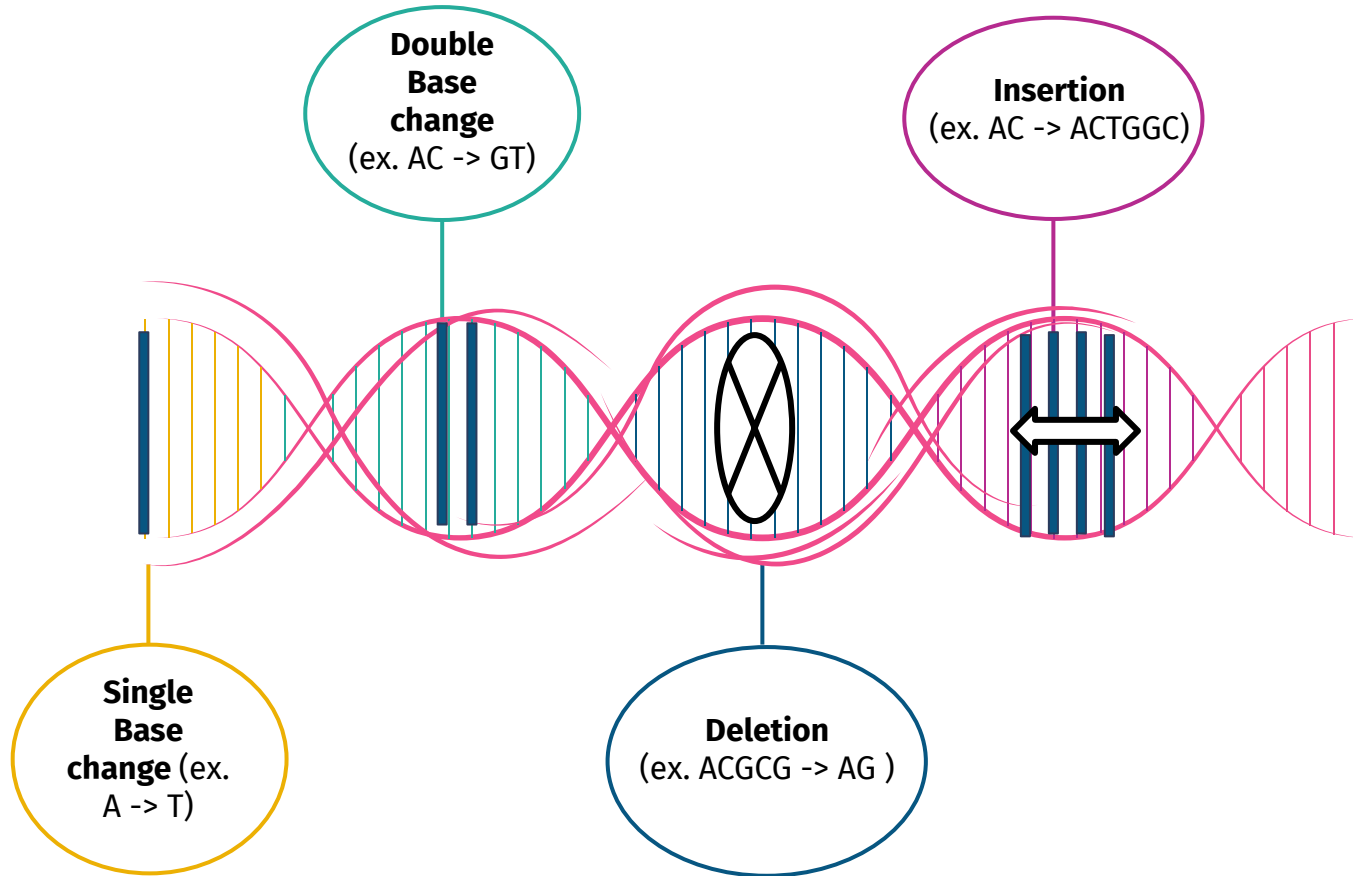




Detecting Tumor Mutational Signatures with a CNN

Beth Baumann, Metis Final Project, Fall 2020

Types of single nucleotide variants (SNVs) and indels



Variant: Occurs in healthy human population

Mutation: Occurs in tumor, but not common in healthy population

What are Mutational Signatures?

