DEPARTMENT OF COMPUTER APPLICATIONS

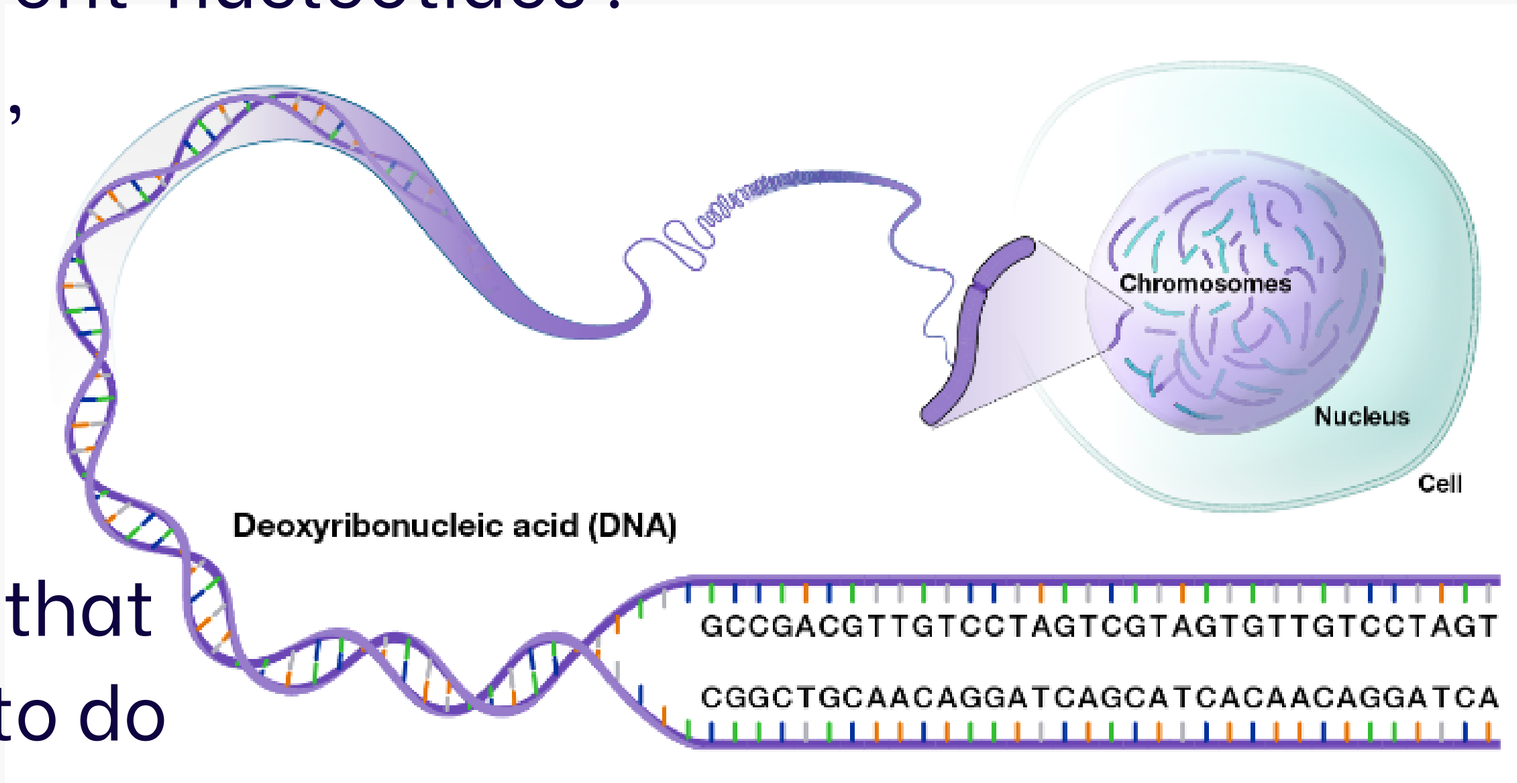
MINI-PROJECT
VARIANT CALLING IN GENOMIC SEQUENCING DATA USING DEEP LEARNING

PRESENTED BY
ABIJITH T K
DEVIKA MOHAN
AJITH PRASAD
SHIBU C

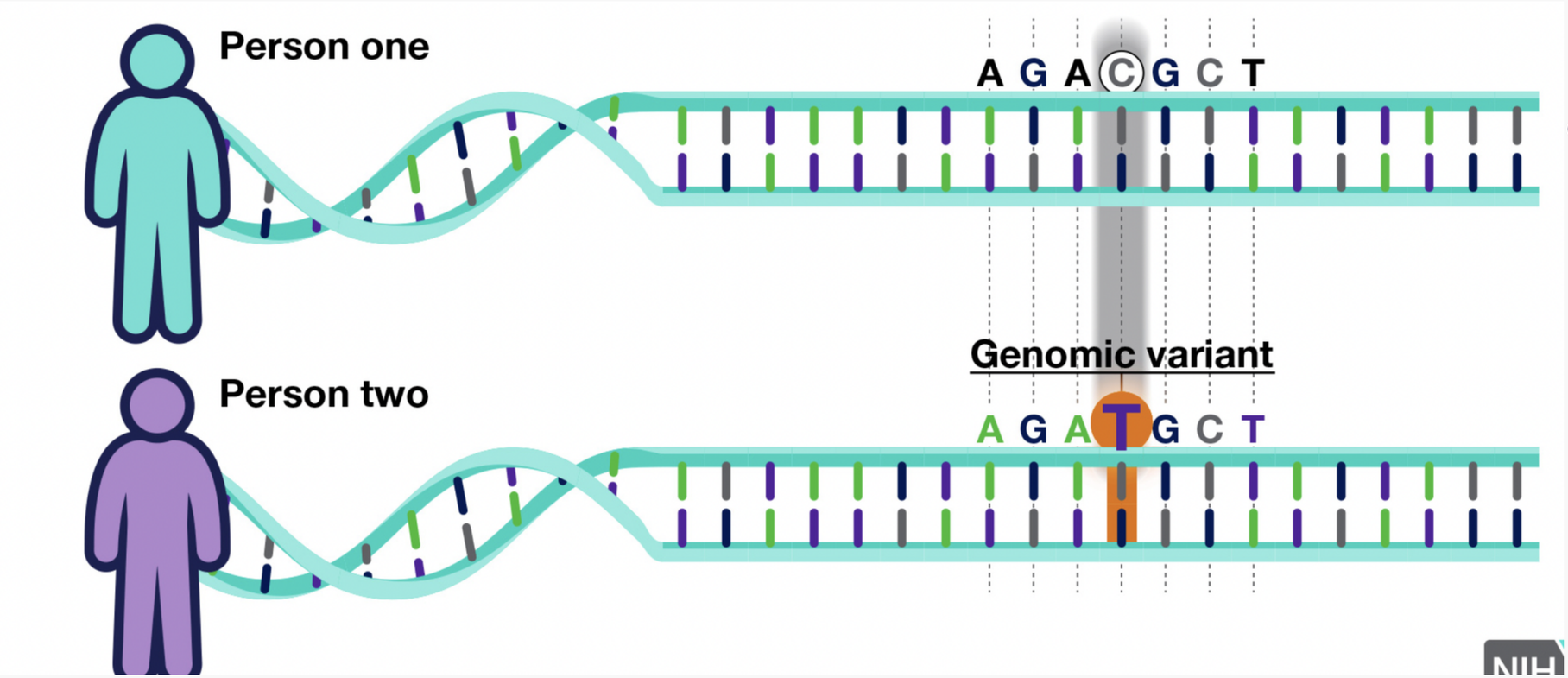
30-11-2023

HUMAN GENOME

- A genome is the complete set of DNA instructions found in every cell
- DNA is made of four different nucleotides :
  adenine (A), thymine (T),
  cytosine (C) and
  guanine (G).
- The order of these letters
  (i.e., the DNA sequence)
  encodes the information that
  instructs each cell what to do
  and when to do it.



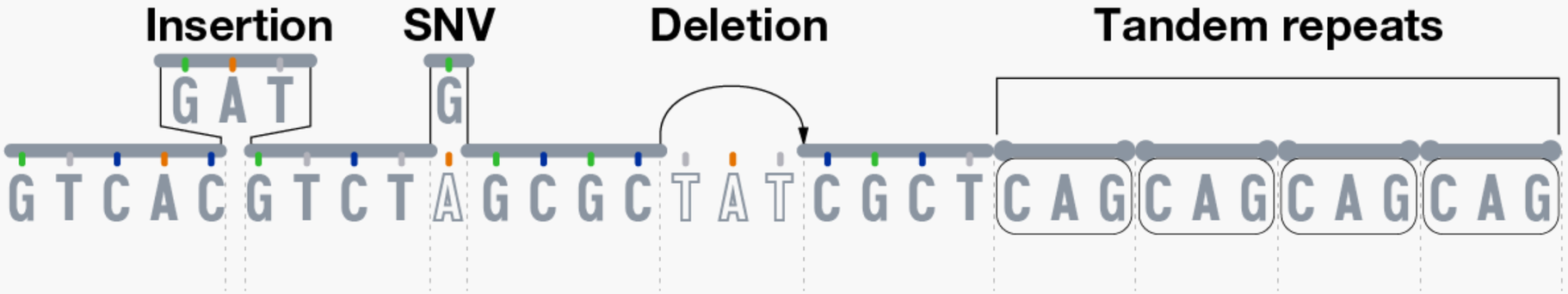
Human Genomic Variation



- The vast majority of the DNA letters in peoples' genomes is identical, but a small fraction of those letters varies

- This genomic variation accounts for some of the differences among people, including important aspects of their health and susceptibility to diseases.

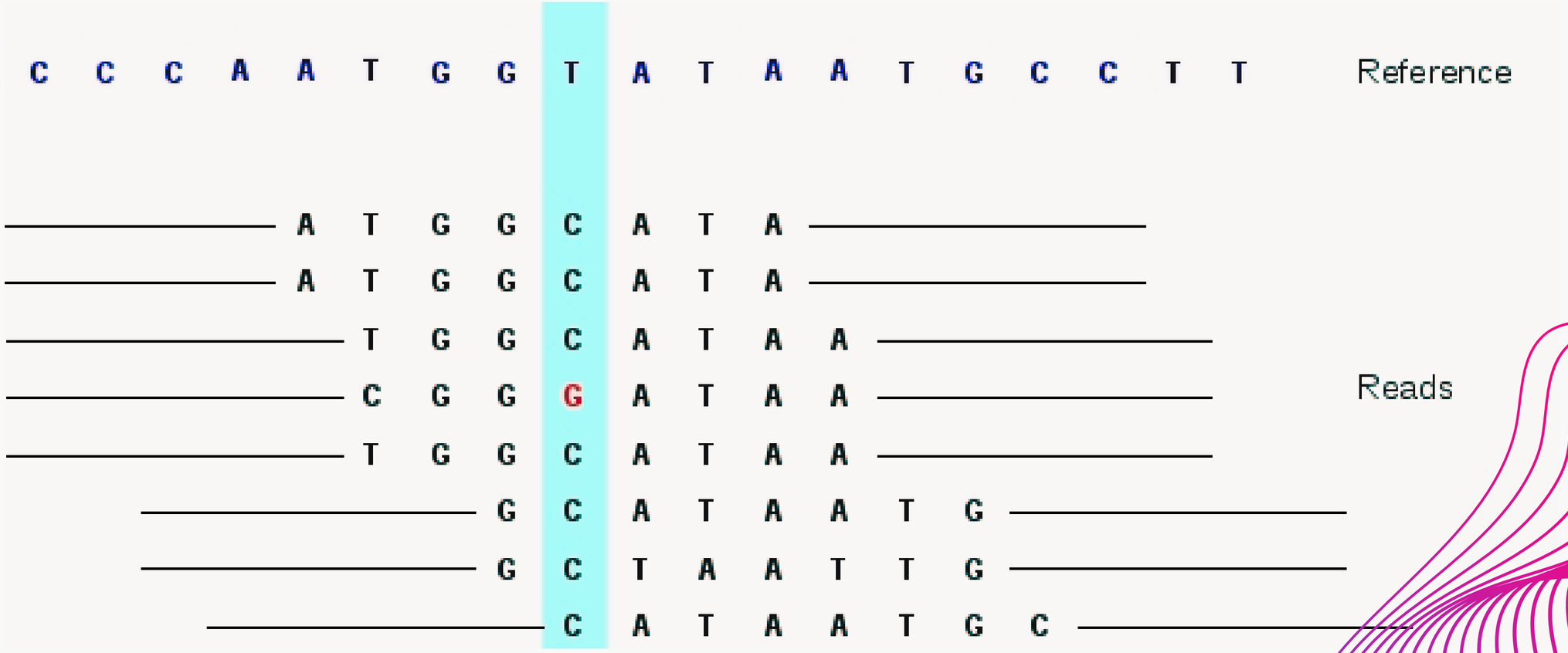
Different Genomic Variations



- The smallest genomic variants are single-nucleotide variants (SNVs). Each SNV reflects a difference in a single nucleotide (or letter).
    - An insertion is a variation in which a specific nucleotide sequence is present in DNA
    - A deletion is a type of mutation that involves the loss of one or more nucleotides from a segment of DNA

