Mahdi_Anvari.compG_HW3

July 3, 2025

1 Computational Genomics - HW3

To begin, we extracted chromosome 1 from the hg38 reference genome to create a distinct reference sequence. This is useful for reducing computational complexity and focusing analysis on a specific region of interest. After generating the chromosome 1 FASTA file, we indexed it to ensure compatibility with downstream tools that require fast sequence lookup.

[10]:
```
mkdir Reference
cd Reference
```

[11]:
```
samtools faidx /mnt/d/NGS/References/hg38.fa chr1 > chr1.fa
```

[13]:
```
samtools faidx chr1.fa
```

[14]:
```
ls
```

```
chr1.fa  chr1.fa.fai
```

[15]:
```
cd ../
```

Next, we downloaded the BAM file data from the Illumina Comprehensive Cancer Panel. To reduce computational time and resource usage, we filtered the variants to include only those located on chromosome 1. We then sorted the BAM file to improve performance during downstream analysis. Finally, we indexed the sorted BAM file to enable efficient access to alignment data.

[19]:
```
mkdir Data
cd Data
```

[ ]:
```
wget https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/
 ↪Somatic_Mutation_WG/data/AmpliSeq_bams/AmpliSeq.bwa.HCC1395BL_1.bam
wget https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/
 ↪Somatic_Mutation_WG/data/AmpliSeq_bams/AmpliSeq.bwa.HCC1395BL_1.bam.bai
wget https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/
 ↪Somatic_Mutation_WG/data/AmpliSeq_bams/AmpliSeq.bwa.HCC1395_1.bam
wget https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/
 ↪Somatic_Mutation_WG/data/AmpliSeq_bams/AmpliSeq.bwa.HCC1395_1.bam.bai
```

[24]:
```
ls
```

```
AmpliSeq.bwa.HCC1395BL_1.bam
AmpliSeq.bwa.HCC1395_1.bam
AmpliSeq.bwa.HCC1395BL_1.bam.bai
AmpliSeq.bwa.HCC1395_1.bam.bai
```

[30]:
```
samtools view -b AmpliSeq.bwa.HCC1395_1.bam chr1 -o AmpliSeq.bwa.HCC1395_1_chr1.
 ↪bam
samtools view -b AmpliSeq.bwa.HCC1395BL_1.bam chr1 -o AmpliSeq.bwa.
 ↪HCC1395BL_1_chr1.bam
```

[31]:
```
ls
```

```
AmpliSeq.bwa.HCC1395BL_1.bam
AmpliSeq.bwa.HCC1395_1.bam
AmpliSeq.bwa.HCC1395BL_1.bam.bai
AmpliSeq.bwa.HCC1395_1.bam.bai
AmpliSeq.bwa.HCC1395BL_1_chr1.bam
AmpliSeq.bwa.HCC1395_1_chr1.bam
```

[32]:
```
samtools sort AmpliSeq.bwa.HCC1395_1_chr1.bam -o AmpliSeq.bwa.
 ↪HCC1395_1_chr1_sorted.bam
samtools sort AmpliSeq.bwa.HCC1395BL_1_chr1.bam -o AmpliSeq.bwa.
 ↪HCC1395BL_1_chr1_sorted.bam
```

[33]:
```
ls
```

```
AmpliSeq.bwa.HCC1395BL_1.bam
AmpliSeq.bwa.HCC1395BL_1.bam.bai
AmpliSeq.bwa.HCC1395BL_1_chr1.bam
AmpliSeq.bwa.HCC1395BL_1_chr1_sorted.bam
AmpliSeq.bwa.HCC1395_1.bam
AmpliSeq.bwa.HCC1395_1.bam.bai
AmpliSeq.bwa.HCC1395_1_chr1.bam
AmpliSeq.bwa.HCC1395_1_chr1_sorted.bam
```

[34]:
```
samtools index AmpliSeq.bwa.HCC1395_1_chr1_sorted.bam
samtools index AmpliSeq.bwa.HCC1395BL_1_chr1_sorted.bam
```