**Ultraviolet
Light
Exposure**

C -> T

**BRCA1
Mutations**

Deletions

**DNA
mismatch
repair
defect**

Deletions of
homopolymers
(TTTTTT)

**Tobacco
carcinogens**

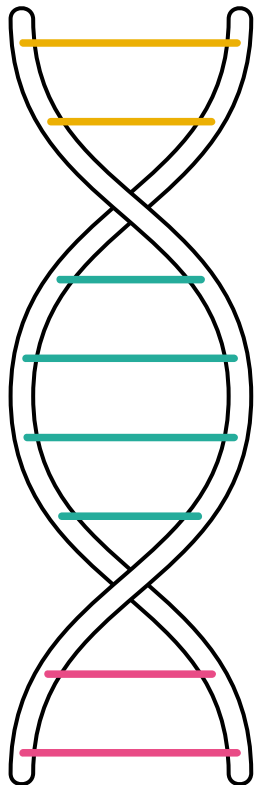
C -> A

**APOBEC
activity**

TC -> TT



Project Goals and Applications



1

Train a neural network to identify a tumor type from short aligned DNA sequences

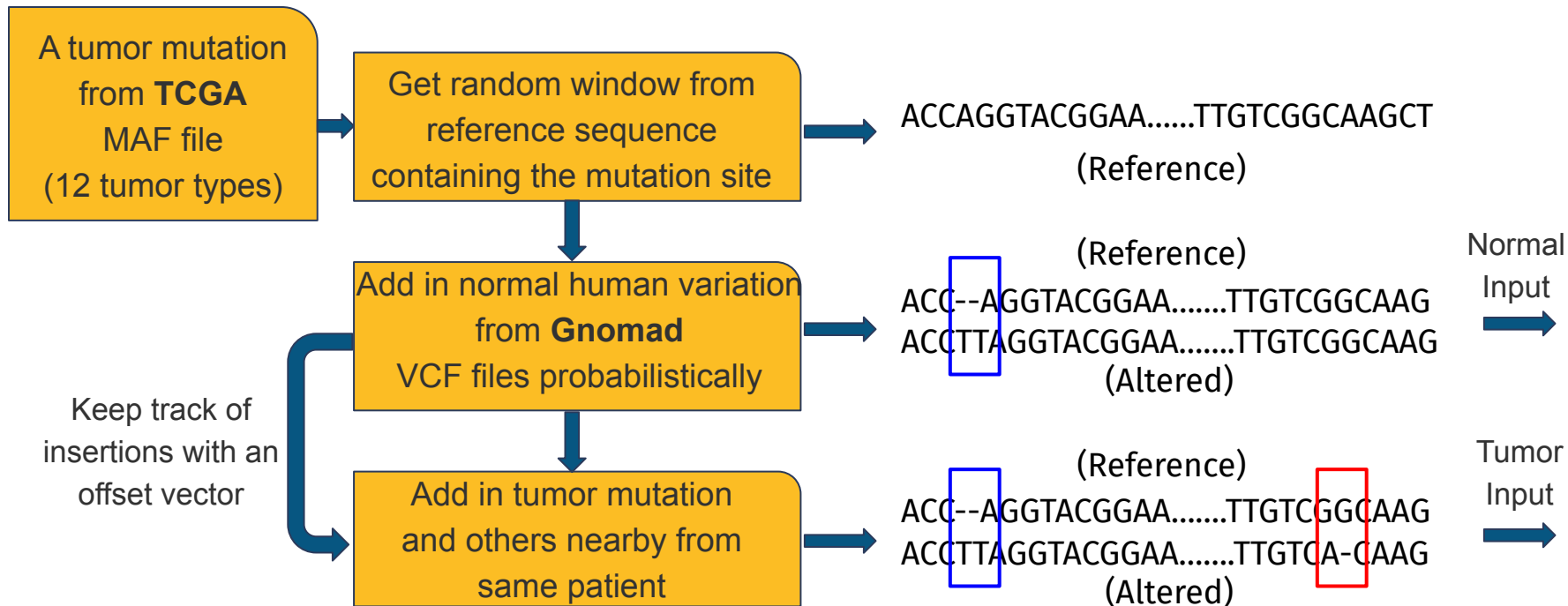
