

virus Hunters

- Cosmo Mielke
- Onno Faber

Finding Virus data in Tumor and Normal WES/WGS/RNA

- How much viral data is there in both tissue types
- Using unmapped data from the BAM files as well as mapped data as a control
- Using a pre-trained model from ViraMiner called bestViraMinerend2end.hdf5

Tools

- Google Colab notebook, importing relevant **fasta** files and **models**

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 300, 5)	0	
conv1d_1 (Conv1D)	(None, 293, 1000)	41000	input_1[0][0]
conv1d_2 (Conv1D)	(None, 290, 1200)	67200	input_1[0][0]
global_average_pooling1d_1 (GlobalAveragePooling1D)	(None, 1000)	0	conv1d_1[0][0]
global_max_pooling1d_1 (GlobalMaxPooling1D)	(None, 1200)	0	conv1d_2[0][0]
dropout_1 (Dropout)	(None, 1000)	0	global_average_pooling1d_1[0][0]
dropout_2 (Dropout)	(None, 1200)	0	global_max_pooling1d_1[0][0]
fc_layer1 (Dense)	(None, 1000)	1001000	dropout_1[0][0]
fc_layer2 (Dense)	(None, 1200)	1441200	dropout_2[0][0]
concatenate_1 (Concatenate)	(None, 2200)	0	fc_layer1[0][0] fc_layer2[0][0]
drop_fc1 (Dropout)	(None, 2200)	0	concatenate_1[0][0]
dense_1 (Dense)	(None, 1)	2201	drop_fc1[0][0]

```
[4] 1  # HELPER FUNCTIONS
2
3  def DNA_to_onehot(dna_line):
4      options_onehot = {'A': [1,0,0,0,0], 'C': [0,1,0,0,0], 'G': [0,0,1,0,0], 'T': [0,0,0,1,0], 'N': [0,0,0,0,0]}
5      onehot_data = map(lambda e: options_onehot[e], dna_line)
6      onehot_data = np.array(onehot_data)
7
8  return onehot_data
```

```
results.md — ViraMiner
results.md x sample2.fa x Find Results x fullset_test.csv exp3_2013_H4.csv x

# TUMOR UNMAPPED

Percentage of Viral: *11%*
{0.0: 28369, 1.0: 3575}

## Top 100 Virus reads

[('TACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAA
 0.9262255430221558),
 ('TACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAA
 0.9262255430221558),
 ('TACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAA
 0.9262255430221558),
 ('CATACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAA
 0.9195188879966736),
 ('CATACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAA
 0.9195188879966736),
 ('TAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAACGCTACCTGTA
 0.9162285923957825),
 ('TAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAACGCTACCTGTA
 0.9162285923957825),
 ('TGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAACGCT
 0.9148574471473694),
 ('GATGGGCATACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAG
 0.9135288000106812),
 ('GATGGGCATACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAG
 0.9135288000106812),
 ('GAAACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAAC
 0.912702739238739),
 ('ACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAACG
 0.9120316505432129),
 ('ATGGGCATACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAG
 0.9107353091239929),
 ('TGGGCATACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAG
 0.9103326797485352),
 ('ATACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAAC
 0.906359851360321)
```

Tumor Tissue Unmapped

Percentage of Viral: 11%

{0.0: 28369, 1.0: 3575}

```
results.md — ViraMiner
results.md x sample2.fa x Find Results x fullset_test.csv exp3_2013_H4.csv x

# NORMAL UNMAPPED
Percentage of Viral reads: *9.5%*
{0.0: 31480, 1.0: 3321}

## Top 100
[('TACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAA
 0.9262255430221558),
 ('TAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAACGCTACCTGTA
 0.9162285923957825),
 ('TAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAACGCTACCTGTA
 0.9162285923957825),
 ('TGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAACG
 0.9148574471473694),
 ('TGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAACG
 0.9148574471473694),
 ('TGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAACG
 0.9148574471473694),
 ('GATGGGCATACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATT
 0.914459526538489),
 ('GATGGGCATACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATT
 0.9135288000106812),
 ('GATGGGCATACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATT
 0.9135288000106812),
 ('GATGGGCATACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATT
 0.9135288000106812),
 ('ACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAAC
 0.9120316505432129),
 ('ACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAAC
 0.9120316505432129),
 ('ACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAGGGTCAAC
 0.9120316505432129),
 ('TGGGCATACTGTAACCATAAGGCCACGTATTTCAAGCTATTAACCGCGATTGCGTACCCGACGACCAAAATTAG
 0.910326797485352)
```