[35]:
```
ls
```

```
AmpliSeq.bwa.HCC1395BL_1.bam
AmpliSeq.bwa.HCC1395BL_1.bam.bai
AmpliSeq.bwa.HCC1395BL_1_chr1.bam
AmpliSeq.bwa.HCC1395BL_1_chr1_sorted.bam
AmpliSeq.bwa.HCC1395BL_1_chr1_sorted.bam.bai
AmpliSeq.bwa.HCC1395_1.bam
AmpliSeq.bwa.HCC1395_1.bam.bai
AmpliSeq.bwa.HCC1395_1_chr1.bam
```

```
AmpliSeq.bwa.HCC1395_1_chr1_sorted.bam
AmpliSeq.bwa.HCC1395_1_chr1_sorted.bam.bai
```

[38]:
```
cd ../
```

With the data prepared, we proceeded to run the VarNet inference notebook on Google Colab. To enable the analysis, we uploaded the necessary input files to our Google Drive:

1. AmpliSeq.bwa.HCC1395BL_1_chr1_sorted.bam — the normal sample
2. AmpliSeq.bwa.HCC1395_1_chr1_sorted.bam — the tumor sample
3. chr1.fa and chr1.fa.fai — the reference genome (chromosome 1 only)

[40]:
```
cd varnet_outputs
```

[41]:
```
ls
```

```
HCC1395
```

[42]:
```
cd HCC1395
```

[45]:
```
ls
```

```
HCC1395.vcf  candidates  predictions
```

[46]:
```
cat HCC1395.vcf
```

```
##fileformat=VCFv4.2
##fileDate=2025June28, 19:56:28
##source=VarNet v1.1.0
##reference=/content/VarNet/data/chr1.fa
##normalBAM=/content/VarNet/data/AmpliSeq.bwa.HCC1395BL_1_chr1_sorted.bam
##tumorBAM=/content/VarNet/data/AmpliSeq.bwa.HCC1395_1_chr1_sorted.bam
##INFO=<ID=TYPE,Number=.,Type=String,Description="Type of Somatic Event INDEL or
SNV">
##INFO=<ID=SCORE,Number=1,Type=Float,Description="Prediction probability score">
##FILTER=<ID=PASS,Description="Accept as somatic mutation with probability score
at least 0.5">
##FILTER=<ID=REJECT,Description="Reject somatic mutation with probability score
value below 0.5">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth in the tumor">
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation
count in the tumor">
##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observation
count in the tumor">
##FORMAT=<ID=AF,Number=1,Type=Float,Description="Allele fractions of alternate
alleles in the tumor">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SAMPLE
```

```
chr1    169764520       .       G       T       .       PASS
TYPE=SNV;SCORE=0.9998;DP=845;RO=255;AO=589;AF=0.697;    GT:DP:RO:AO:AF
0/1:845:255:589:0.697
chr1    149034140       .       G       T       .       PASS
TYPE=SNV;SCORE=0.9493;DP=1608;RO=1094;AO=512;AF=0.3184; GT:DP:RO:AO:AF
0/1:1608:1094:512:0.3184
chr1    11387326        .       T       A       .       PASS
TYPE=SNV;SCORE=0.9798;DP=870;RO=706;AO=164;AF=0.1885;   GT:DP:RO:AO:AF
0/1:870:706:164:0.1885
chr1    236923385       .       A       G       .       PASS
TYPE=SNV;SCORE=0.6584;DP=718;RO=599;AO=119;AF=0.1657;   GT:DP:RO:AO:AF
0/1:718:599:119:0.1657
chr1    172452442       .       C       G       .       PASS
TYPE=SNV;SCORE=0.9998;DP=704;RO=481;AO=223;AF=0.3168;   GT:DP:RO:AO:AF
0/1:704:481:223:0.3168
chr1    83230081        .       C       G       .       PASS
TYPE=SNV;SCORE=0.9293;DP=464;RO=271;AO=192;AF=0.4138;   GT:DP:RO:AO:AF
0/1:464:271:192:0.4138
chr1    87329086        .       C       T       .       REJECT
TYPE=SNV;SCORE=0.3242;DP=593;RO=252;AO=340;AF=0.5734;   GT:DP:RO:AO:AF
0/1:593:252:340:0.5734
chr1    242432018       .       C       T       .       PASS
TYPE=SNV;SCORE=0.9999;DP=1646;RO=721;AO=925;AF=0.562;   GT:DP:RO:AO:AF
0/1:1646:721:925:0.562
chr1    143775490       .       A       G       .       PASS