2

Discover mutational signatures and insights about tumor types

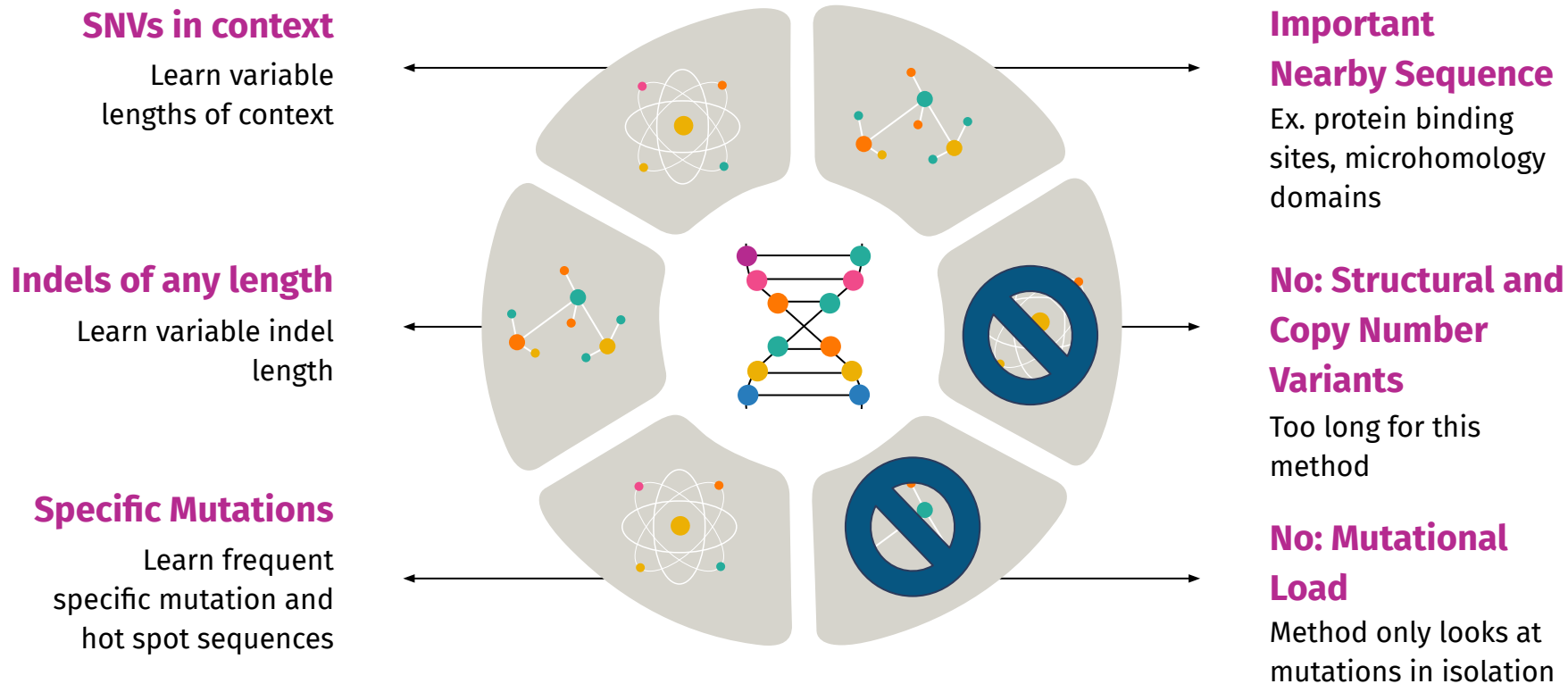
3

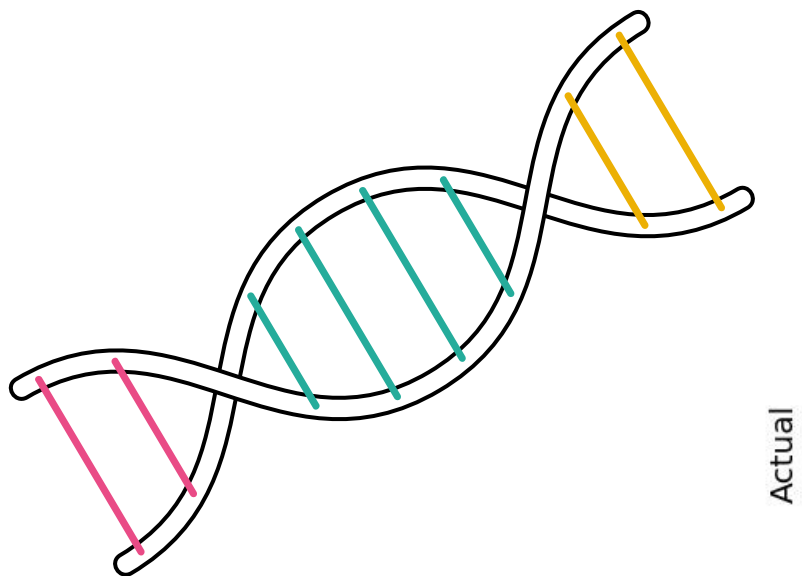
Predict tumor origin of sequencing reads from liquid biopsy
(if model well-developed with a high F1 score)

Generating *synthetic but realistic* 100bp exome sequencing read alignments



What signatures can a Convolutional Neural Network learn about?





