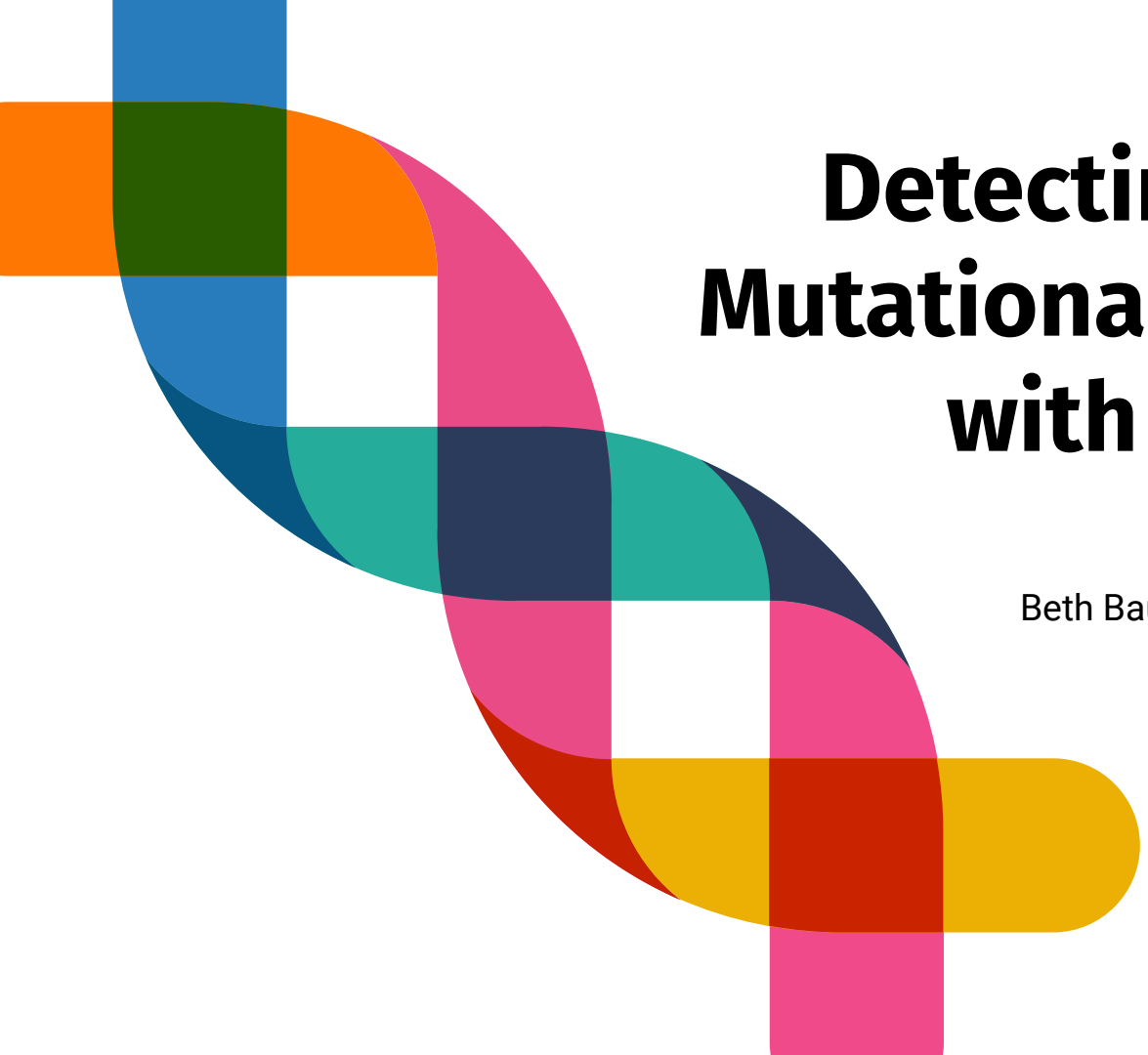
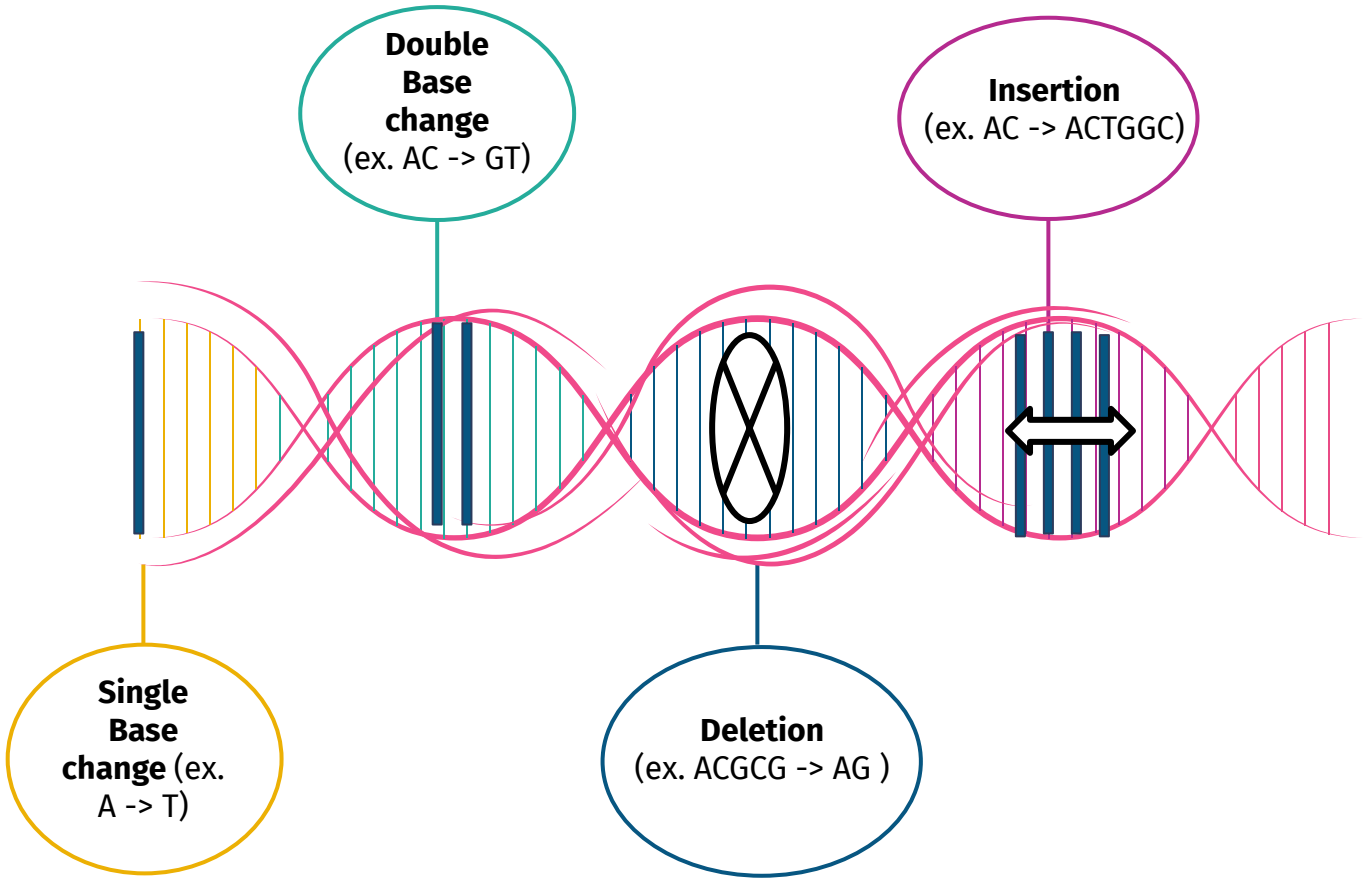