TYPE=SNV;SCORE=0.7586;DP=1393;RO=910;AO=482;AF=0.346;   GT:DP:RO:AO:AF
0/1:1393:910:482:0.346
chr1    229871997       .       G       A       .       PASS
TYPE=SNV;SCORE=0.9999;DP=620;RO=508;AO=111;AF=0.179;    GT:DP:RO:AO:AF
0/1:620:508:111:0.179
chr1    19209141        .       C       T       .       PASS
TYPE=SNV;SCORE=0.9998;DP=809;RO=1;AO=808;AF=0.9988;     GT:DP:RO:AO:AF
0/1:809:1:808:0.9988
chr1    236752160       .       G       T       .       PASS
TYPE=SNV;SCORE=0.867;DP=585;RO=541;AO=44;AF=0.0752;     GT:DP:RO:AO:AF
0/1:585:541:44:0.0752
chr1    149808218       .       A       T       .       PASS
TYPE=SNV;SCORE=0.9263;DP=981;RO=573;AO=408;AF=0.4159;   GT:DP:RO:AO:AF
0/1:981:573:408:0.4159
chr1    68773236        .       C       G       .       PASS
TYPE=SNV;SCORE=0.9993;DP=719;RO=452;AO=267;AF=0.3713;   GT:DP:RO:AO:AF
0/1:719:452:267:0.3713
chr1    75106326        .       C       A       .       PASS
TYPE=SNV;SCORE=0.9981;DP=445;RO=0;AO=445;AF=1.0;        GT:DP:RO:AO:AF
0/1:445:0:445:1.0
chr1    169375018       .       T       G       .       PASS
TYPE=SNV;SCORE=0.6743;DP=1532;RO=941;AO=591;AF=0.3858;  GT:DP:RO:AO:AF
0/1:1532:941:591:0.3858
```

```
chr1    102718461       .       G       C       .       PASS
TYPE=SNV;SCORE=0.5455;DP=8;RO=5;AO=3;AF=0.375;  GT:DP:RO:AO:AF  0/1:8:5:3:0.375
chr1    158128325       .       G       C       .       PASS
TYPE=SNV;SCORE=0.9982;DP=596;RO=197;AO=399;AF=0.6695;   GT:DP:RO:AO:AF
0/1:596:197:399:0.6695
chr1    2101556 .       C       G       .       PASS
TYPE=SNV;SCORE=0.9897;DP=245;RO=204;AO=41;AF=0.1673;    GT:DP:RO:AO:AF
0/1:245:204:41:0.1673
chr1    2298602 .       A       C       .       REJECT
TYPE=SNV;SCORE=0.3879;DP=748;RO=677;AO=71;AF=0.0949;    GT:DP:RO:AO:AF
0/1:748:677:71:0.0949
chr1    91578606        .       G       T       .       PASS
TYPE=SNV;SCORE=0.9996;DP=830;RO=214;AO=613;AF=0.7386;   GT:DP:RO:AO:AF
0/1:830:214:613:0.7386
chr1    96802788        .       G       C       .       PASS
TYPE=SNV;SCORE=0.9962;DP=243;RO=208;AO=35;AF=0.144;     GT:DP:RO:AO:AF
0/1:243:208:35:0.144
chr1    203517608       .       C       A       .       PASS
TYPE=SNV;SCORE=0.8722;DP=2808;RO=2565;AO=243;AF=0.0865; GT:DP:RO:AO:AF
0/1:2808:2565:243:0.0865
chr1    143621384       .       A       G       .       PASS
TYPE=SNV;SCORE=0.9951;DP=696;RO=220;AO=476;AF=0.6839;   GT:DP:RO:AO:AF
0/1:696:220:476:0.6839
chr1    227880191       .       GAA     G       .       PASS
TYPE=INDEL;SCORE=0.9704;DP=968;RO=559;AO=408;AF=0.4215; GT:DP:RO:AO:AF
0/1:968:559:408:0.4215
chr1    158128509       .       CT      C       .       PASS
TYPE=INDEL;SCORE=0.8445;DP=689;RO=624;AO=63;AF=0.0914;  GT:DP:RO:AO:AF
0/1:689:624:63:0.0914
chr1    235532297       .       CAT     C       .       PASS
TYPE=INDEL;SCORE=0.7107;DP=1030;RO=908;AO=114;AF=0.1107;        GT:DP:RO:AO:AF
0/1:1030:908:114:0.1107
chr1    19209119        .       GTAACAAATAGCAATTTT      G       .       REJECT
TYPE=INDEL;SCORE=0.3404;DP=810;RO=4;AO=805;AF=0.9938;   GT:DP:RO:AO:AF
0/1:810:4:805:0.9938
```

The output of the VarNet model is the HCC1395.vcf file, which contains the somatic variants detected between the tumor and normal samples. In total, 28 variants were identified, of which 3 were rejected by the model's prediction. We filtered out the rejected variants, leaving 25 high-confidence somatic variants.