Normal Tissue Unmapped

Percentage of Viral: 9.5%

{0.0: 31480, 1.0: 3321}

results.md — ViraMiner

results.md sample2.fa Find Results fullset_test.csv exp3_2013_H4.csv

```

# TUMOR MAPPED

Percentage of tumor reads: *0.7%*

## Top 100 virus reads

[('TGTAGACTGAGAAAAATTATAAGCTTCAGTGGTTAACAGTAGCTTCTAAACTGGTTCTGTTACAGTATATTACTT
 0.8685897588729858),
 ('TGTAGACTGAGAAAAATTATAAGCTTCAGTGGTTAACAGTAGCTTCTAAACTGGTTCTGTTACAGTATATTACTT
 0.8685897588729858),
 ('GTAGACTGAGAAAAATTATAAGCTTCAGTGGTTAACAGTAGCTTCTAAACTGGTTCTGTTACAGTATATTACTT
 0.8499968647956848),
 ('GTAGACTGAGAAAAATTATAAGCTTCAGTGGTTAACAGTAGCTTCTAAACTGGTTCTGTTACAGTATATTACTT
 0.8499968647956848),
 ('GTAGACTGAGAAAAATTATAAGCTTCAGTGGTTAACAGTAGCTTCTAAACTGGTTCTGTTACAGTATATTACTT
 0.8499968647956848),
 ('TTGTAGACTGAGAAAAATTATAAGCTTCAGTGGTTAACAGTAGCTTCTAAACTGGTTCTGTTACAGTATATTACTT
 0.8487735986709595),
 ('TTGTAGACTGAGAAAAATTATAAGCTTCAGTGGTTAACAGTAGCTTCTAAACTGGTTCTGTTACAGTATATTACTT
 0.8487735986709595),
 ('GTTTGTAGACTGAGAAAAATTATAAGCTTCAGTGGTTAACAGTAGCTTCTAAACTGGTTCTGTTACAGTATATTTA
 0.8029335141181946),
 ('GTTAACAGTGTAGCTTCTAAACTGGTTCTGTTACAGTATATTACTTGAAGCAATTGGGGGGTACTTCTAAACAG
 0.7962791919708252),
 ('GTTAACAGTGTAGCTTCTAAACTGGTTCTGTTACAGTATATTACTTGAAGCAATTGGGGGGTACTTCTAAACAG
 0.7962791919708252),
 ('GTTTAAGAGAAGGCAAAACACTGTTAAGAAAGTACCCCCCAATTGCTCAAGTAATATACTGTAACAGAAACAGT
 0.7816153764724731),
 ('AGTTTAGACTGAGAAAAATTATAAGCTTCAGTGGTTAACAGTAGCTTCTAAACTGGTTCTGTTACAGTATATT
 0.755334832191467),
 ('TAACCCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAAC
 0.7537426352500916),
 ('TAACCCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAACCTAAC
 0.7537426352500916),
 ('TTTGTAGACTGAGAAAAATTATAAGCTTCAGTGGTTAACAGTAGCTTCTAAACTGGTTCTGTTACAGTATATTAC
 0.7461342811584473)

```

Tumor Tissue Mapped

We took out reads from the BAM file that have been mapped to the human reference genome.

Percentage of Viral: 0.7%

{0.0: 43072, 1.0: 293}

virus-hunter | Group Over | How to Rev | NCBI Blast | NCBI Multi | Complete | Genotyping | Staphylococcus | Nucleotide | blast — Help | FASTA form | blastn notes | +

Find Virus in BLAST

blast.ncbi.nlm.nih.gov/Blast.cgi

short

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments

Download Manage Columns Show 100 ?

CCACGTATTGCAAGCTATTAACGGCGCGATTGCGTAC
CCGACGACCAAAATTAGG

	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/> Hordeum vulgare subsp. vulgare genome assembly, chromosome: 0H	109	109	100%	7e-21	100.00%	LR722623.1
<input checked="" type="checkbox"/> Fusarium proliferatum ET1 A' protein (FPRO_16161), partial mRNA	109	109	100%	7e-21	100.00%	XM_031225275.1
<input checked="" type="checkbox"/> Shigella phage SGF3, complete genome	109	109	100%	7e-21	100.00%	MN266305.1
<input checked="" type="checkbox"/> Escherichia virus phiX174, complete genome	109	109	100%	7e-21	100.00%	MN385565.1
<input checked="" type="checkbox"/> Uncultured prokaryote 16S OTU:542 gene for 16S rRNA, partial sequence	109	109	100%	7e-21	100.00%	LC403762.1
<input checked="" type="checkbox"/> Uncultured prokaryote 16S OTU:2816 gene for 16S rRNA, partial sequence	109	109	100%	7e-21	100.00%	LC402901.1
<input checked="" type="checkbox"/> Uncultured prokaryote 16S OTU:2081 gene for 16S rRNA, partial sequence	109	109	100%	7e-21	100.00%	LC402450.1
<input checked="" type="checkbox"/> Escherichia virus phiX174 strain evolved J1, complete genome	109	109	100%	7e-21	100.00%	MH378443.1
<input checked="" type="checkbox"/> Escherichia virus phiX174 strain evolved J2 line, complete genome	109	109	100%	7e-21	100.00%	MH378442.1
<input checked="" type="checkbox"/> Escherichia virus phiX174 strain evolved J3 line, complete genome	109	109	100%	7e-21	100.00%	MH378441.1
<input checked="" type="checkbox"/> Andersenella sp. Alg231_50 genome assembly, chromosome: VII	109	109	100%	7e-21	100.00%	LT703009.1
<input checked="" type="checkbox"/> Sphingorhabdus sp. Alg231_15 genome assembly, chromosome: II	109	218	100%	7e-21	100.00%	LT703002.1
<input checked="" type="checkbox"/> Erythrobacter sp. Alg231_14 genome assembly, chromosome: II	109	109	100%	7e-21	100.00%	LT703000.1
<input checked="" type="checkbox"/> Plasmopara halstedii replication-associated protein a (PHALS_06510), partial mRNA	109	109	100%	7e-21	100.00%	XM_024719804.1
<input checked="" type="checkbox"/> Escherichia coli DH5alpha plasmid p301-4 contig COV35TF1_c7 genomic sequence	109	109	100%	7e-21	100.00%	MG692646.1
<input checked="" type="checkbox"/> Staphylococcus haemolyticus isolate Staphylococcus haemolyticus K8 genome assembly, chromosome: I	109	109	100%	7e-21	100.00%	LT963441.1
<input checked="" type="checkbox"/> Staphylococcus cohnii isolate Staphylococcus cohnii ATCC 29974 genome assembly, chromosome: I	109	109	100%	7e-21	100.00%	LT963440.1
<input checked="" type="checkbox"/> Staphylococcus xylosus isolate Staphylococcus xylosus ATCC 29971 genome assembly, chromosome: I	109	109	100%	7e-21	100.00%	LT963439.1
<input checked="" type="checkbox"/> Staphylococcus simulans isolate Staphylococcus simulans ATCC 27848 genome assembly, chromosome: I	109	109	100%	7e-21	100.00%	LT963435.1
<input checked="" type="checkbox"/> Culicoides sonorensis genome assembly, scaffold: scaffold781	109	109	100%	7e-21	100.00%	LN484131.1
<input checked="" type="checkbox"/> Synthetic Enterobacteria phage CryptX174, complete genome	109	109	100%	7e-21	100.00%	MF426915.1

Feedback

Lucid-0.90.1ka Hackathon Leila's Case, 2020-01-12

Show All 7

Next Steps

1. Map reads of suggested viruses to actual virus
2. Frequency of the viruses
3. Run it on other data (RNA, WGS)