Variant Calling

- Align the sequences to a reference genome
- Identify where the aligned reads differ from the reference genome



Limitations of traditional variant calling methods

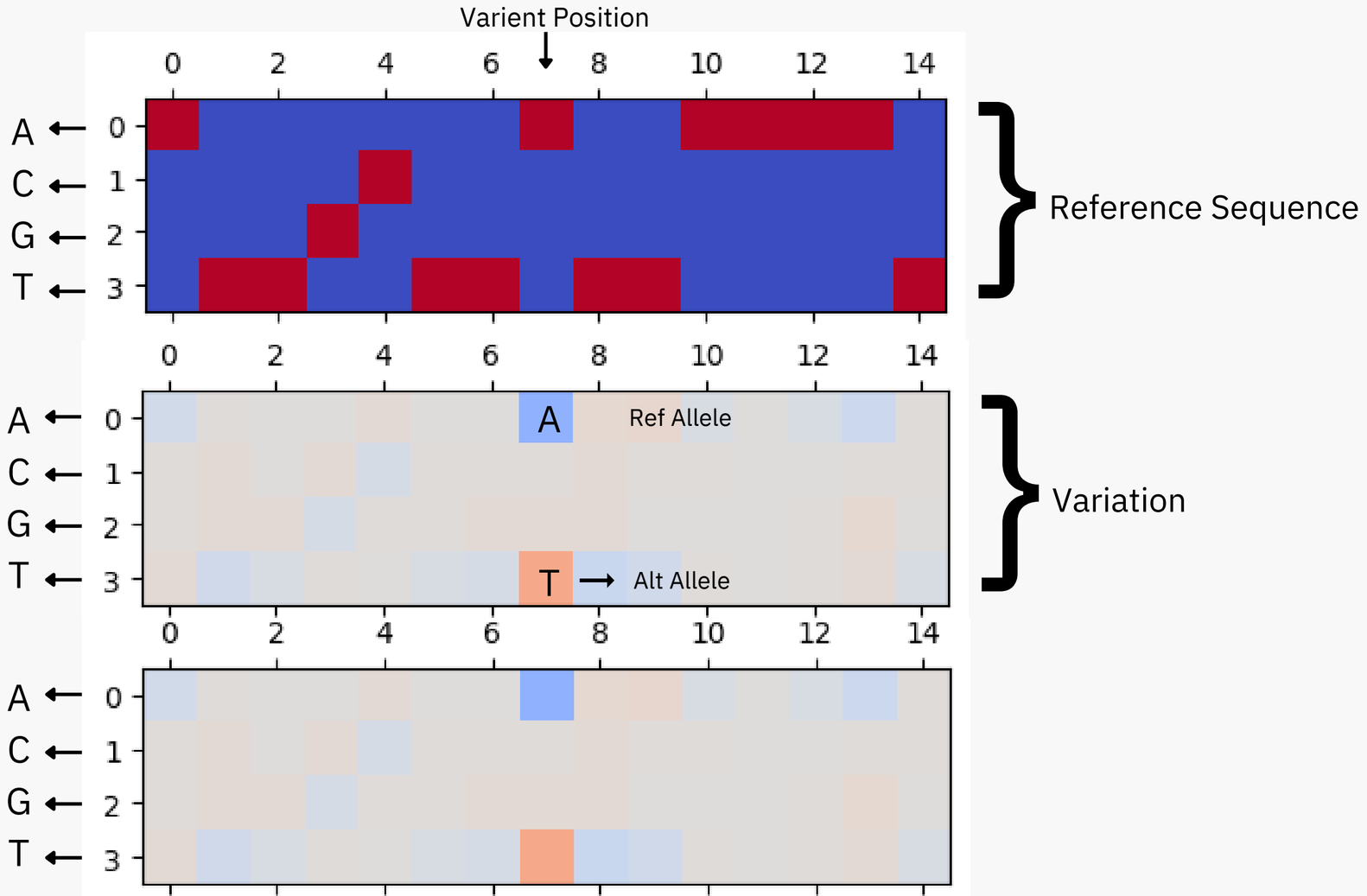
- **Inaccurate for complex variants:** Traditional methods struggle to identify and classify complex variants like indels and multi-allelic variants.
- **Prone to sequencing errors:** Traditional methods are susceptible to false positive variant calls due to sequencing errors.
- **Limited to specific variant types:** Traditional methods are often tailored to specific variant types, limiting their applicability.
- **Scalability issues with large datasets:** Traditional methods can become inefficient when processing large volumes of sequencing data.

How the model overcome the issues

- **Inaccurate for complex variants** The model uses a deep learning approach that is able to learn complex patterns in the data.
- **Prone to sequencing errors** The model uses a variety of techniques to filter out sequencing errors, such as using quality scores and base calling consensus.
- **Limited to specific variant types** The model is able to handle a wide range of variant types, including SNPs, indels, and multi-allelic variants.
- **Scalability issues with large datasets** The model is able to efficiently process large volumes of sequencing data.

How we identified the variants

- The alignments are converted to three 15 by 4 matrices for training the network and calling variants.



- Encode all alignments to a 15*4*3 tensor

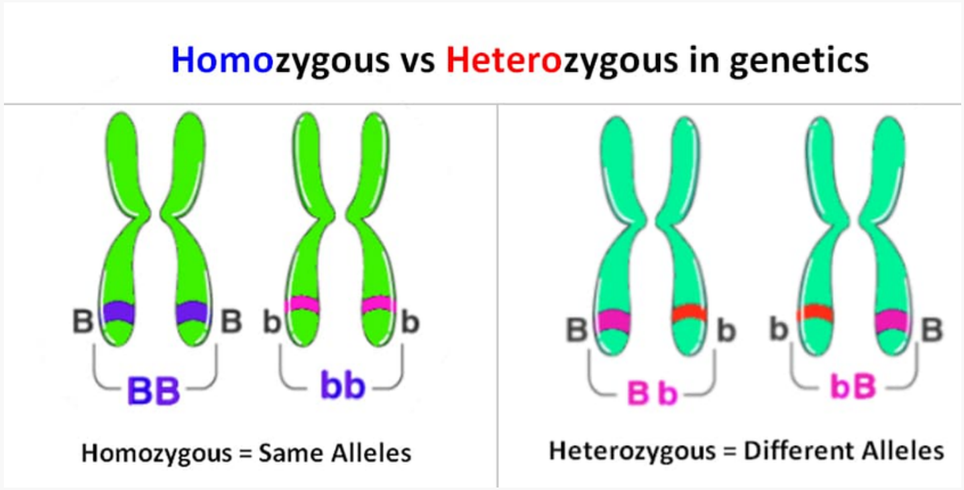
- we train the neural network and classify the called variants into four categories: homozygous variant , heterozygous variant, non-variant or complex varient.
- It is also trained to predict the possible varient base

Yarray - (n, 8) ⟶ [0. 0.5 0. 0.5] [0. 0. 0. 1.]

Genotype (A, C, G, T)

Variant Type (Homozygous, Heterozygous, Non Varient and Complex Varient)

Homozygous variant : A variant that is present on both copies of a chromosome in an individual.



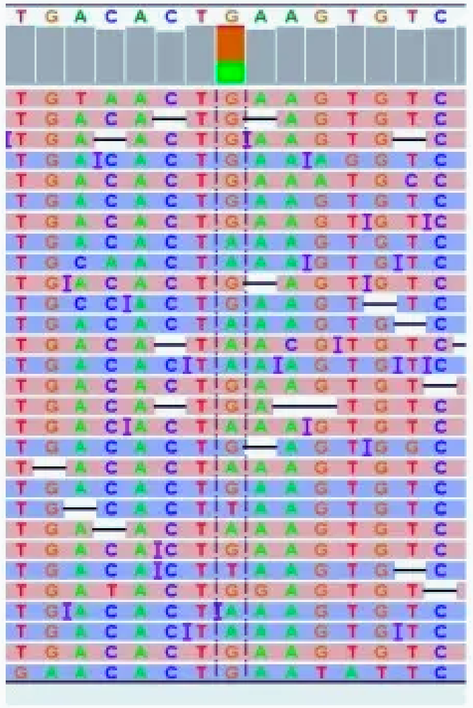
Heterozygous variant : A variant that is present on only one copy of a chromosome in an individual.

Non-variant : A variant that is not present in any of the individuals in the sample that was being sequenced.

Complex variant : A variant that is difficult to classify into one of the other categories.

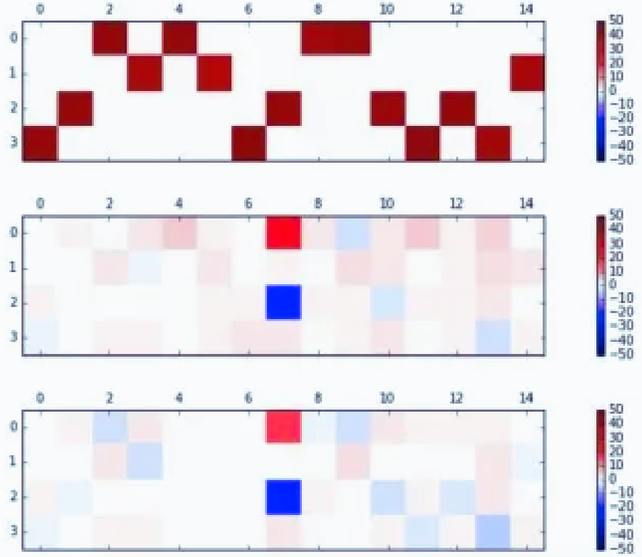
Variant Calling Flowchart

Sequence alignments



Each candidate +/- 7 bp

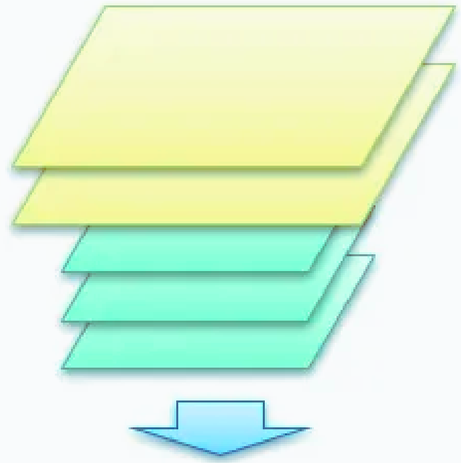
Alignment Tensor



Encode all alignments
to a 15 x 4 x 3 tensor

2 convolution layers

4 full connected layers



Genotype

[0.5, 0.0, 0.5, 0.0]

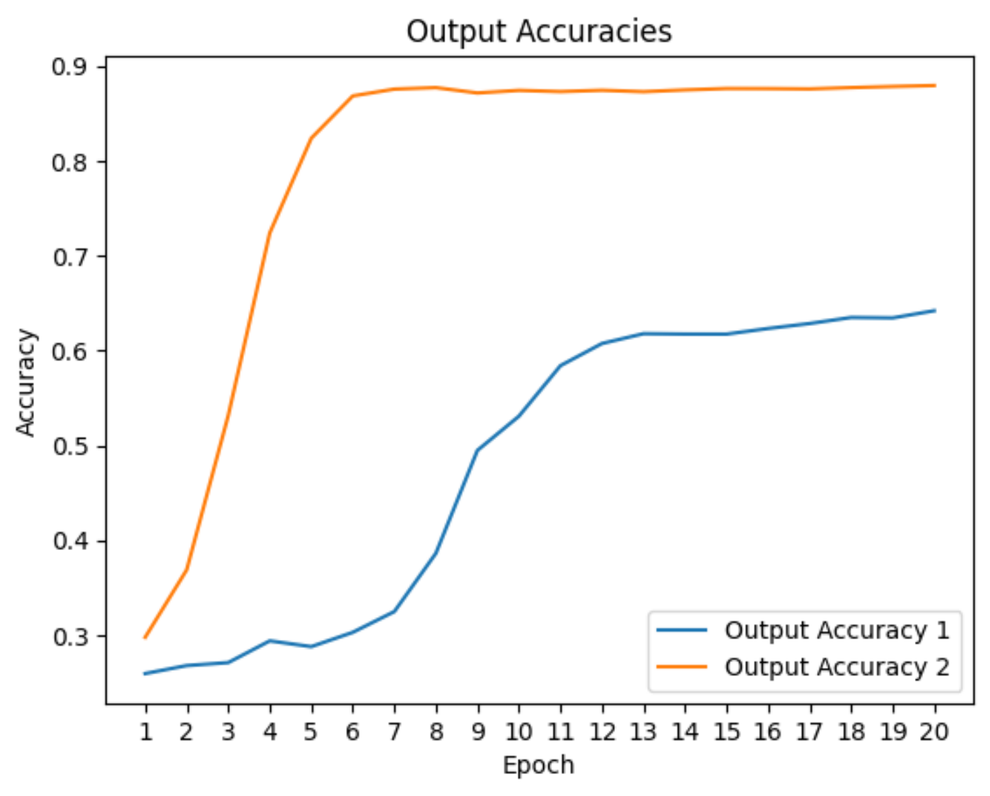
A C G T

Softmax output

[0.98, 0, 0, 0.02]

het none hom complex

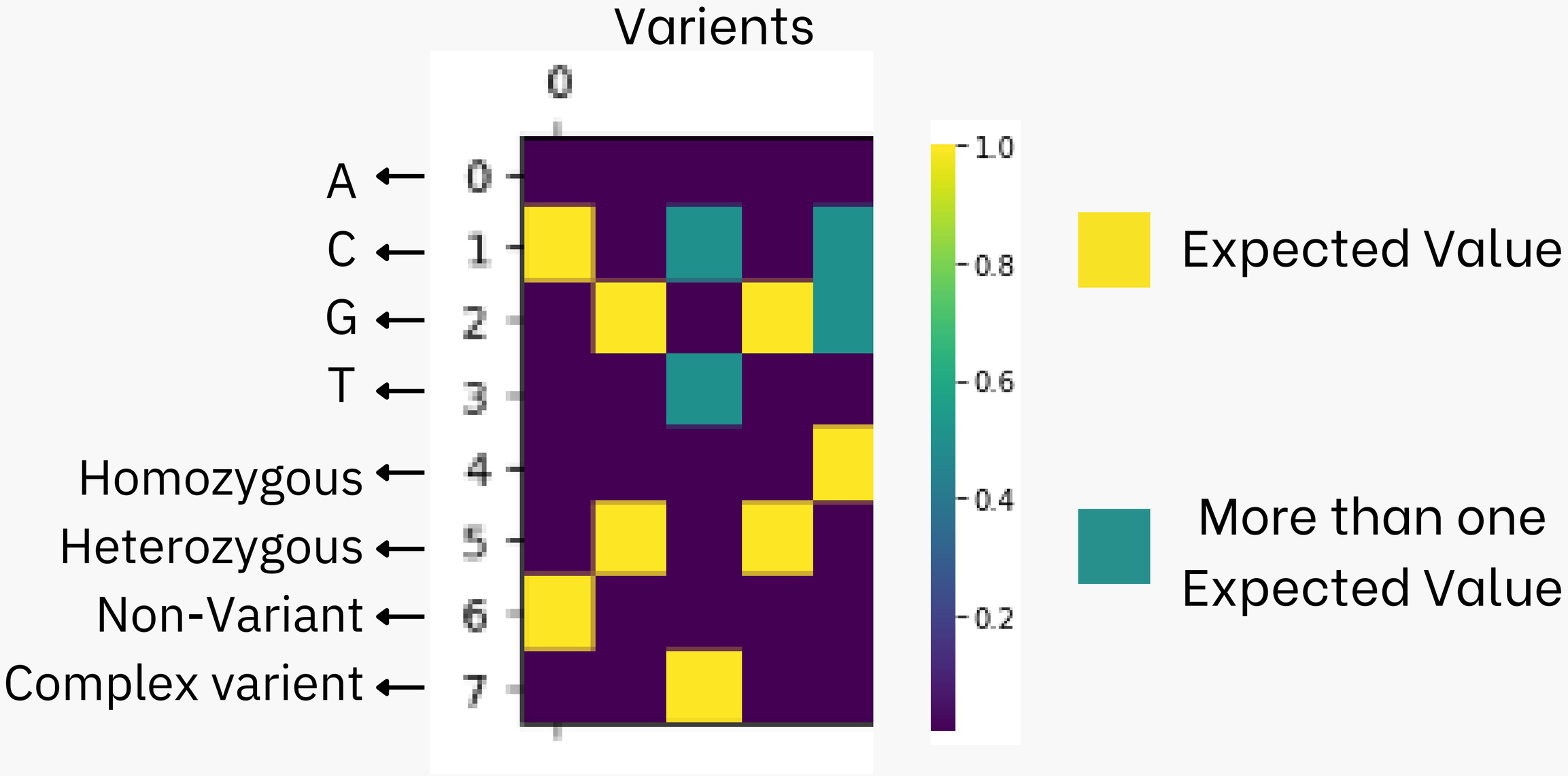
Model Accuracy



Output 1 - Bases(A/C/G/T)

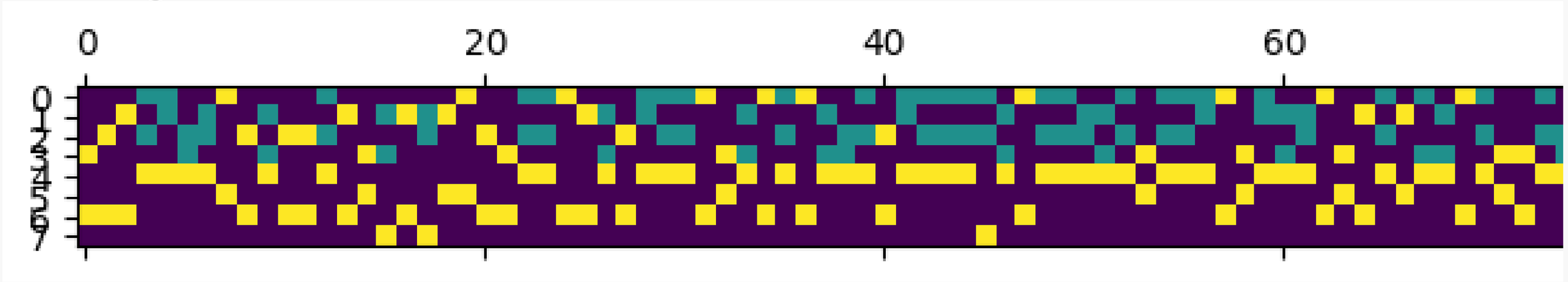
Output 2 - Varient Type (hom/het/non/com)

Understanding the Predictions

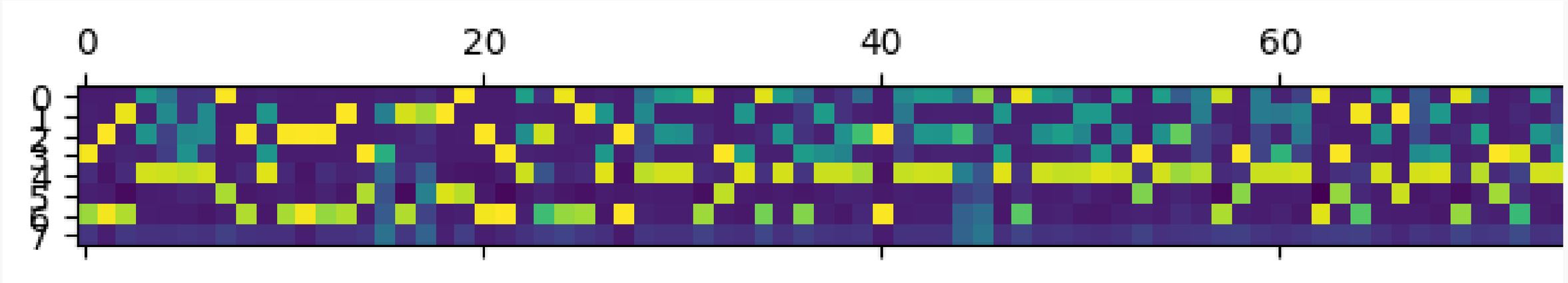


Predictions

Original variant call



Predicted variant call



Classification report

	precision	recall	f1-score	support
0	0.96	0.97	0.97	21809
1	0.96	0.99	0.97	9691
2	0.94	0.98	0.98	12609
3	0.90	0.90	0.54	3125
accuracy			0.95	47234
acro avg	0.94	0.87	0.89	47234
weighted avg	0.95	0.95	0.95	47234

THANK YOU