CNN Confusion Matrix Normal

Actual \ Predicted	normal	bladder	breast	colorectal	glioma/blastoma	lung	pancreatic	renal	prostate	skin	stomach	uterine	liver
normal	18625	336	23	36	137	472	512	465	945	648	410	229	1162

CNN Confusion Matrix Tumor Types

Actual \ Predicted	normal	bladder	breast	colorectal	glioma/blastoma	lung	pancreatic	renal	prostate	skin	stomach	uterine	liver
bladder	42	774	22	4	29	179	116	99	103	301	45	108	178
breast	64	384	12	11	23	198	138	188	193	253	101	151	284
colorectal	36	77	2	32	63	113	275	130	374	166	257	250	225
glioma/blastoma	8	114	9	7	59	180	243	180	354	233	97	213	303
lung	30	210	13	5	33	610	172	151	100	206	67	83	320
pancreatic	7	97	2	8	53	143	555	68	380	212	130	195	150
renal	14	124	7	9	43	221	159	485	166	192	82	92	406
prostate	8	94	9	11	52	149	306	133	554	182	166	131	205
skin	30	192	3	2	33	78	108	37	93	1184	37	60	143
stomach	36	75	7	25	44	148	290	141	366	155	342	127	244
uterine	39	69	3	28	52	119	303	82	286	210	173	431	205
liver	55	104	12	8	31	298	149	209	142	203	101	93	595

Where is the CNN placing importance for classification?

ACGGAGAAATTTATCCATCAGATTTTGCCGTGGAGATACTTTTGGCGAGAAAAATGACTTCCAGTGATGTTGTAGCTGGATCCGATTAAGTATAGCTCCCC
ACGAAGAATTTATTTCATCAGATTTTGCCGTGGAGATACTTTTGGCGAGAAAAATGACTTCCAGTGATGTTGTAGCTGAATCCAATTAAGTATAGCTCCTC

TTTCCTTTAGGCAGAGGTCTATGAACACCTTCAAGGGCTGGCGCTCTCCCATCCTTGGACAGTCTCCTCACTGTCTGCCTCTTACTCATGGCCTCTGGGGA
TTTTCTTTAGGCAGAGGTCTATGAACACCTTCAAGGGCTGGCGCTCTCCCATCCTTGGACAGTCTCCTCACTGTCTGCCTTTTACTCATAACCTCTGGGGA

CCTGTTCCACTAATTTTCTGAGGCTAATTCCTCTTGAGTTCTGGGCTTTCAATGTTGTTTTGCCTTTAAAAAAAAAAAAAAAAAGAAAGAAAGAAAGAA
CCTGTTCCACTAATTTTCTGAGGCTAATTCCTCTTGAGTTCTGAACTTTCAATGTTGTTTTGCCTTTAAAAAAXXXXXXGAAAAAAAAAAAAA

CTTCTCCCAGTATGAATTATCTTATGTTTAGTAAGGGCTGAAAGATGGTTAAAGCTTTGCCACATTCTTACATTTGTAGGTTTTCCCTCCAGTATGA
CTTCTTTCCAGTATGAATTATCTTATGTTTAGTAAGGGCTGAAAAATGGTTAAAGCTTTGCCACATTCTTACATTTGTAGGGTTTTCTCCAGTATGA

TCCATCCGAGAGCAGGGCAGTGGGAGGAGACGCCATGACCCCATCTCACGGTCTGATCTGTCTCGGTGAGATTGAAGAGGGAGGGGAGCTTCTAA
TCCATCCGAGAGCAGAGCAGTGGGAGGAGACGCTATGACCCCATCTCACAGTCTGATCTGTCTCGGTGAGATTGAAGAGGGAGAGAGCTTCTAA

CTTGGTCTAATTGTTCTCATCTGGAAGACCCTCACCTTCATATCCCAATGTACTTATTCTTGGGAGTTTAGCCTTTGTGGATGCTTTCGTTATCATCCA
CTTGGTCTAATTGTTCTCATCTGGAAGACCCTCACCTTCATATCCCAATGTACTTATTCTTGGAGTTTAGCTTTGTGGATGCTTGTATCATCTA

AGACGTTAATCACGTTTCATGCATCTCCAATCATCATGTTCTAATCTGCCCTCCGGAGGAGGAACAGGTAAGGATTATCCACCTGACGATACAGACXXX
AGACGTTAATCACGTTTCATACATCTCCAATCATCATATCCTAATCTGTCTCTCAAAGAAAGAACAAGTAAGGATTATCCCACTTGACAATACAAGCAAA

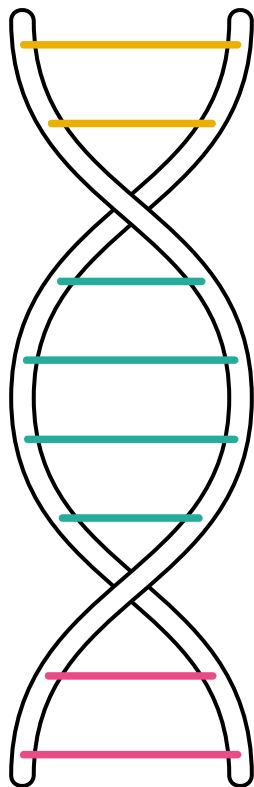
AGACGTTAATCACGTTTCATGCATCTCCAATCATCATGTTCTAATCTGCCCTCCGGAGGAGGAACAGGTAAGGATTATCCACCTGACGATACAGACXXX
AGACGTTAATCACGTTTCATACATCTCCAATCATCATATCCTAATCTGTCTCTCAAAGAAAGAACAAGTAAGGATTATCCCACTTGACAATACAAGCAAA

GCAGTGGCTGCAGGGAGTCACAGAAGGGCAGGACCTGAACGCTGTCTGCTTCCCTGGAATCCAAGATGCTGAGTGGAAGTGGAACCTGGGTGGGCCCGGC
GCAGTGGCTGCAGGAAGTCACAGAAGGGCAGGACCTGAACGCTGTCTGCTTTCTGGAATCCAAGATGCTGAGTGGAAGTGGAACCTGGGTGGGCCCGGC

TCCATCCGAGAGCAGGGCAGTGGGAGGAGACGCCATGACCCCATCTCACGGTCTGATCTGTCTCGGTGAGATTGAAGAGGGAGGGGAGCTTCTAA
TCCATCCGAGAGCAGAGCAGTGGGAGGAGACGCTATGACCCCATCTCACAGTTCTGATCTGTCTCGGTGAGATTGAAGAGGGAGAGAGCTTCTAA

Top Skin Tumor-associated sequences

C to T change is known common UV-related mutation

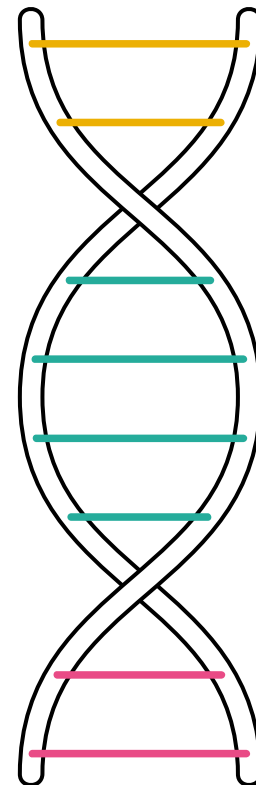


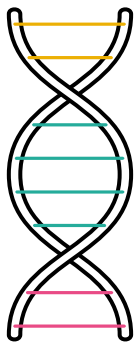
Reference	Altered	Avg Importance	#	Avg Importance x #
CTCCCTCC	CTTTCTTT	273.136314	6	1638.817886
CTTCCTTCCTTT	CTTTCTTTTTTT	302.177643	4	1208.710571
TTCC	TTTC	83.493270	13	1085.412514
CTTC	CTTT	91.110146	11	1002.211605
TCTCCTTT	TCTTTTTT	161.271838	5	806.359192
TTTCTTTC	TTTTTTTT	168.812729	4	675.250916
CTCC	CTTT	132.571730	5	662.858650
CCGGCCGG	CCAACCAA	218.585027	3	655.755081
AAATGGGA	AAATAAGA	160.861458	4	643.445831
TTTCCCTC	TTTTTCTC	159.043461	4	636.173843

Top Lung Tumor-associated Sequences

G to T change is known mutation from benzo(a)pyrene in tobacco

Reference	Altered	Avg Importance	#	Avg Importance x #
GGGG	GTGG	44.803573	8	358.428585
GGTCGGTCCGTG	GGTAGGTAAGTG	159.669815	2	319.339630
GGTG	GTTG	38.266541	8	306.132324
GGCCCCAT	GGCCAAAT	89.676956	3	269.030869
TCCCTCCCCAGGTCAT	TCCATCCAAAGGTCAT	126.131210	2	252.262421
GCCTGCCTCGGCCACCCGCG	GCCTGCCTAAGCCACCAGCG	124.771141	2	249.542282
ACGATGATGAGGCCAGGATCTGT	ACGATGATGAGTCCCATTATGTGT	117.515015	2	235.030029
AGCCAGCCCAG	AGCCAGCCAGAG	78.167747	3	234.503242
CCTCCCTCCCAGTTGCGCTGGGTG	CCTCCCTCCAAGTTGAGCTGGGAG	112.450432	2	224.900864
AGAGAGAGGGGCTCACCTGCCGGC	AGAGAGAGTGGCTCACATGCCGGC	110.537506	2	221.075012



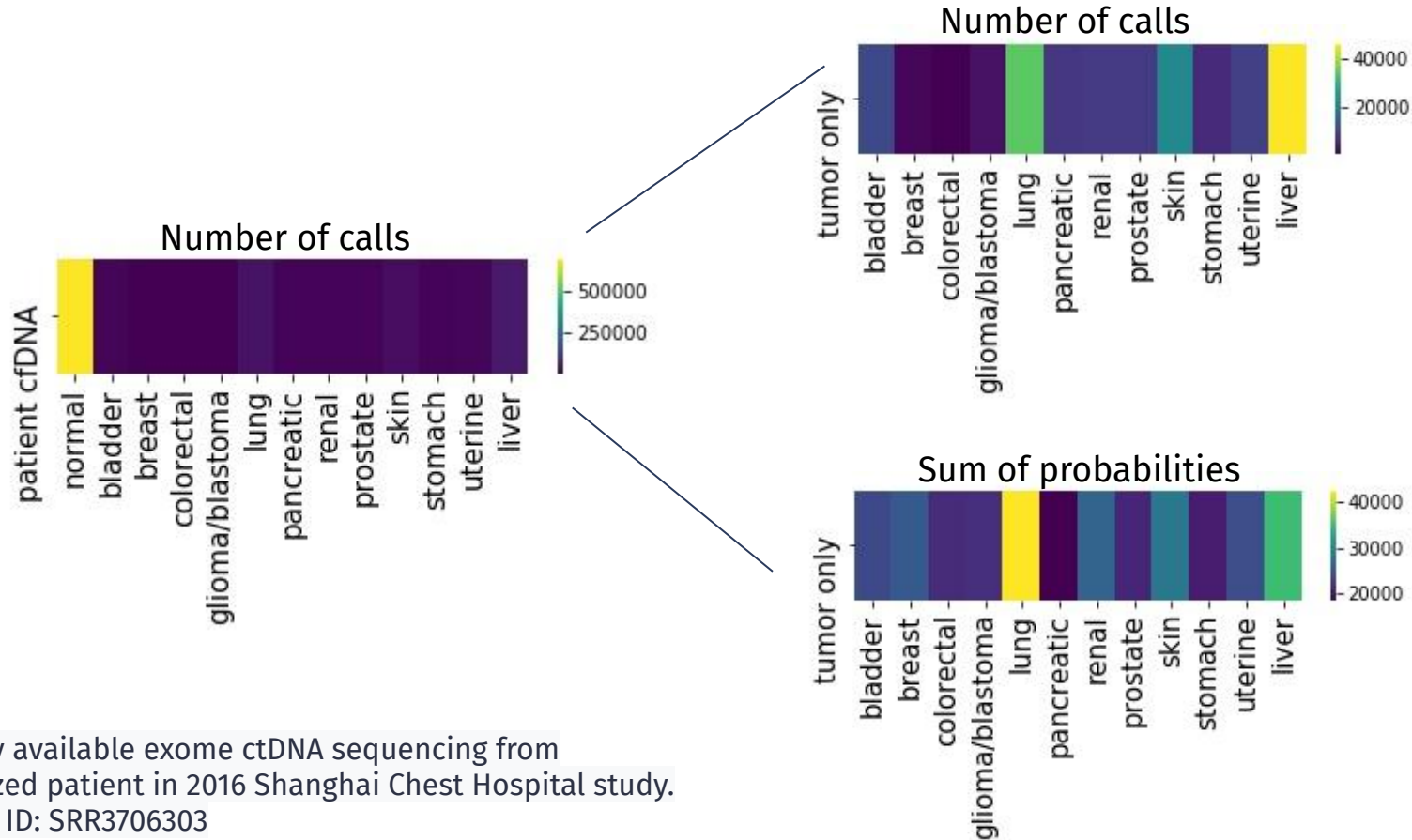


Top Renal Clear Cell Carcinoma-associated Sequences

Insertions and deletions are reported to be most common in this cancer

Reference	Altered	Avg Importance	#	Avg Importance x #
-----	AAAAAAAA	38.893335	13	505.613358
----	AAAA	17.492642	23	402.330760
-----	AAGAAAGAAGAA	37.159491	9	334.435417
CTACCTACCACCACTACCACT-----AATAG	CTACCTACTAC---TACCACTATAATAATAATAATAATAAATAG	128.056763	2	256.113525
----	AAGA	12.586613	19	239.145653
AGATAGATG-----A	AGAAAGAAGGAAAGAAGAAAGAAAGAAGAAAA	107.190369	2	214.380737
AAT-----CA-----G	AATAATGAATGCACATCATG	102.416496	2	204.832993
CCCT-----TATCTATC-----TCCCAG	CCCTTATCTATCTATCTATCTATCTATATATCCCAG	101.835312	2	203.670624
----	TATA	9.229853	22	203.056776
CACAGCTT-----TC-----TTTT	CACAGCTTTTTGAATTAAGTTTGAATTAAGTCTAATGTATTAATGTATTTTT	101.356491	2	202.712982

Applying Model to Lung Cancer Patient cell-free DNA (cfDNA)



Take-Aways and Future Directions

Some tumor types are more distinguishable than others

Skin cancer is very identifiable and breast cancer is not

Generating synthetic reads was a useful alternative

No real patient DNA used for training

CNN can learn mutational signatures

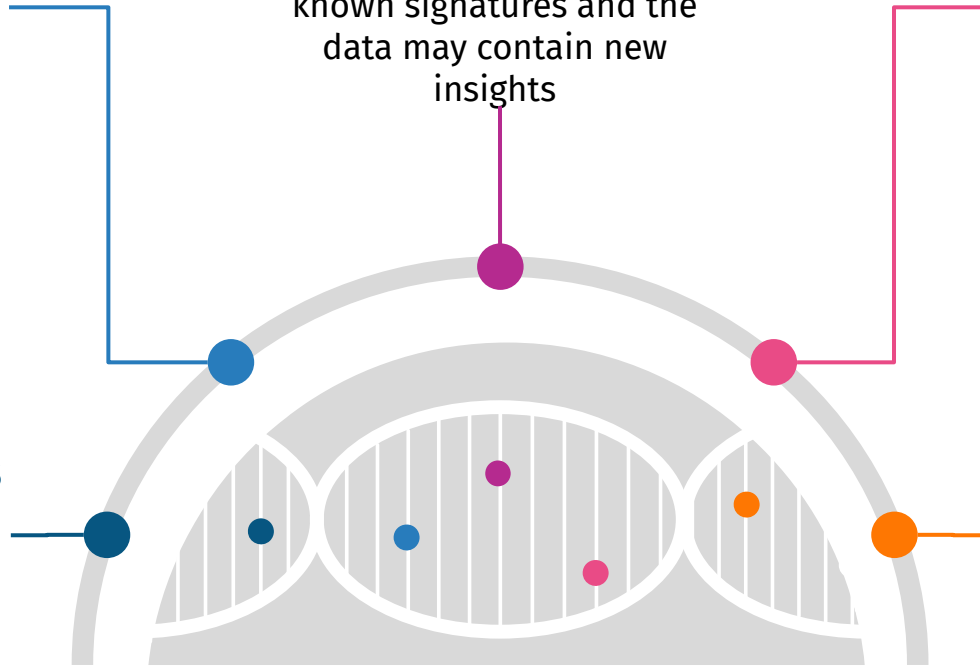
CNN recapitulated some known signatures and the data may contain new insights

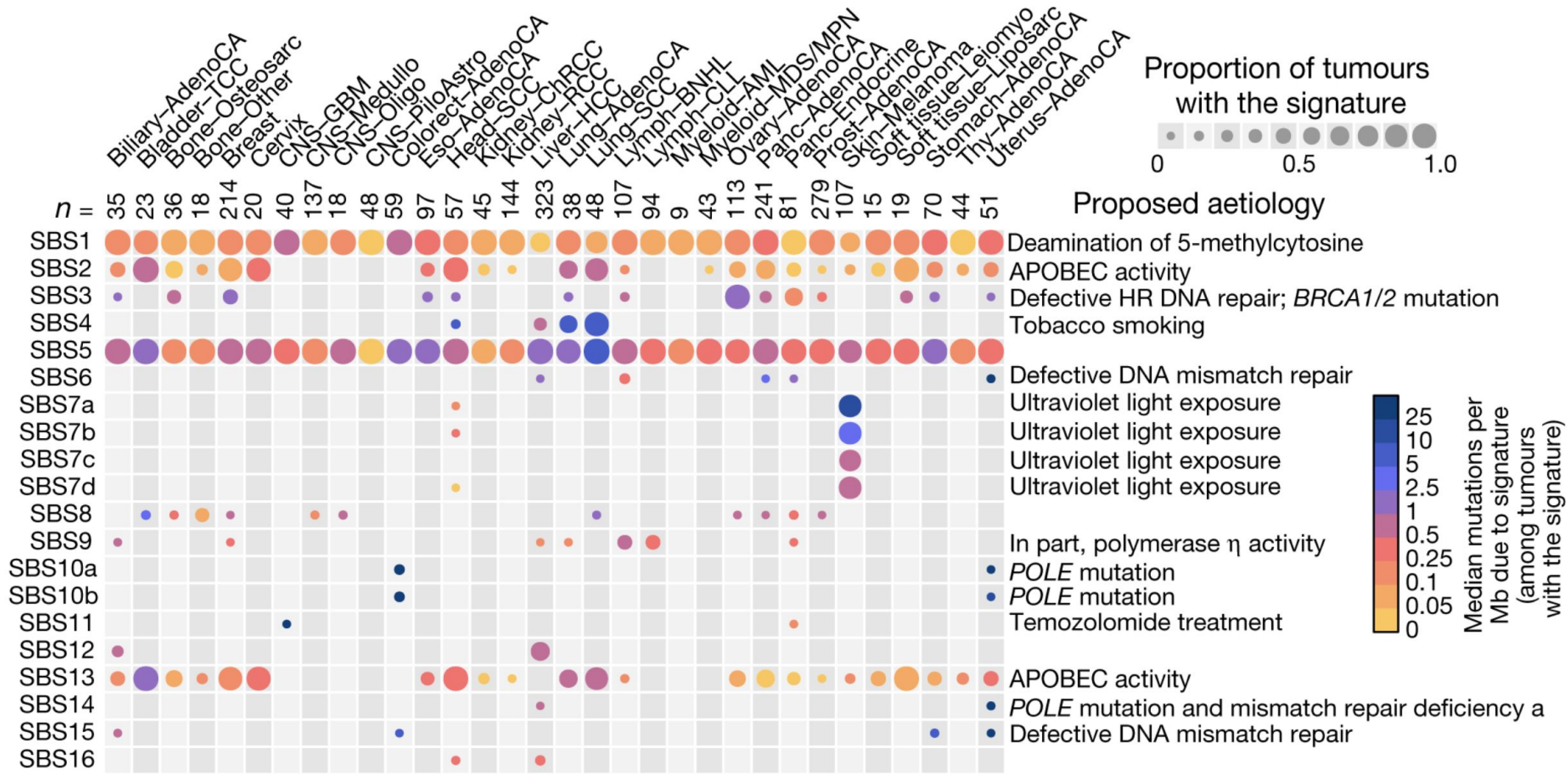
Model can be used with real sequencing data

Real sequences can be format as input and classified

For intratumor signatures:

Train model to predict by patient for finer grain signature learning





Top portion of table from Alexandrov, L.B., Kim, J., Haradhvala, N.J. et al. *Nature* 578, 94–101 (2020).

Instructions for use

In order to use this template, you must credit [Slidesgo](#) and [Freepik](#) in your final presentation and include links to both websites.

You are allowed to:

- Modify this template.
- Use it for both personal and commercial projects.

You are not allowed to:

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute Slidesgo Content unless it has been expressly authorized by Slidesgo.
- Include Slidesgo Content in an online or offline database or file.
- Offer Slidesgo templates (or modified versions of Slidesgo templates) for download.
- Acquire the copyright of Slidesgo Content.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

<https://slidesgo.com/faqs> and <https://slidesgo.com/slidesgo-school>

Infographics

You can add and edit some **infographics** to your presentation to present your data in a visual way.

- Choose your favourite infographic and insert it in your presentation using Ctrl C + Ctrl V or Cmd C + Cmd V in Mac.
- Select one of the parts and **ungroup** it by right-clicking and choosing “Ungroup”.
- **Change the color** by clicking on the paint bucket.
- Then **resize** the element by clicking and dragging one of the square-shaped points of its bounding box (the cursor should look like a double-headed arrow). Remember to hold Shift while dragging to keep the proportions.
- **Group** the elements again by selecting them, right-clicking and choosing “Group”.
- Repeat the steps above with the other parts and when you’re done editing, copy the end result and paste it into your presentation.
- Remember to choose the “**Keep source formatting**” option so that it keeps the design. For more info, please visit **Slidesgo School**.