Detecting Tumor Mutational Signatures with a CNN

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

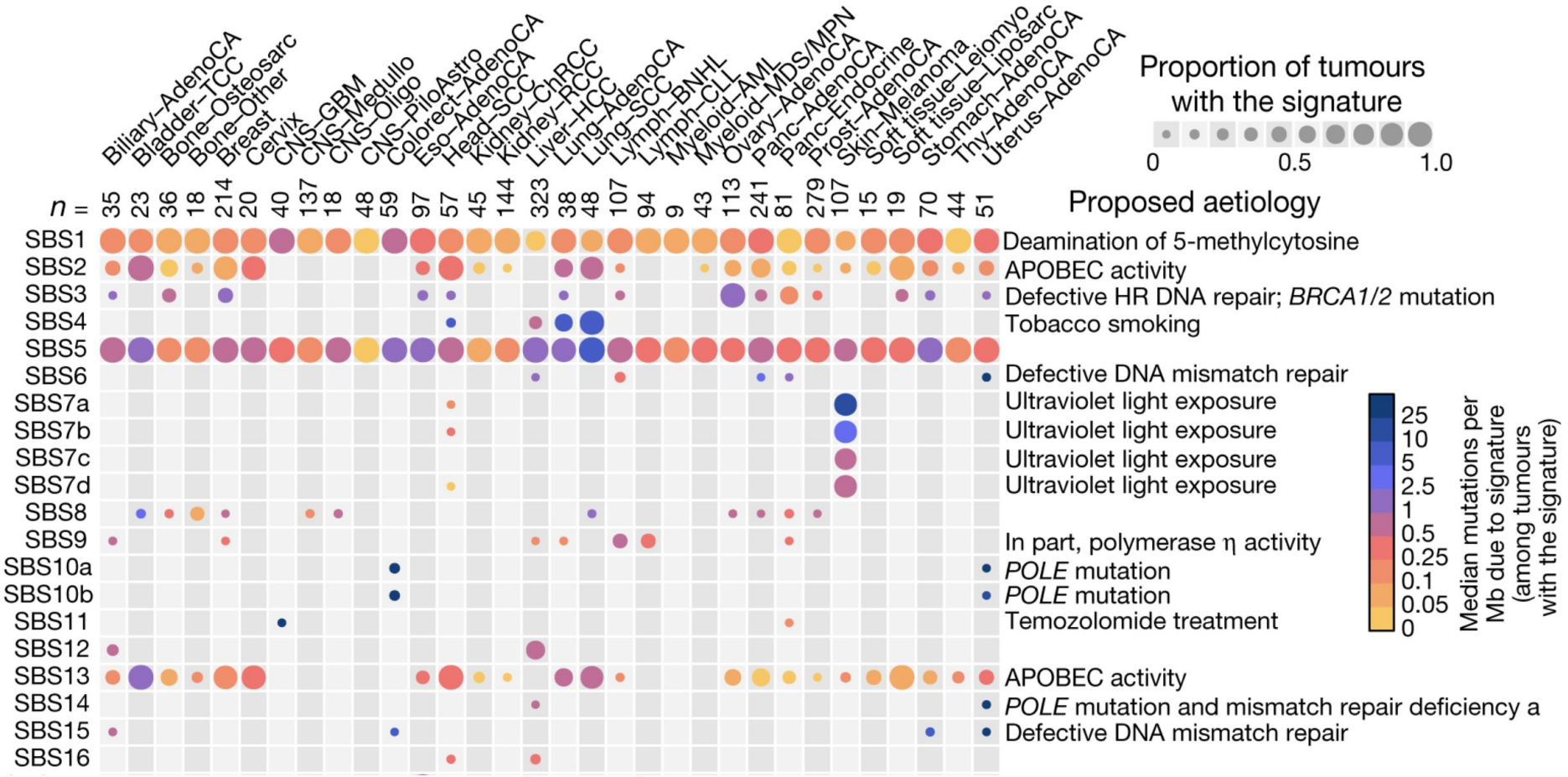
Deletions of homopolymers (TTTTTT)