These filtered variants can then be cross-referenced with public databases (e.g., COSMIC, dbSNP, ClinVar) to determine whether they have been previously reported. Further biological analysis can also be conducted to assess their potential functional impact and relevance to cancer.

[48]:
```
grep -v '^#' HCC1395.vcf | wc -l
```

```
28
```

[49]:
```
grep -v '^#' HCC1395.vcf | awk '$7 == "PASS"' | wc -l
```

25

For each of the 25 filtered variants, we queried the dbSNP database to determine whether they had been previously reported. The search was performed using three key fields:

Organism: Homo sapiens Chromosome Number: Chromosome 1 Base Position: The genomic position of the variant

If a match was found in dbSNP, we recorded the corresponding rsID (Reference SNP ID). This step allowed us to annotate known variants and distinguish them from potentially novel mutations, which can be critical for downstream biological interpretation and clinical relevance.

[50]:
```
ls
```

```
HCC1395.vcf  candidates  predictions
```

After querying the 25 high-confidence variants in the dbSNP database using the specified fields (organism, chromosome number, and base position), we found that 4 variants had been previously reported. For these known variants, we retrieved and recorded their corresponding rsIDs.

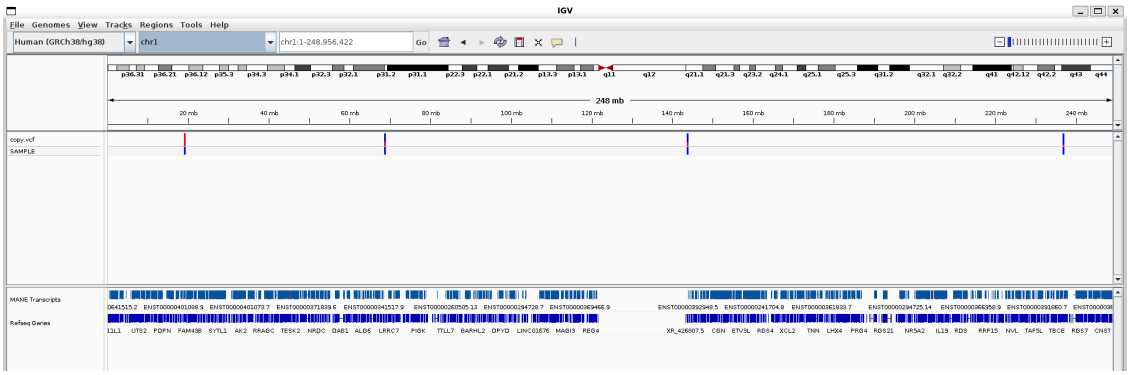
To facilitate further analysis, we created a separate VCF file containing only these 4 annotated variants along with their rsIDs.

[51]:
```
ls
```

```
HCC1395.vcf  candidates  predictions
selected.vcf
```

[52]:
```
cat selected.vcf
```

```
chr1    236923385       rs1667318397       A       G       .       PASS
TYPE=SNV;SCORE=0.6584;DP=718;RO=599;AO=119;AF=0.1657;   GT:DP:RO:AO:AF
0/1:718:599:119:0.1657
chr1    143775490       rs1165169713       A       G       .       PASS
TYPE=SNV;SCORE=0.7586;DP=1393;RO=910;AO=482;AF=0.346;   GT:DP:RO:AO:AF
0/1:1393:910:482:0.346
chr1    19209141         rs534493951       C       T       .       PASS
TYPE=SNV;SCORE=0.9998;DP=809;RO=1;AO=808;AF=0.9988;     GT:DP:RO:AO:AF
0/1:809:1:808:0.9988
chr1    68773236        rs1646637309       C       G       .       PASS
TYPE=SNV;SCORE=0.9993;DP=719;RO=452;AO=267;AF=0.3713;   GT:DP:RO:AO:AF
0/1:719:452:267:0.3713
```



Although none of the 25 detected variants were found to have documented clinical significance in databases such as ClinVar, we selected one variant for further annotation and biological analysis.

rs534493951:

Welcome to the Reference SNP (rs) Report
All alleles are reported in the Forward orientation. Click on the Variant Details tab for details on Genomic Placement, Gene, and Amino Acid changes. HGVS names are in the HGVS tab.

