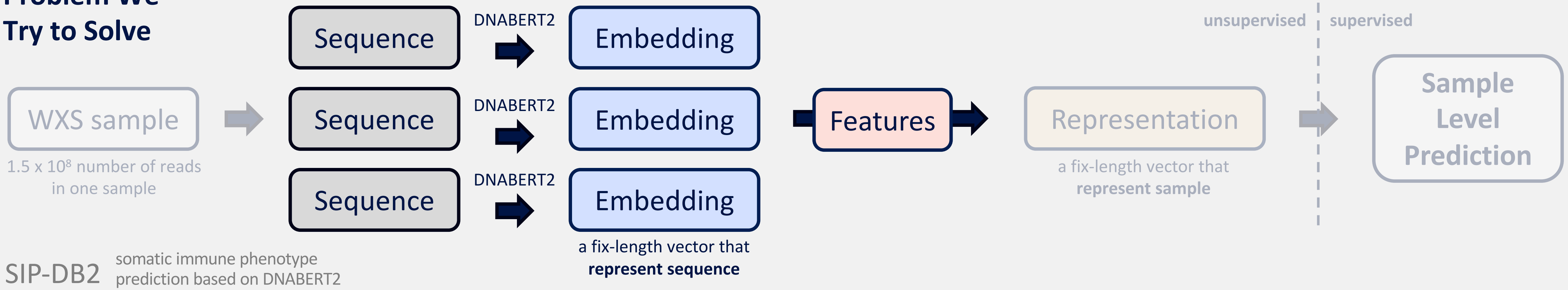
Results: Immune Prediction of Untreated Melanoma Patient from TCGA

- Features select from 23 chromosome of all Stanford data samples
- Projection of embedding to PCA
- Visualization of metadata on PCA (unsupervised)
- 13/17 samples predicted correctly



Challenges: Scaling Up from Stanford (N=23) to TCGA (N=928)

Problem We Try to Solve



Select Features

c: index of chromosome, 1-22 & X
N_c: num of reads for chromosome c, all samples
K_c: num of features(positions) for chromosome c, i.e., $K_c = 0.1 \times N_c$

	Sequence	Position
1	TCTTG…CACA	51442
…	…	…
K_c	CTGGA…AAC	2764524
…	…	…
N_c	AAGCA…CTCAA	6463840

length vary by reads

DNABERT2

	Embedding	Position
1	-0.032, -0.121, …, 0.071	51442
…	…	…
K_c	-0.137, -0.090, …, 0.107	2764524
…	…	…
N_c	0.022, -0.076, …, -0.085	6463840

fix-length vector of 768

Quadratic computing time

max of cross reads distance

	Distance	Position
1	9.7	51442
…	…	…
K_c	6.5	2764524
…	…	…
N_c	12.1	6463840

sort by distance

keep top K_c positions

	Distance	Position
1	12.1	6463840
	…	…
K_c	9.7	51442

we have K_c number of most informative position of chromosome c

RGC

Methods(v2): Bucket Version of Selecting Features

Problem We Try to Solve

WXS sample → Sequence / Sequence / Sequence → (DNABERT2) → Embedding / Embedding / Embedding → Features → Representation → Sample Level Prediction

1.5 x 10^8 number of reads in one sample

a fix-length vector that **represent sequence**

a fix-length vector that **represent sample**

unsupervised | supervised

SIP-DB2

Select Features

	Embedding	Position
1	-0.032, -0.121, ···, 0.071	51,442
···	···	···
K_c	-0.137, -0.090, ..., 0.107	2,764,524
···	···	···
N_c	0.022, -0.076, ···, -0.085	6,463,840

fix-length vector of 768

group by bucket →

Embedding	Position	Bucket
-0.032, -0.121, ···, 0.071 ···	51,400 - 51,499	514
···	···	···
··· -0.137, -0.090, ..., 0.107 ···	2,764,500 - 2,764,599	27,645
···	···	···
··· 0.022, -0.076, ···, -0.085	6,463,800 - 6,463,899	6,463,840

max of cross reads distance →

each bucket calculate independently

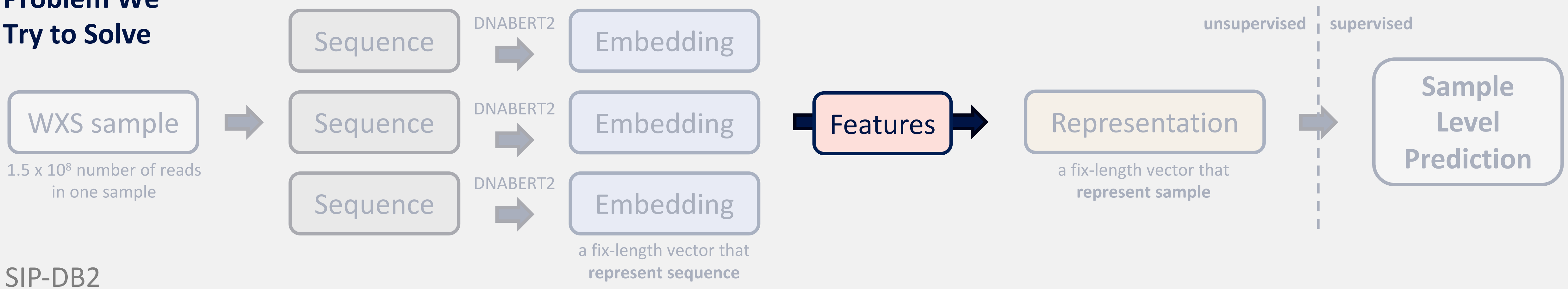
Distance	Bucket
9.7	514
···	···
6.5	27,645
···	···
12.1	6,463,840

select top k buckets with max distance →

RGC

Methods(v2): How Bucket Version Address the Challenges?

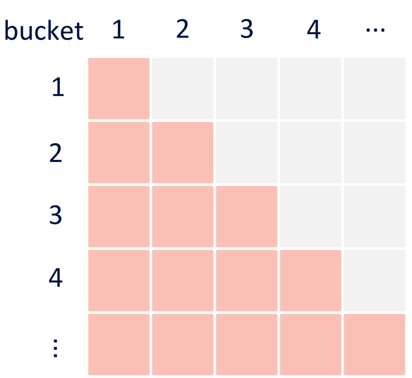
Problem We Try to Solve



SIP-DB2

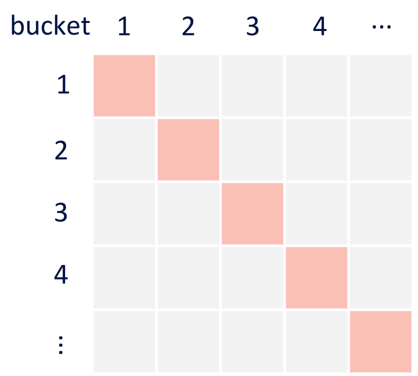
Linear Computing Time

v1

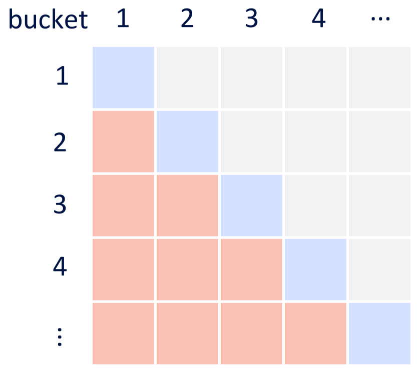


Need to calculate distance between all reads/buckets

v2



Only calculate distance within each buckets



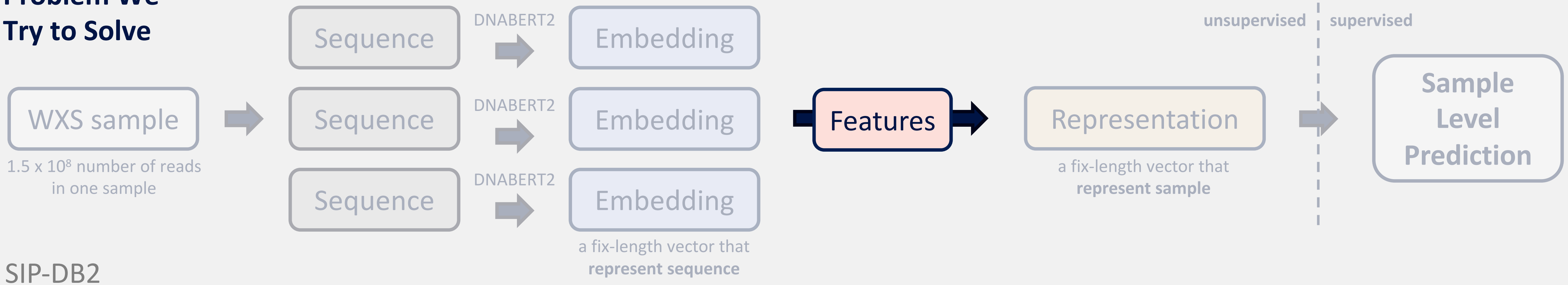
identify key buckets within genes of the same function but vary across samples

identify key buckets that are most dissimilar gene acriss regions

Only calculate distance within buckets still capture information we need

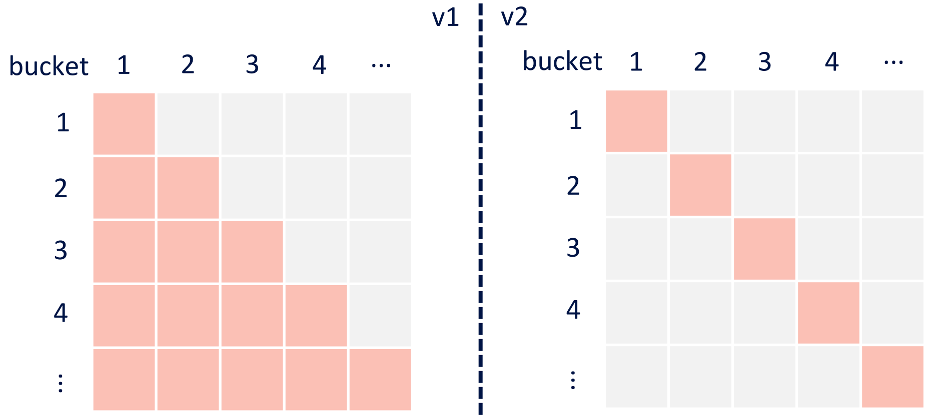
Methods(v2): How Bucket Version Address the Challenges?

Problem We Try to Solve



SIP-DB2

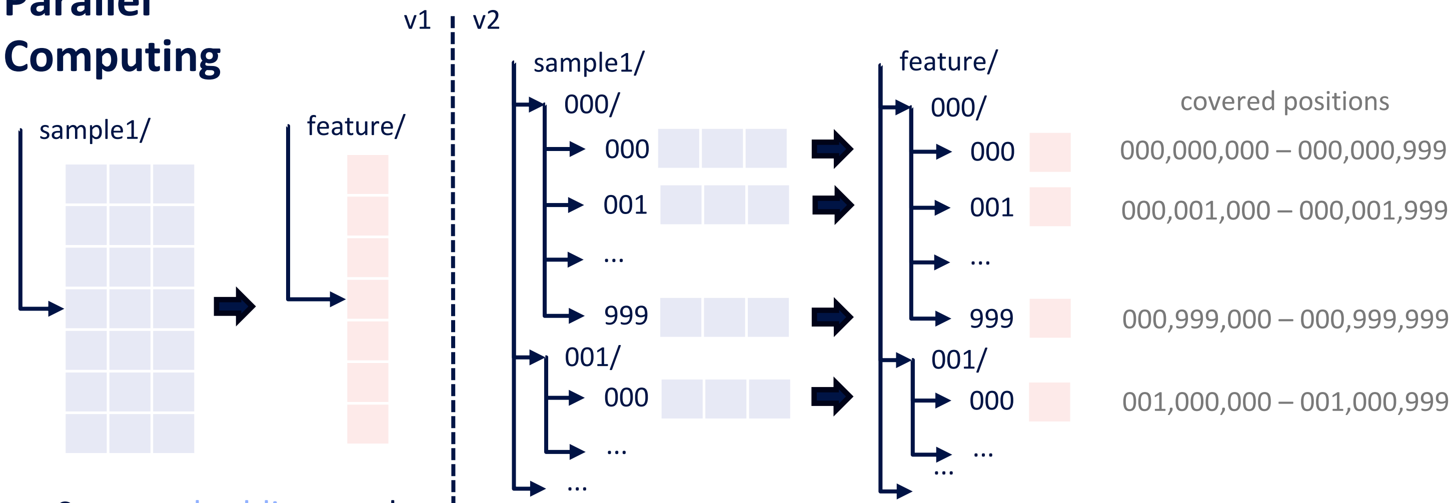
Linear Computing Time



Need to calculate distance between all reads/buckets

Only calculate distance within each buckets

Parallel Computing

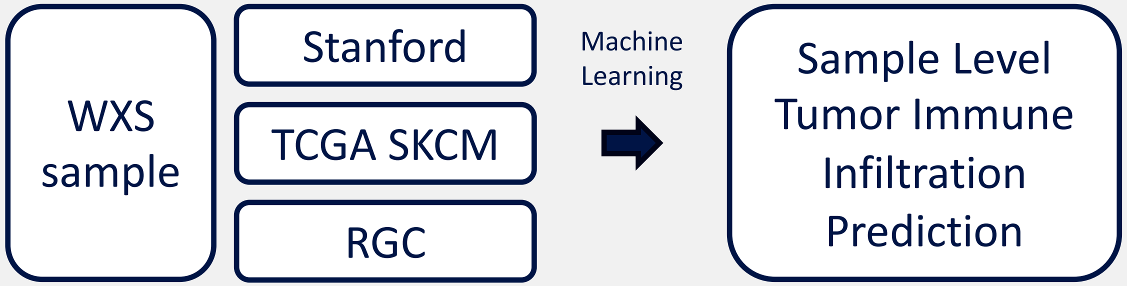


Store embeddings and features in a single file

Store and process each bucket independently

RGC

Next Steps



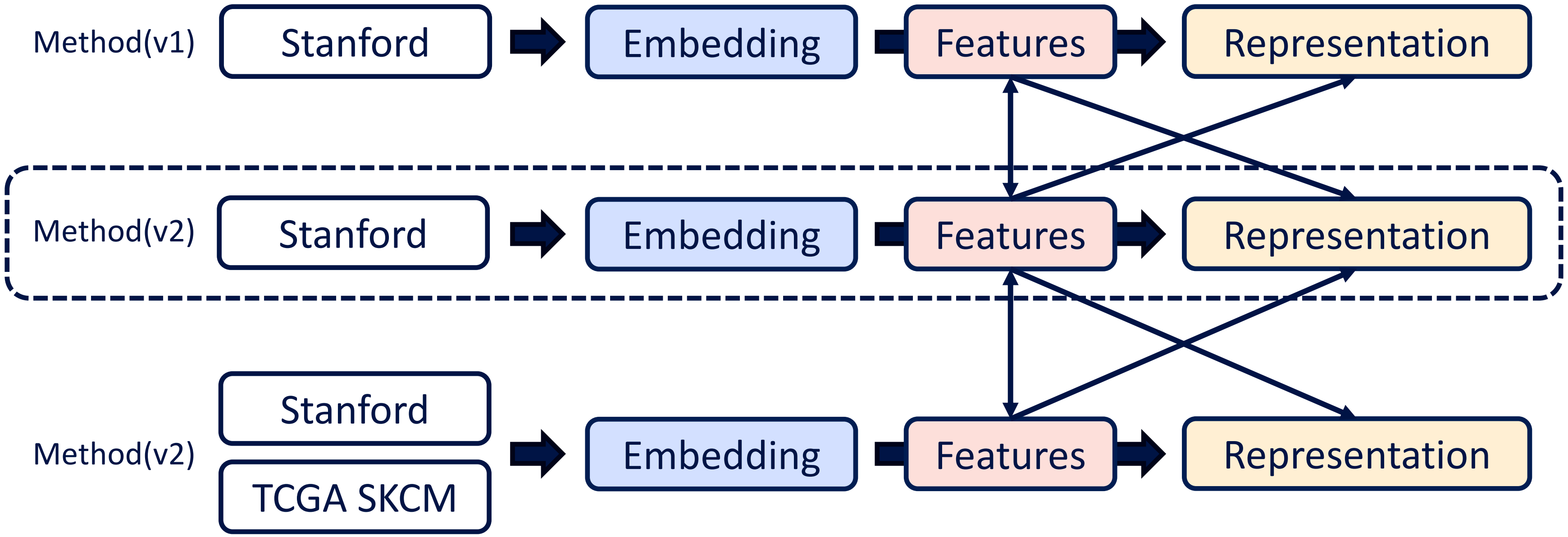
- Goal 1: Viability of predict tumor immune infiltration by DNA sequence on Stanford data
- Goal 2: Finetune the pipeline on TCGA SKCM data
- Goal 3: Predict on RGC data

Verify Our Assumption

Differences of positions/buckets between
- v1 and v2 features
- Stanford and TCGA features

by
- checking covered regions
- cross validating the pipeline

before jump to prediction



Next Steps

WXS sample

Stanford
TCGA SKCM
RGC

Machine Learning ➡

Sample Level Tumor Immune Infiltration Prediction

- Goal 1: Viability of predict tumor immune infiltration by DNA sequence on Stanford data
- Goal 2: Finetune the pipeline on TCGA SKCM data
- Goal 3: Predict on RGC data

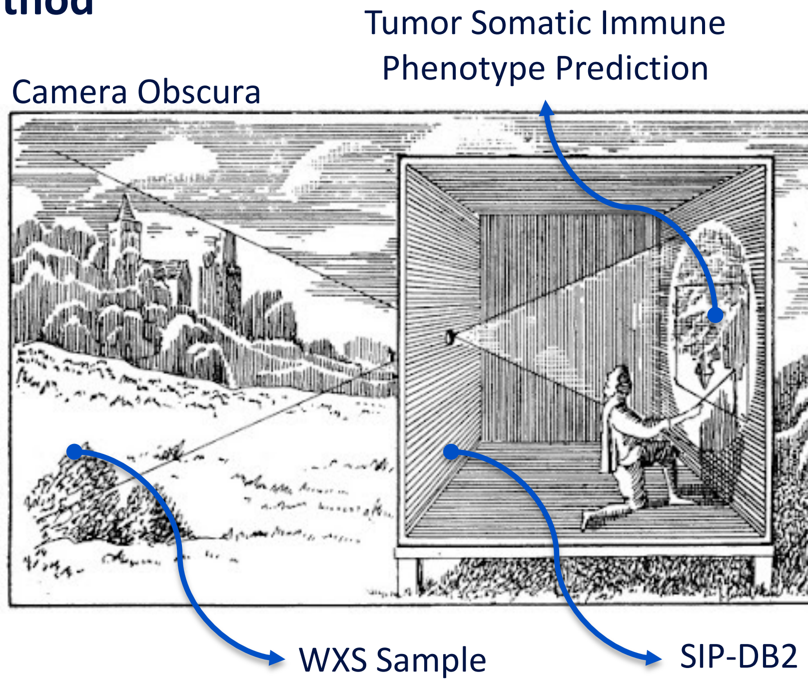
Vertify Our Assumption

Differences of positions/buckets between
- v1 and v2 features
- Stanford and TCGA features

by
- checking covered regions
- cross validating the pipeline

before jump to prediction

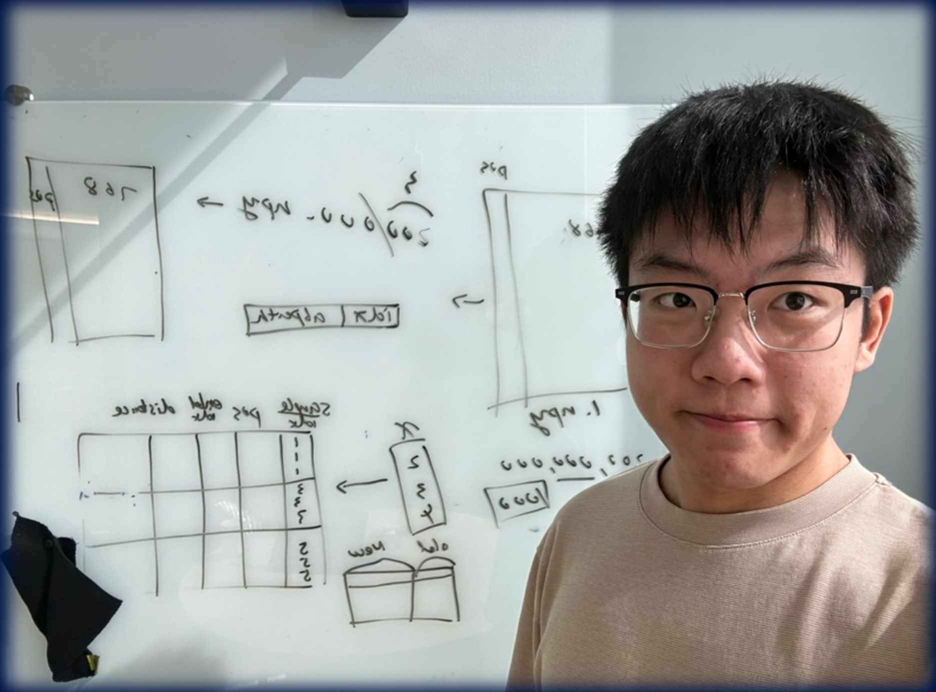
Method



By vectorizing WXS samples:
- reduce computational demand
- efficient similarity searches
- scalability and multimodality

RGC

RGC



- Exposed me to data volumes far surpassing those in prior research, sharpening my skills in algorithm design, engineering, and optimization.
- It instilled a translational mindset, an awareness of scalability and practical application, that I aim to bring into optical imaging.