Tobacco carcinogens

G -> T

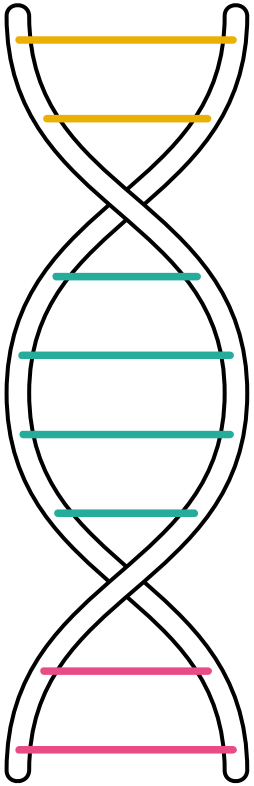
APOBEC activity

TC -> TT



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

Project Goals and Applications

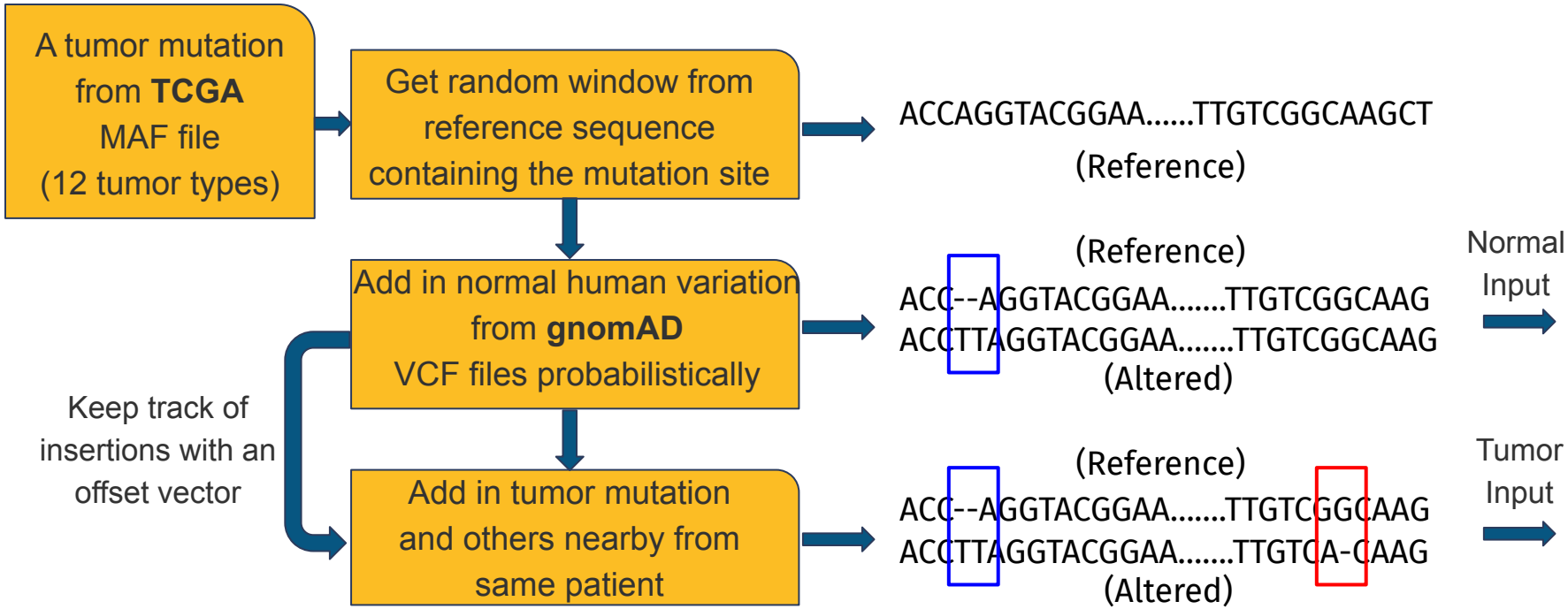


1. **Train** a neural network to identify a tumor type from short aligned DNA sequences

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?

SNVs in context

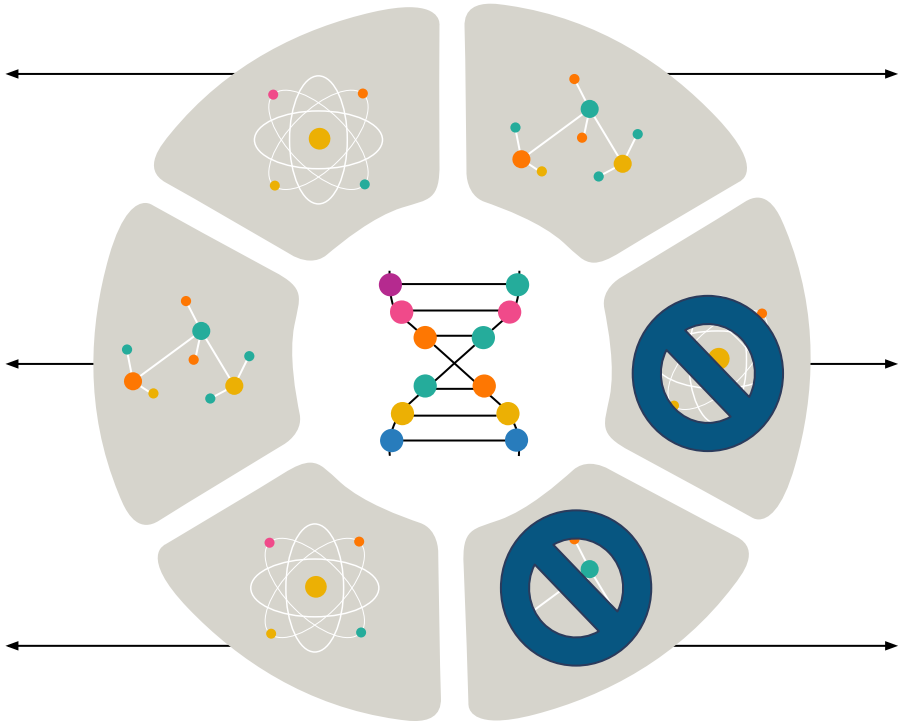
Learn variable lengths of context

Indels of any length

Learn variable indel length

Specific Mutations

Learn frequent specific mutation and hot spot sequences



Important Nearby Sequence

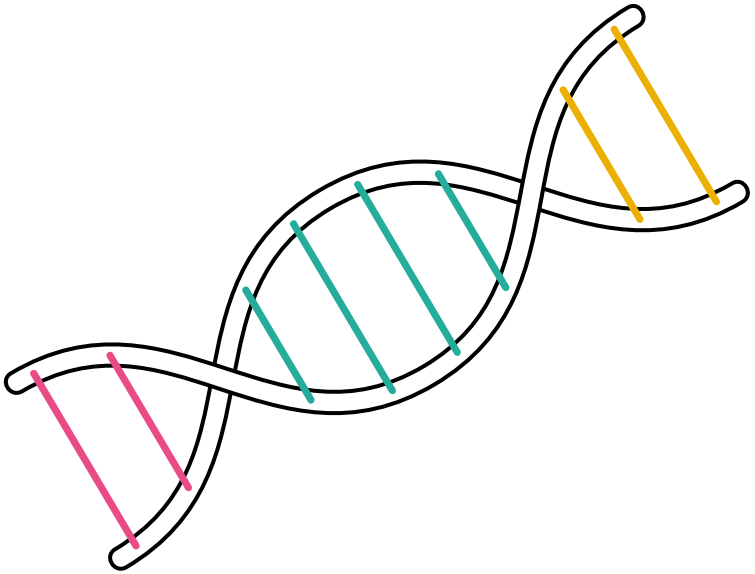
Ex. protein binding sites, microhomology domains

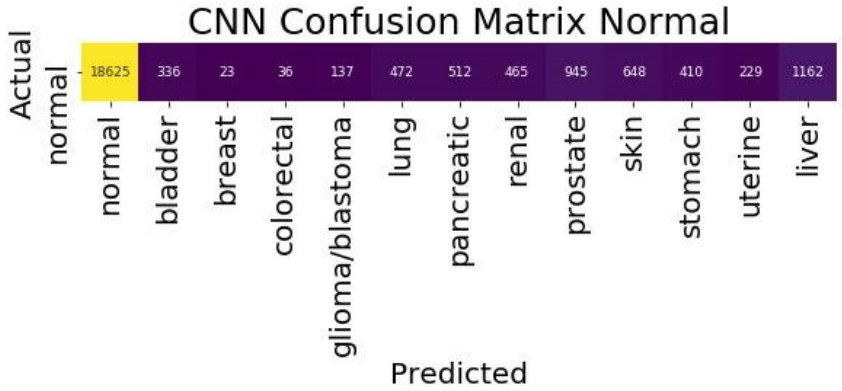
No: Structural and Copy Number Variants

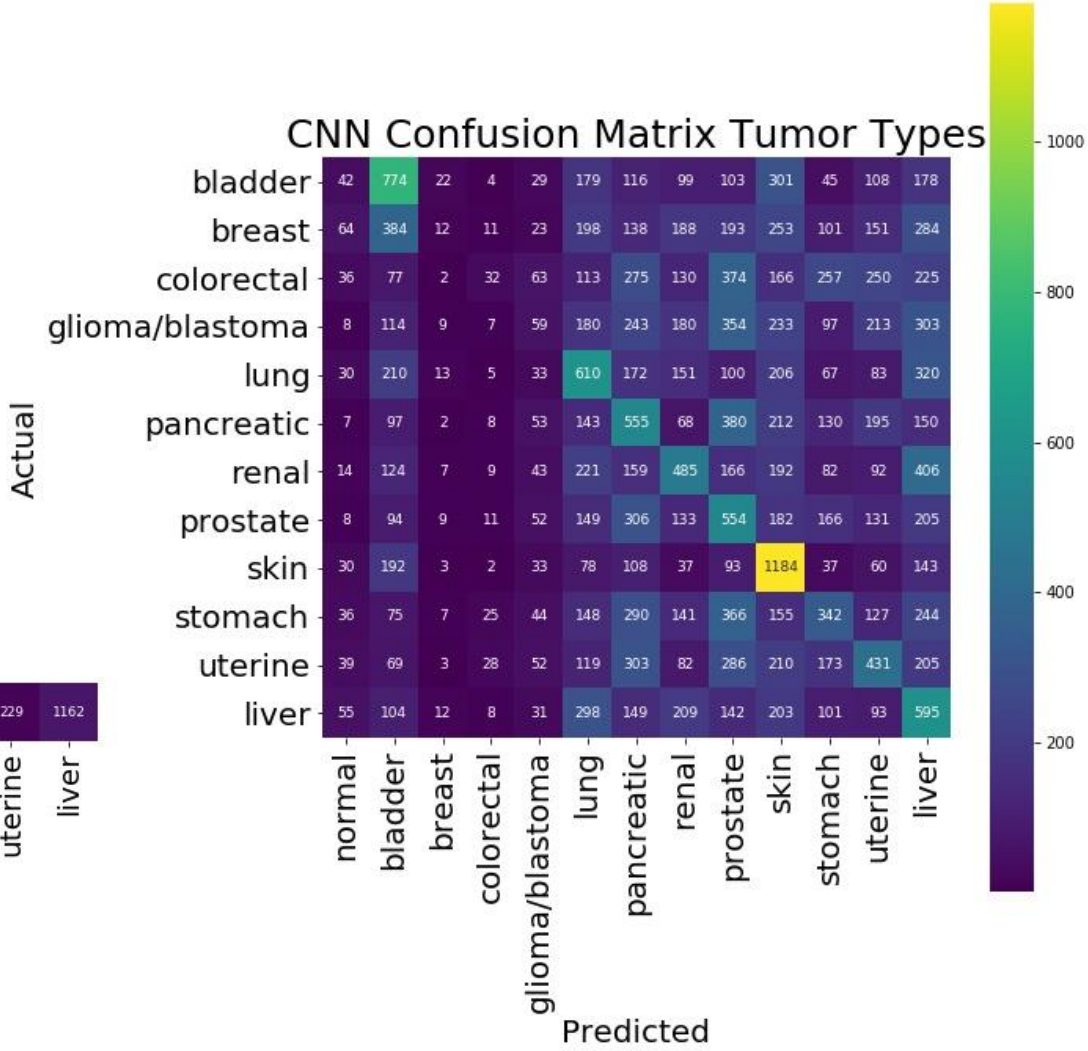
Too long for this method

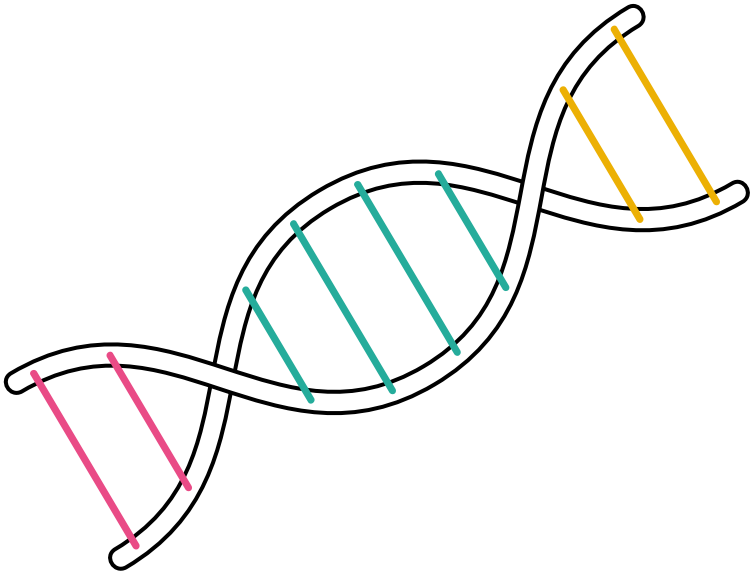
No: Mutational Load

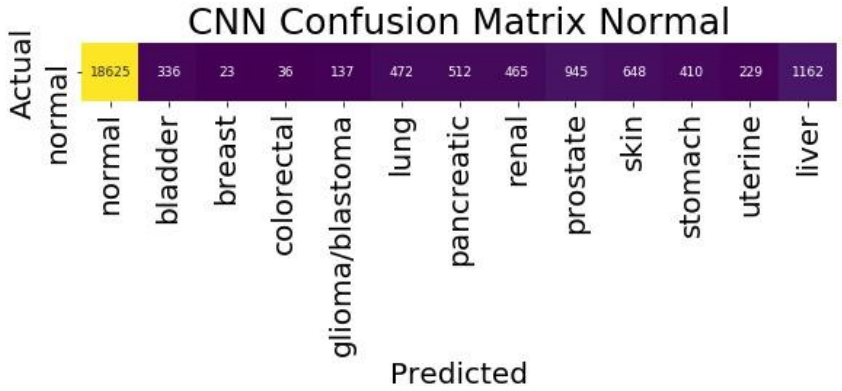
Method only looks at mutations in isolation

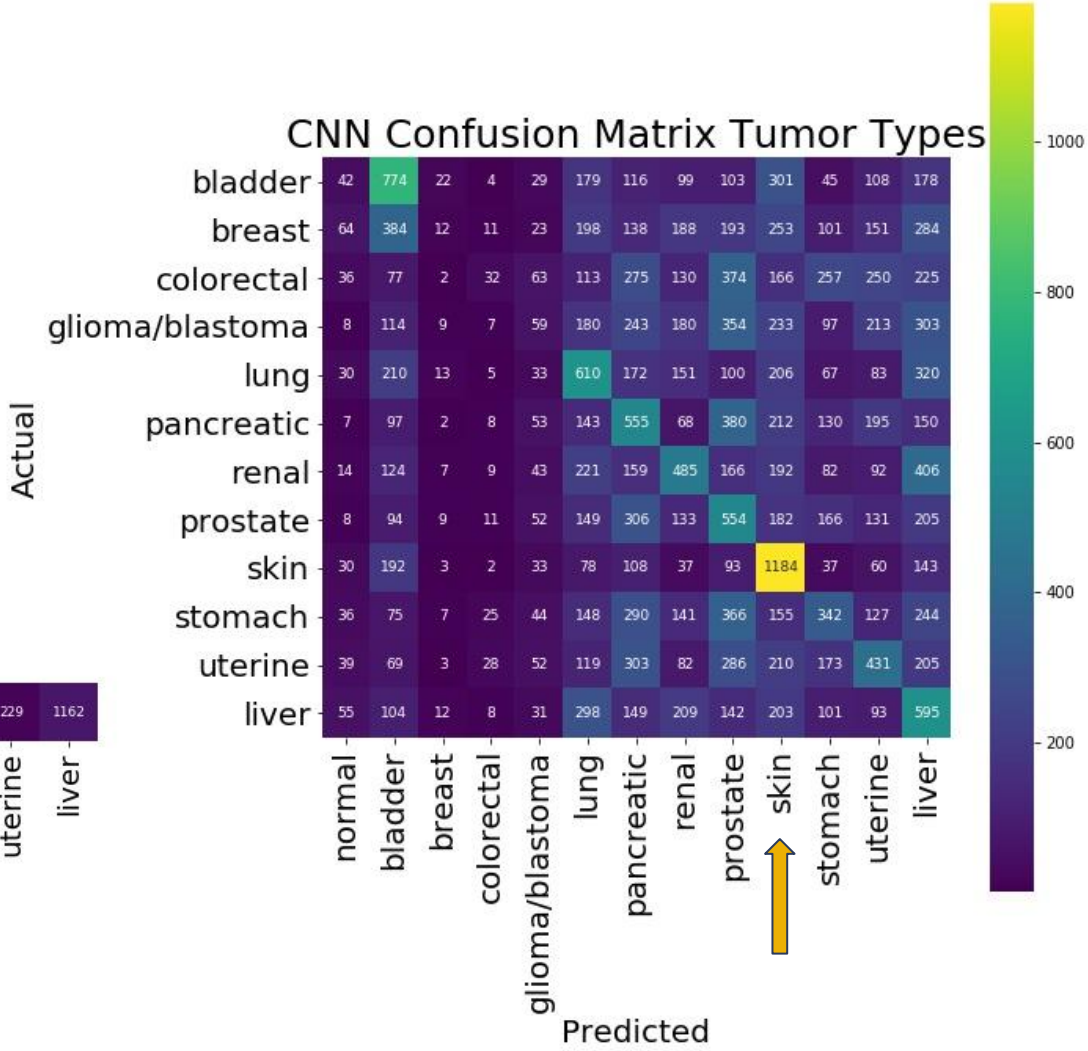


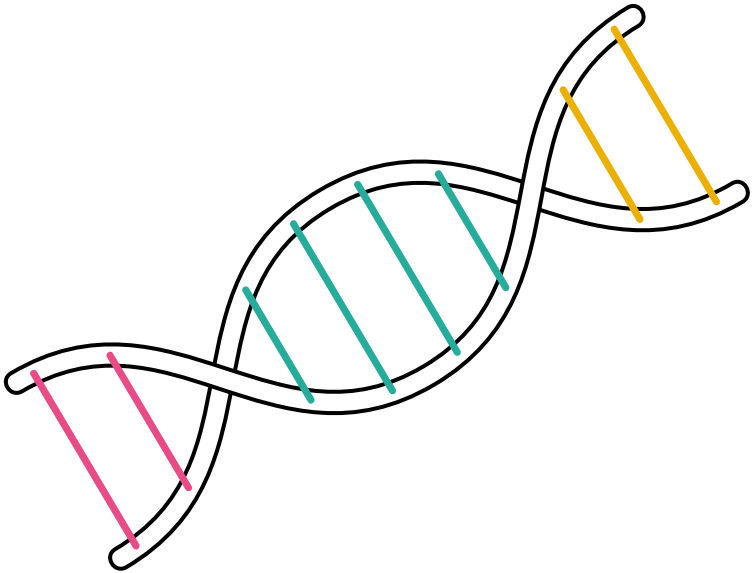

CNN Confusion Matrix Normal
Actual
normal
18625 336 23 36 137 472 512 465 945 648 410 229 1162
normal
bladder
breast
colorectal
glioma/blastoma
lung
pancreatic
renal
prostate
skin
stomach
uterine
liver
Predicted

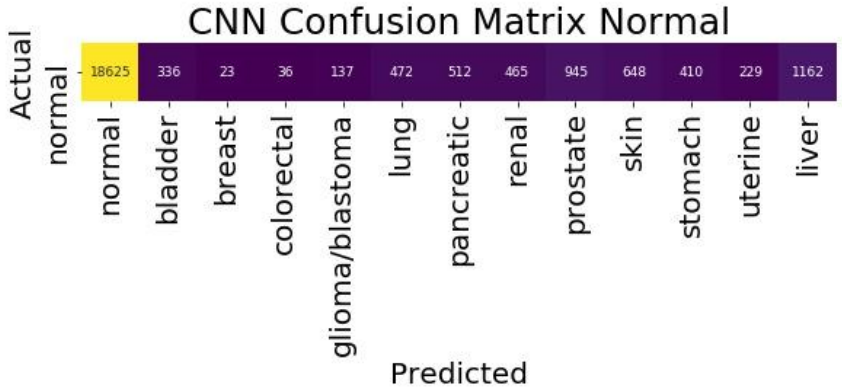

CNN Confusion Matrix Tumor Types
Actual
bladder 42 774 22 4 29 179 116 99 103 301 45 108 178
breast 64 384 12 11 23 198 138 188 193 253 101 151 284
colorectal 36 77 2 32 63 113 275 130 374 166 257 250 225
glioma/blastoma 8 114 9 7 59 180 243 180 354 233 97 213 303
lung 30 210 13 5 33 610 172 151 100 206 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
normal
bladder
breast
colorectal
glioma/blastoma
lung
pancreatic
renal
prostate
skin
stomach
uterine
liver
Predicted
1000
800
600
400
200

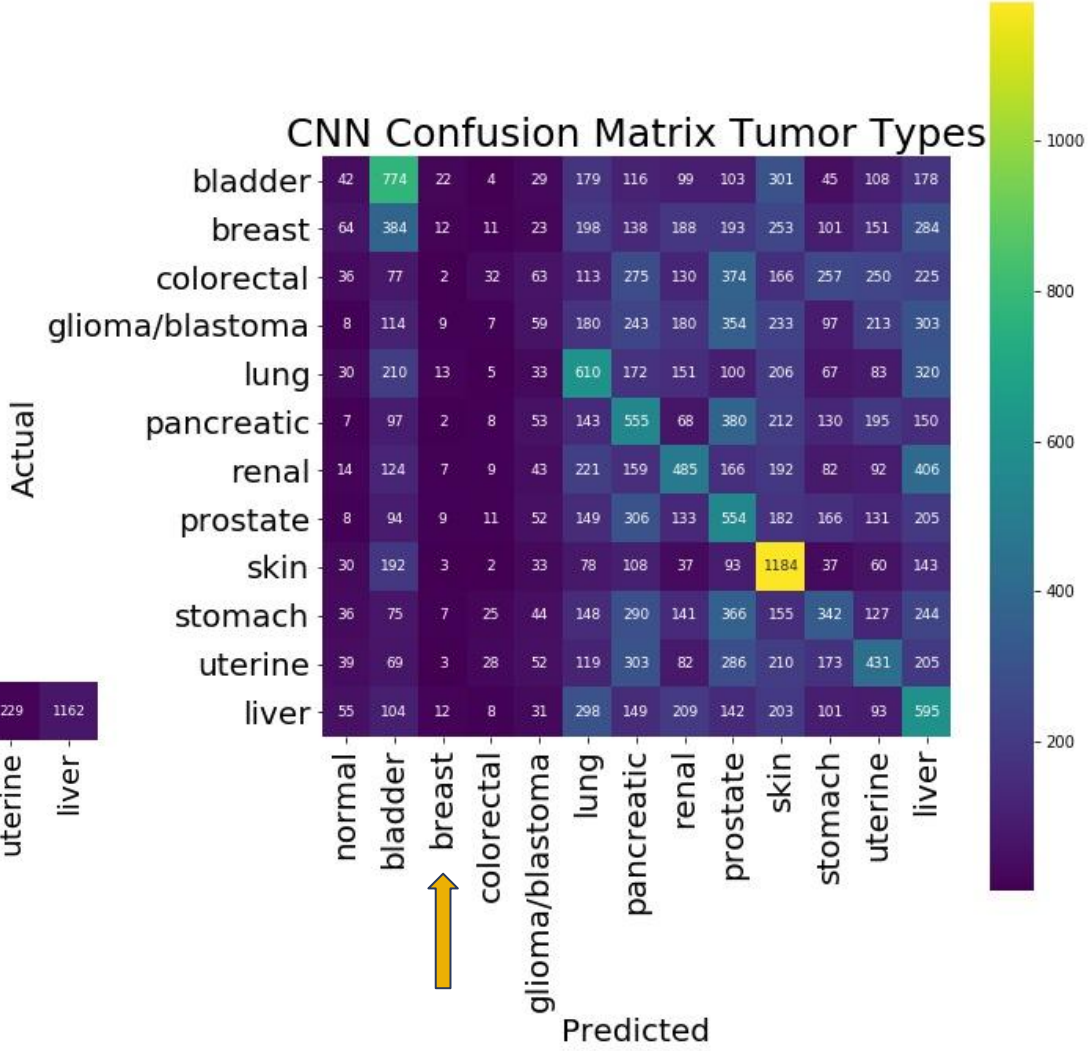



CNN Confusion Matrix Normal
Actual
normal
18625 336 23 36 137 472 512 465 945 648 410 229 1162
normal
bladder
breast
colorectal
glioma/blastoma
lung
pancreatic
renal
prostate
skin
stomach
uterine
liver
Predicted

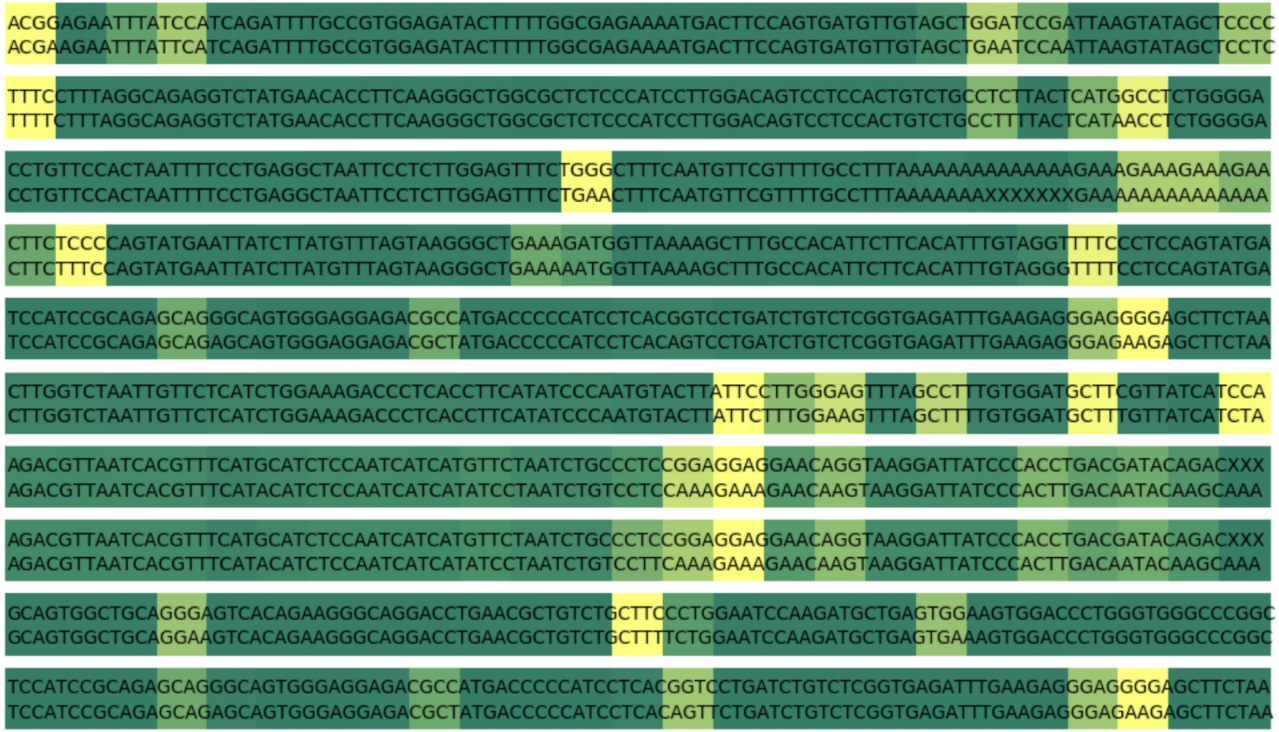

CNN Confusion Matrix Tumor Types
Actual
bladder 42 774 22 4 29 179 116 99 103 301 45 108 178
breast 64 384 12 11 23 198 138 188 193 253 101 151 284
colorectal 36 77 2 32 63 113 275 130 374 166 257 250 225
glioma/blastoma 8 114 9 7 59 180 243 180 354 233 97 213 303
lung 30 210 13 5 33 610 172 151 100 206 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
normal
bladder
breast
colorectal
glioma/blastoma
lung
pancreatic
renal
prostate
skin
stomach
uterine
liver
Predicted
1000
800
600
400
200




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

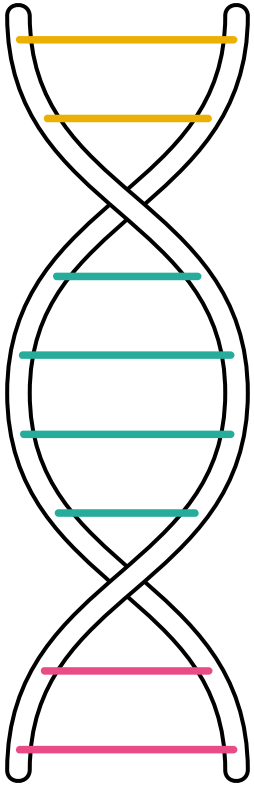

CNN Confusion Matrix Tumor Types
Actual
bladder 42 774 22 4 29 179 116 99 103 301 45 108 178
breast 64 384 12 11 23 198 138 188 193 253 101 151 284
colorectal 36 77 2 32 63 113 275 130 374 166 257 250 225
glioma/blastoma 8 114 9 7 59 180 243 180 354 233 97 213 303
lung 30 210 13 5 33 610 172 151 100 206 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
normal
bladder
breast
colorectal
glioma/blastoma
lung
pancreatic
renal
prostate
skin
stomach
uterine
liver
Predicted
1000
800
600
400
200

Where is the CNN placing importance for classification?



Top Skin Tumor-associated sequences

C to T change is known common UV-related mutation

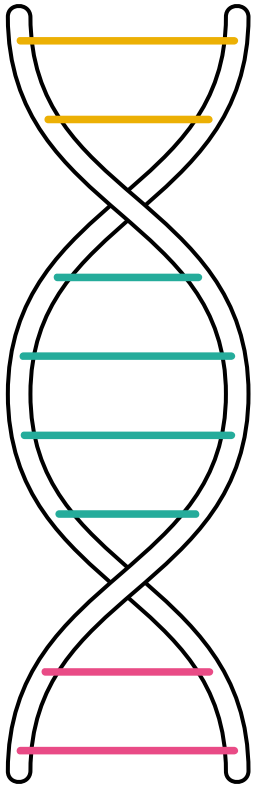


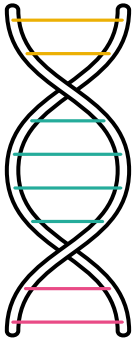
Reference	Altered	Avg Importance	#	Avg Importance x #
CTCCCTCC	**CTTTCTTT**	273.136314	6	1638.817886
CTTCCTTCCTTT	**CTTTCTTTTTTT**	302.177643	4	1208.710571
TTCC	**TTTC**	83.493270	13	1085.412514
CTTC	**CTTT**	91.110146	11	1002.211605
TCTCCTTT	**TCTTTTTT**	161.271838	5	806.359192
TTTCTTTC	**TTTTTTTT**	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	GGTAGGTAACTG	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
ACGATGATGAGGCCCAGGATCTGT	ACGATGATGAGTCCCATTATGTGT	117.515015	2	235.030029
AGCCAGCCCGAG	AGCCAGCCAGAG	78.167747	3	234.503242
CCTCCCTCCCACTTGCGCTGGGTG	CCTCCCTCCAACTTGAGCTGGGAG	112.450432	2	224.900864
AGAGAGAGGGGCTCACCTGCCGGC	AGAGAGAGTGGCTCACATGCCGGC	110.537506	2	221.075012



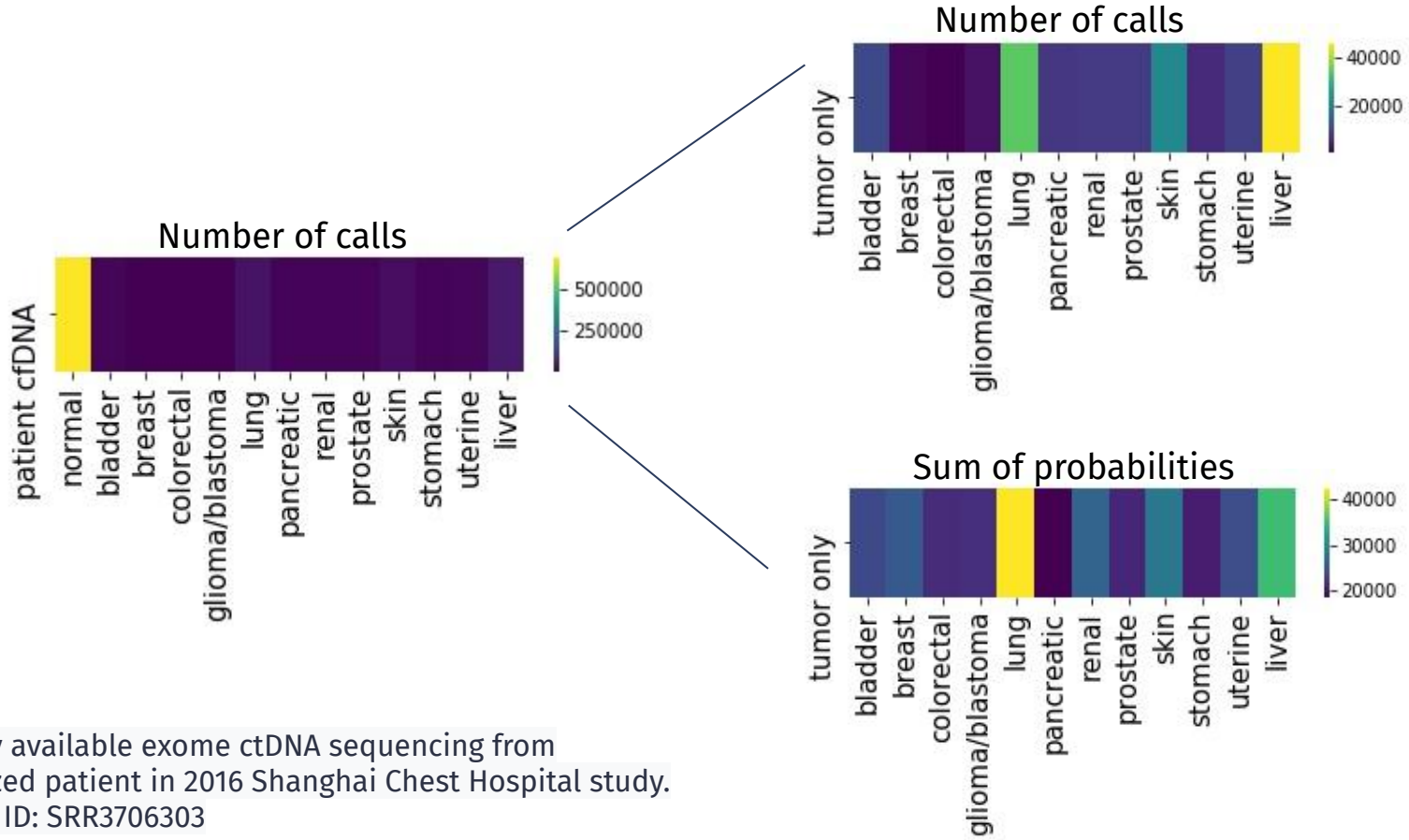


Top Renal Clear Cell Carcinoma-associated Sequences

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---TACCACTATAATAATAATAATAAATAG	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)



Publically available exome ctDNA sequencing from anonymized patient in 2016 Shanghai Chest Hospital study.
NCBI SRA ID: SRR3706303

Take-Aways and Future Directions

CNN can learn mutational signatures

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

Some tumor types are more distinguishable than others

Skin cancer is very identifiable and breast cancer is not

Model can be used with real sequencing data

Real sequences can be format as input and classified

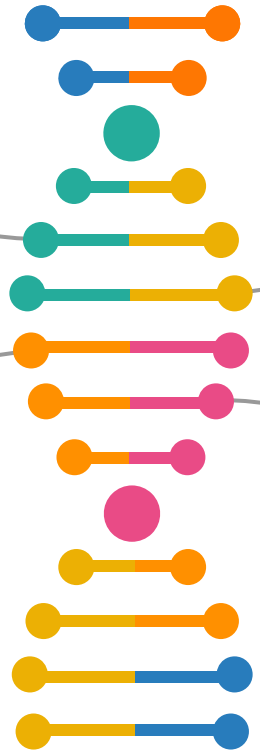
Generating synthetic reads was a useful alternative

No real patient DNA used for training

For intratumor signatures:

Train model to predict by patient for finer-grain signature learning

Thank you!



E-mail

Baumann.Bethany
@gmail.com

GitHub

github/Beth526

Blog

medium.com/
beth-blog

Linkedin

linkedin.com/in/
baumannbethany