Reference SNP (rs) Report

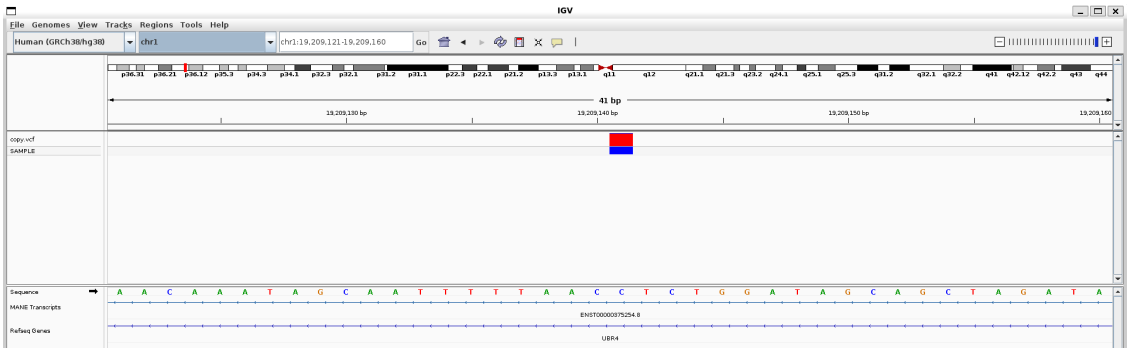
Download

rs534493951

Current Build 157
Released September 3, 2024

Organism	*Homo sapiens*	**Clinical Significance**	Not Reported in ClinVar
Position	chr1:19209141 (GRCh38.p14)	**Gene : Consequence**	UBR4 : Intron Variant EMC1-AS1 : 2KB Upstream Variant
Alleles	C>T	**Publications**	0 citations
Variation Type	SNV Single Nucleotide Variation	**Genomic View**	See rs on genome
Frequency	T=0.000013 (2/149250, GnomAD_genomes) T=0.0002 (1/6404, 1000G_30X) T=0.0002 (1/5008, 1000G) (+ 2 more)		

The variant rs534493951 is located within an intronic region of the UBR4 gene and lies approximately 2 kilobases upstream of the EMC1-AS1 gene. While there is currently no confirmed clinical association with breast cancer reported in databases such as ClinVar, the genomic context suggests potential regulatory relevance. UBR4 encodes a ubiquitin ligase involved in protein quality control, and dysregulation of the ubiquitin-proteasome system has been implicated in several cancers, including breast cancer. EMC1-AS1, a long non-coding RNA, may also play a role in gene expression regulation in nearby regions. Although the functional impact of rs534493951 remains uncertain, its position within regulatory and non-coding regions warrants further investigation, particularly in the context of transcriptional control and epigenetic modulation in breast cancer biology.



2 VarNet CNN Architecture Summary

VarNet uses deep convolutional neural networks to call somatic mutations directly from raw sequencing data. Two models were built: one for SNVs and one for indels.

2.1 Input Encoding

Sequencing reads from tumor and matched normal samples are converted into 5-channel image-like tensors encoding: - Base identity
- Base quality
- Mapping quality
- Strand bias
- Reference base

Shapes: - **SNVs**: `(100, 70, 5)` over a 30-bp window; candidate site repeated 5×
- **Indels**: `(140, 150, 5)` over a 70-bp window; variable-length indels encoded in-place

2.2 SNV Model

- Custom CNN with 10 convolutional blocks:
  - Conv → ReLU → BatchNorm
- Two average-pooling layers
- Three dense layers: 256 → 128 → 64 units
- Final sigmoid output layer
- ~3.5 million trainable parameters

2.3 Indel Model

- Based on **InceptionV3** to capture complex patterns
- Supports longer context due to indel variability

2.4 Training Configuration

- **Optimizer**: Adam (`lr=1e-4`)
- **Batch size**: 32
- **Framework**: TensorFlow
- **Hardware**: Nvidia Titan-X GPU

VarNet learns mutation-relevant features directly from alignments, eliminating the need for hand-crafted filters and enabling broad generalization across cancer types and sequencing platforms.

[ ]: