Bioinformatics for Beginners

Genomics Practice

Tugce Bilgin Sonay, ZHAW, May 2023

The Study:

PNAS

Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA

Boris Rebolledo-Jaramillo[a,1], Marcia Shu-Wei Su[b,1], Nicholas Stoler[a], Jennifer A. McElhoe[c], Benjamin Dickins[d], Daniel Blankenberg[a], Thorfinn S. Korneliussen[e,f], Francesca Chiaromonte[g], Rasmus Nielsen[e], Mitchell M. Holland[c], Ian M. Paul[h], Anton Nekrutenko[a,2], and Kateryna D. Makova[b,2]

Departments of [a]Biochemistry and Molecular Biology, [b]Biology, and [g]Statistics, [c]Forensic Science Program, Pennsylvania State University, University Park, PA 16802; [d]School of Science and Technology, Nottingham Trent University, Nottingham NG1 4BU, United Kingdom; [e]Department of Integrative Biology, University of California, Berkeley, CA 94720; [f]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, DK-1350 Copenhagen, Denmark; and [h]Department of Pediatrics, College of Medicine, Pennsylvania State University, Hershey, PA 17033

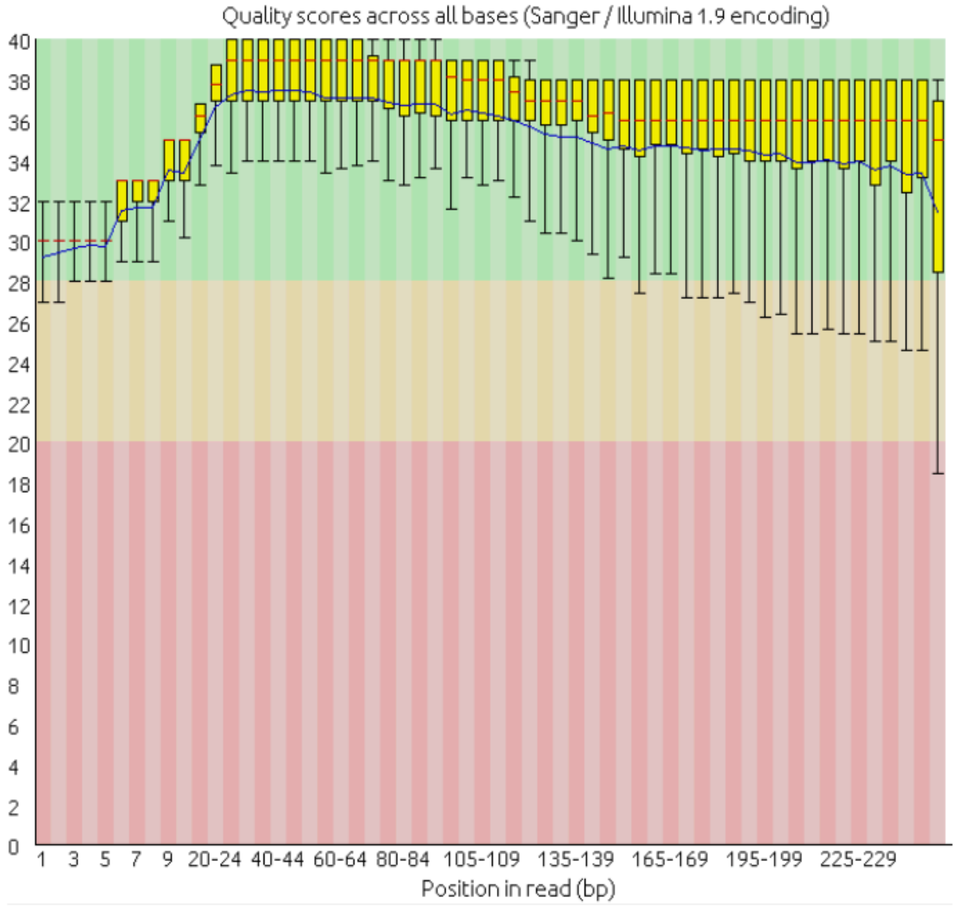
Edited by Michael Lynch, Indiana University, Bloomington, IN, and approved September 8, 2014 (received for review May 20, 2014)

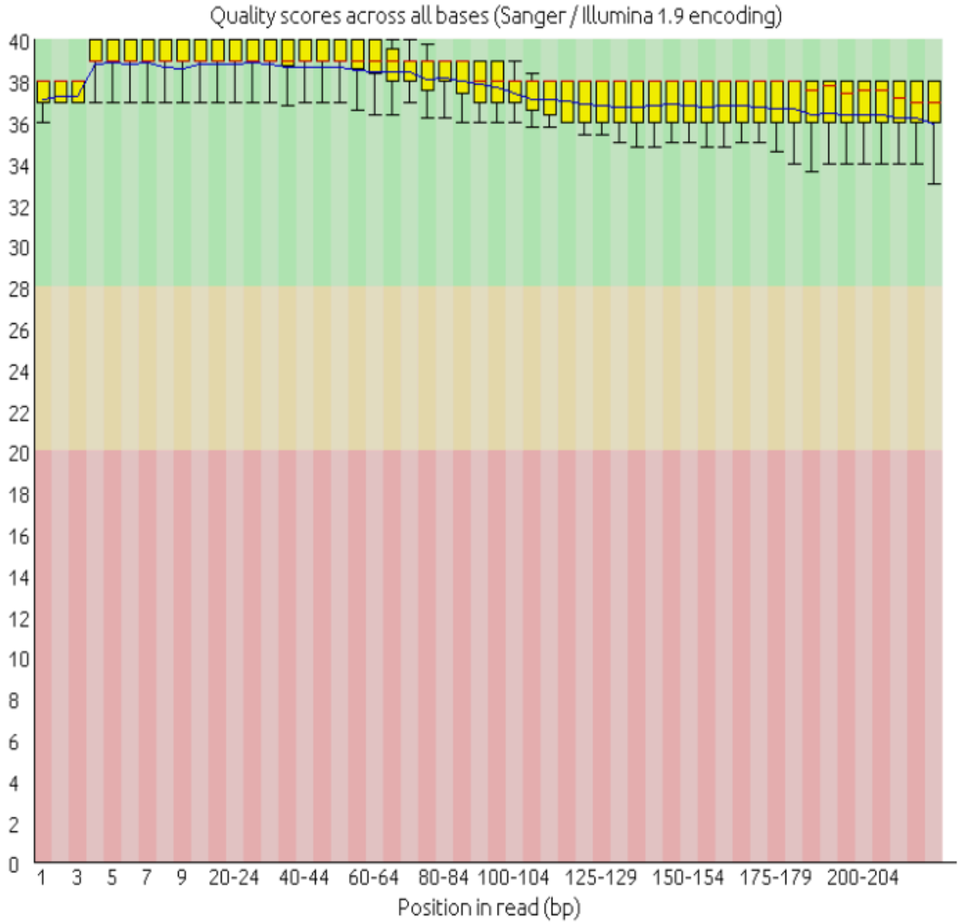
Filtering:

- Each base call has an associated base call quality
  - What is the chance that the base call is incorrect?
    - Illumina evidence: intensity values + cycle
  - Phred values (log scale)
    - Q10 = 1 in 10 chance of base call incorrect
    - Q20 = 1 in 100 chance of base call incorrect
  - Accurate base qualities essential measure in variant calling

- **Rule of thumb: Anything less than Q20 is not useful data**

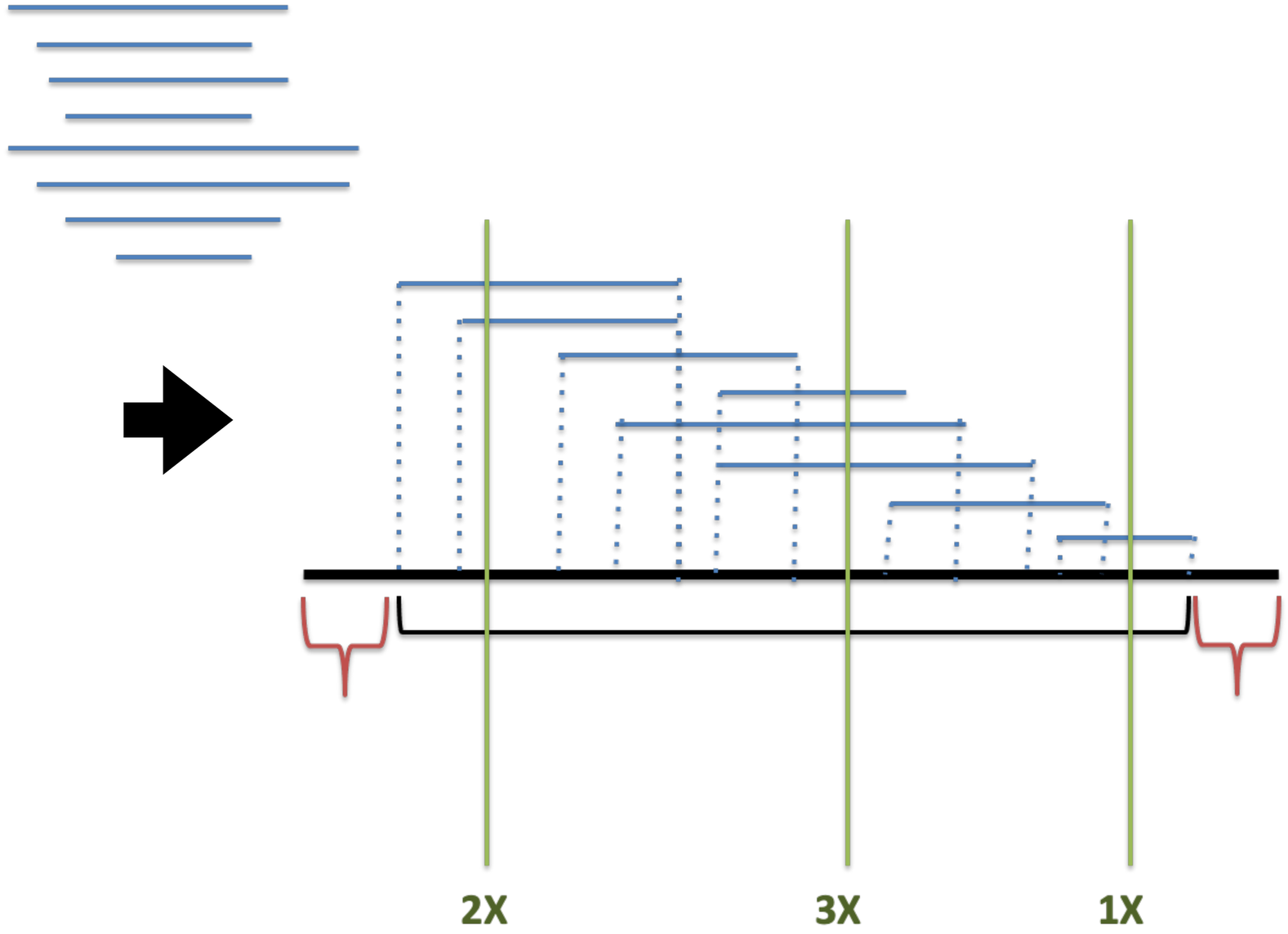
Before and After:



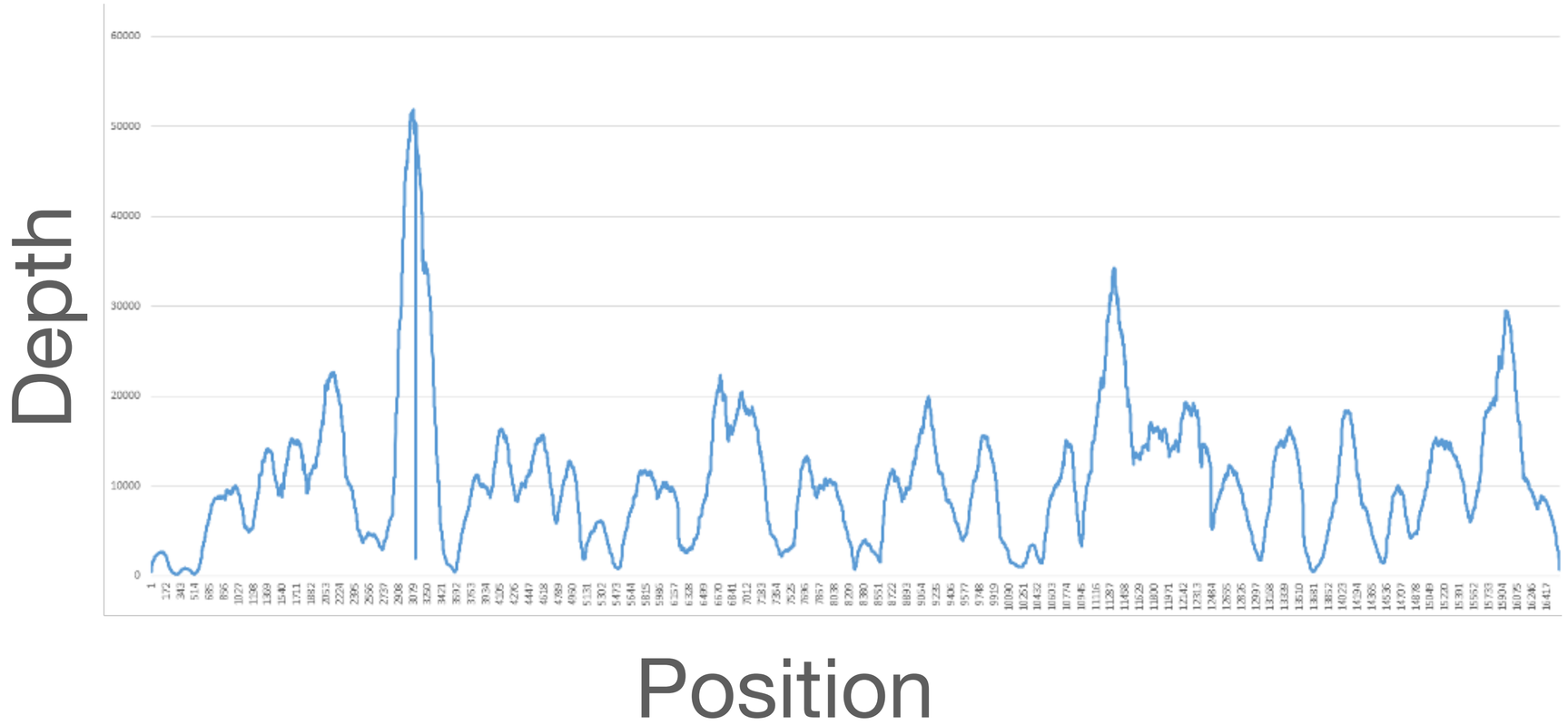


Alignment File:

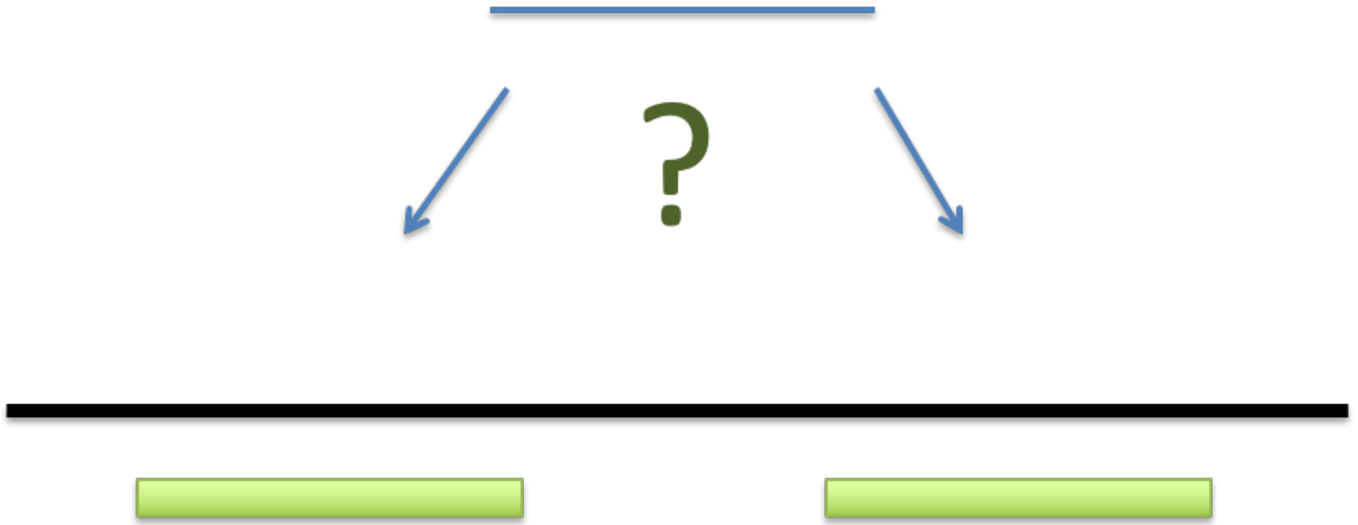
Coverage and Depth:



Depth:



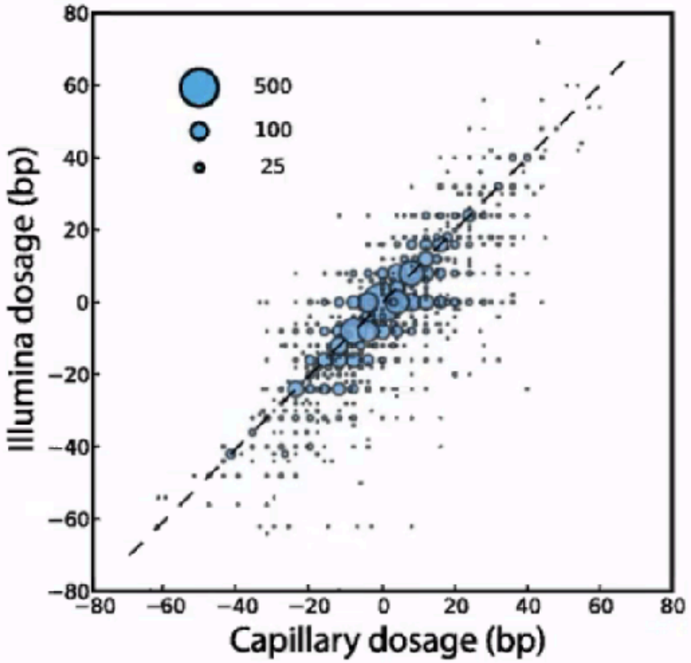
When there are repeats



STR calling has gotten better!

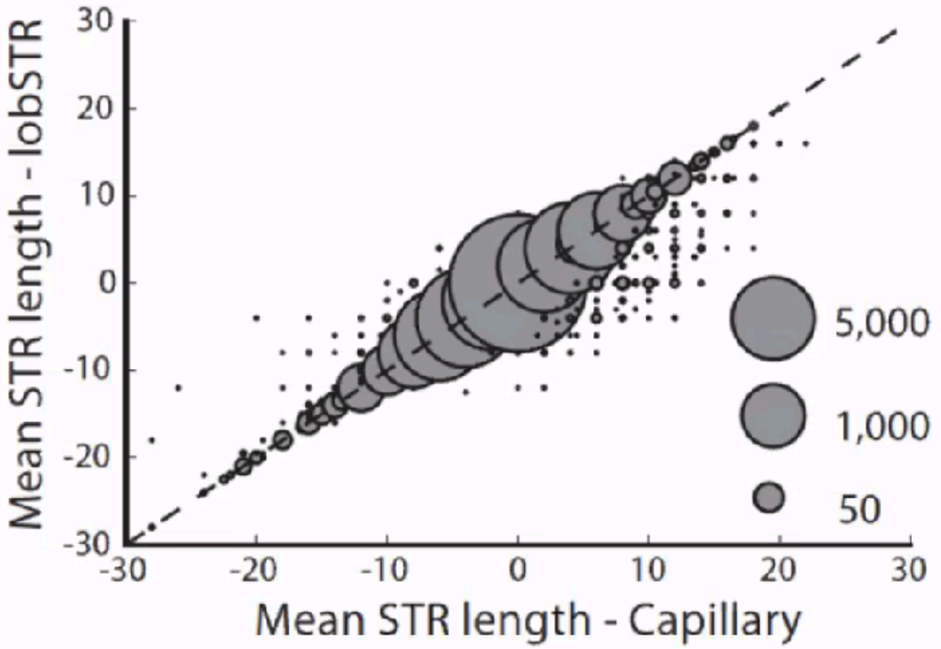
1000 Genomes (4x, lobSTR) (2014)

41% concordance with capillary data



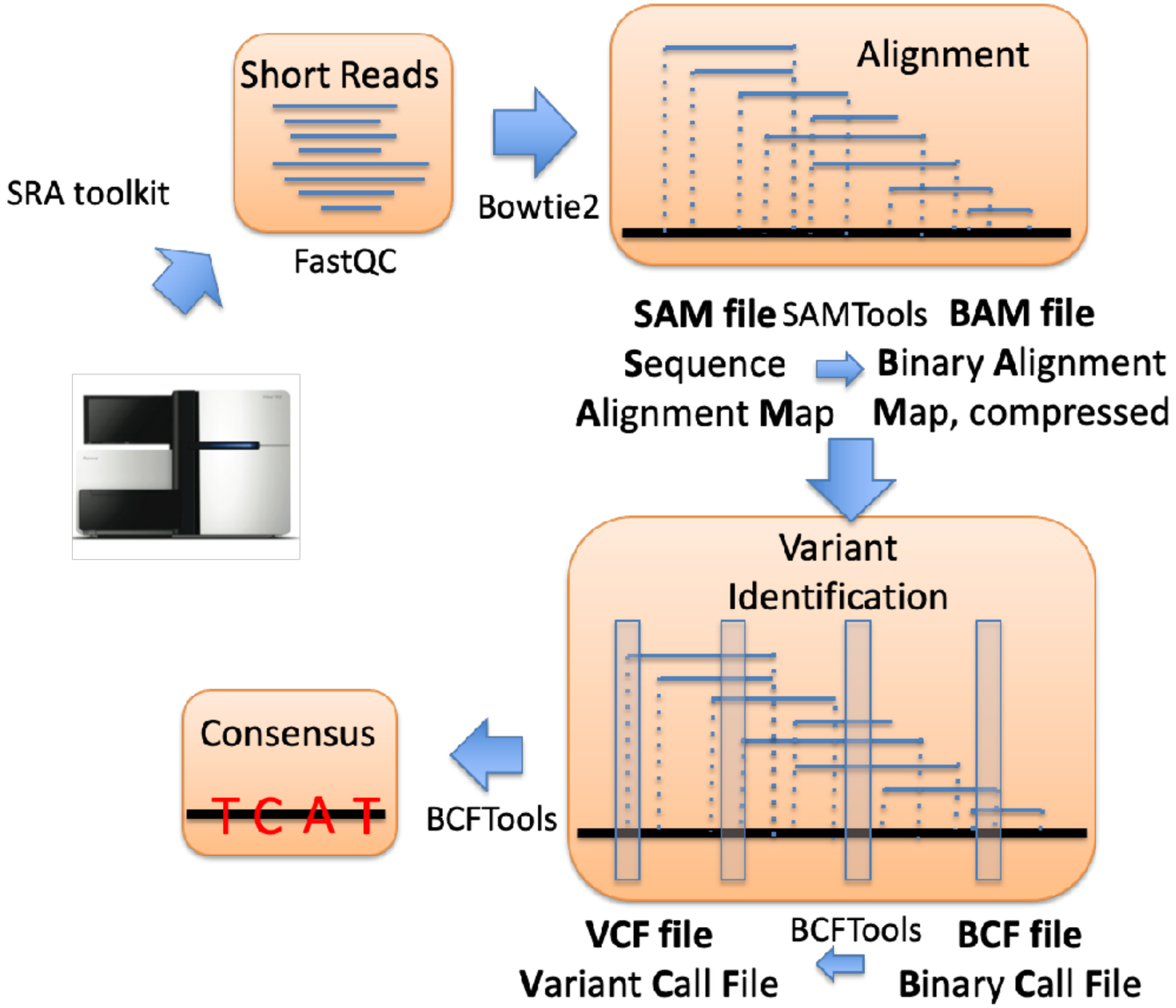
SGDP (30x, PCR free, lobSTR) (2016)

93% concordance with capillary data

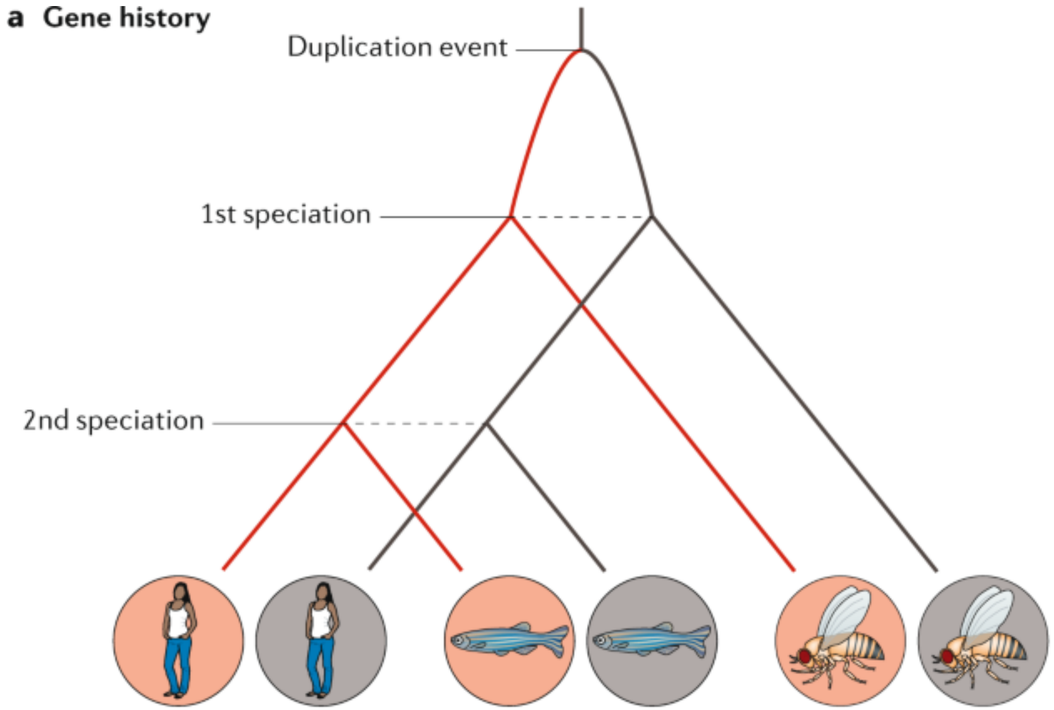


SGDP (30x, PCR free, **HipSTR**): 98.5% concordance on duplicated samples (2017)

Overview


Short Reads
SRA toolkit
FastQC
Bowtie2
Alignment
SAM file SAMTools BAM file
Sequence Binary Alignment
Alignment Map Map, compressed
Variant
Identification
Consensus
T C A T
BCFTools
VCF file
Variant Call File
BCFTools
BCF file
Binary Call File

Phylogenetic Trees



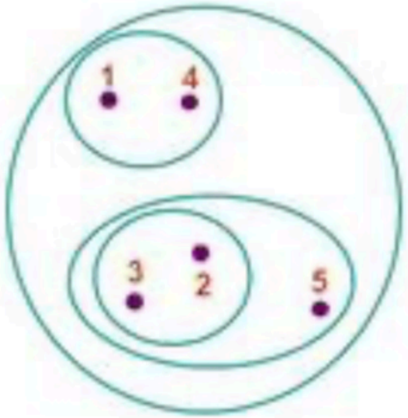
A Phylogeny (Phylogenetic tree) or Evolutionary tree represents the evolutionary relationships among

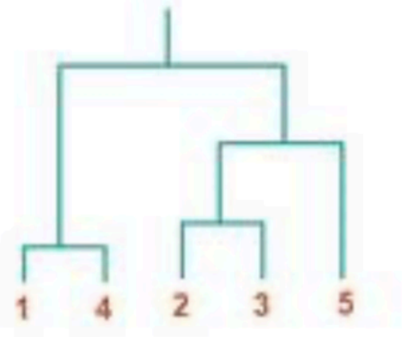
- a set of organisms or groups of organisms
- a family related nucleic acid or protein sequences

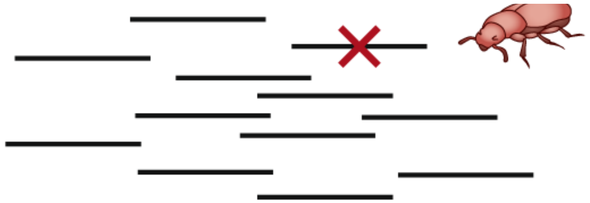
Every phylogenetic tree is an hypothesis about relationships

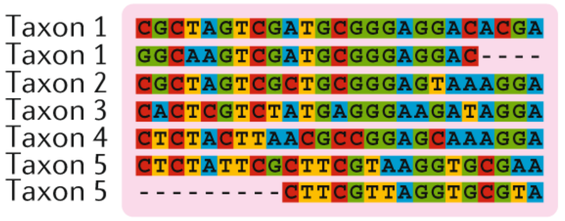
- some are well supported by data
- others are not

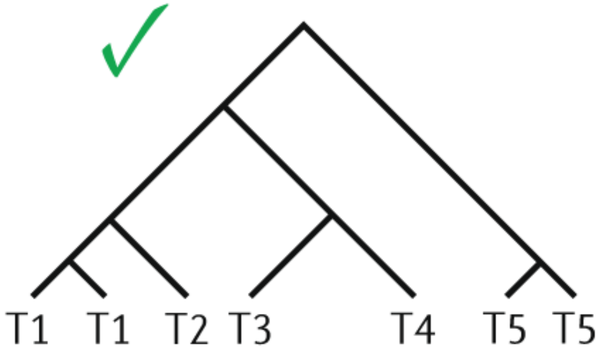
Phylogeny Tree Reconstruction











Dnaml -- DNA Maximum Likelihood program

This program implements the maximum likelihood method for DNA sequences. The present version is faster than earlier versions of Dnaml. Details of the algorithm are published in the paper by Felsenstein and Churchill (1996). The model of base substitution allows the expected frequencies of the four bases to be unequal, allows the expected frequencies of transitions and transversions to be unequal, and has several ways of allowing different rates of evolution at different sites.

The assumptions of the present model are:

1. Each site in the sequence evolves independently.
2. Different lineages evolve independently.
3. Each site undergoes substitution at an expected rate which is chosen from a series of rates (each with a probability of occurrence) which we specify.
4. All relevant sites are included in the sequence, not just those that have changed or those that are "phylogenetically informative".
5. A substitution consists of one of two sorts of events:

   (a)
   The first kind of event consists of the replacement of the existing base by a base drawn from a pool of purines or a pool of pyrimidines (depending on whether the base being replaced was a purine or a pyrimidine). It can lead either to no change or to a transition.

   (b)
   The second kind of event consists of the replacement of the existing base by a base drawn at random from a pool of bases at known frequencies, independently of the identity of the base which is being replaced. This could lead either to a no change, to a transition or to a transversion.

   The ratio of the two purines in the purine replacement pool is the same as their ratio in the overall pool, and similarly for the pyrimidines.

   The ratios of transitions to transversions can be set by the user. The substitution process can be diagrammed as follows: Suppose that you specified A, C, G, and T base frequencies of 0.24, 0.28, 0.27, and 0.21.

   - First kind of event:

     1. Determine whether the existing base is a purine or a pyrimidine.
     2. Draw from the proper pool:

```
Purine pool:                Pyrimidine pool:
```

Questions?

Genomics Practice

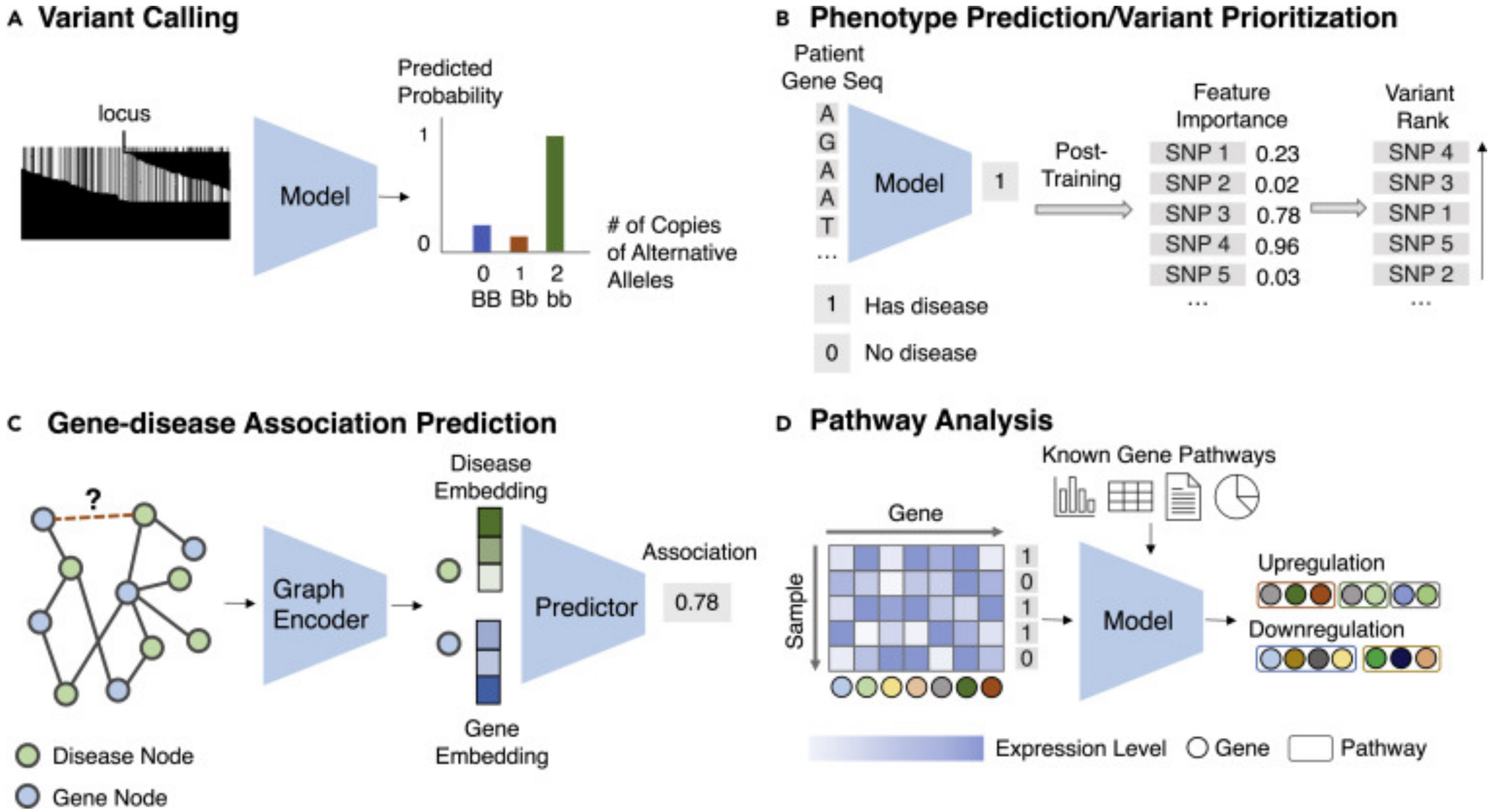
Tugce Bilgin Sonay, ZHAW, May 2023

Bioinformatics for Beginners

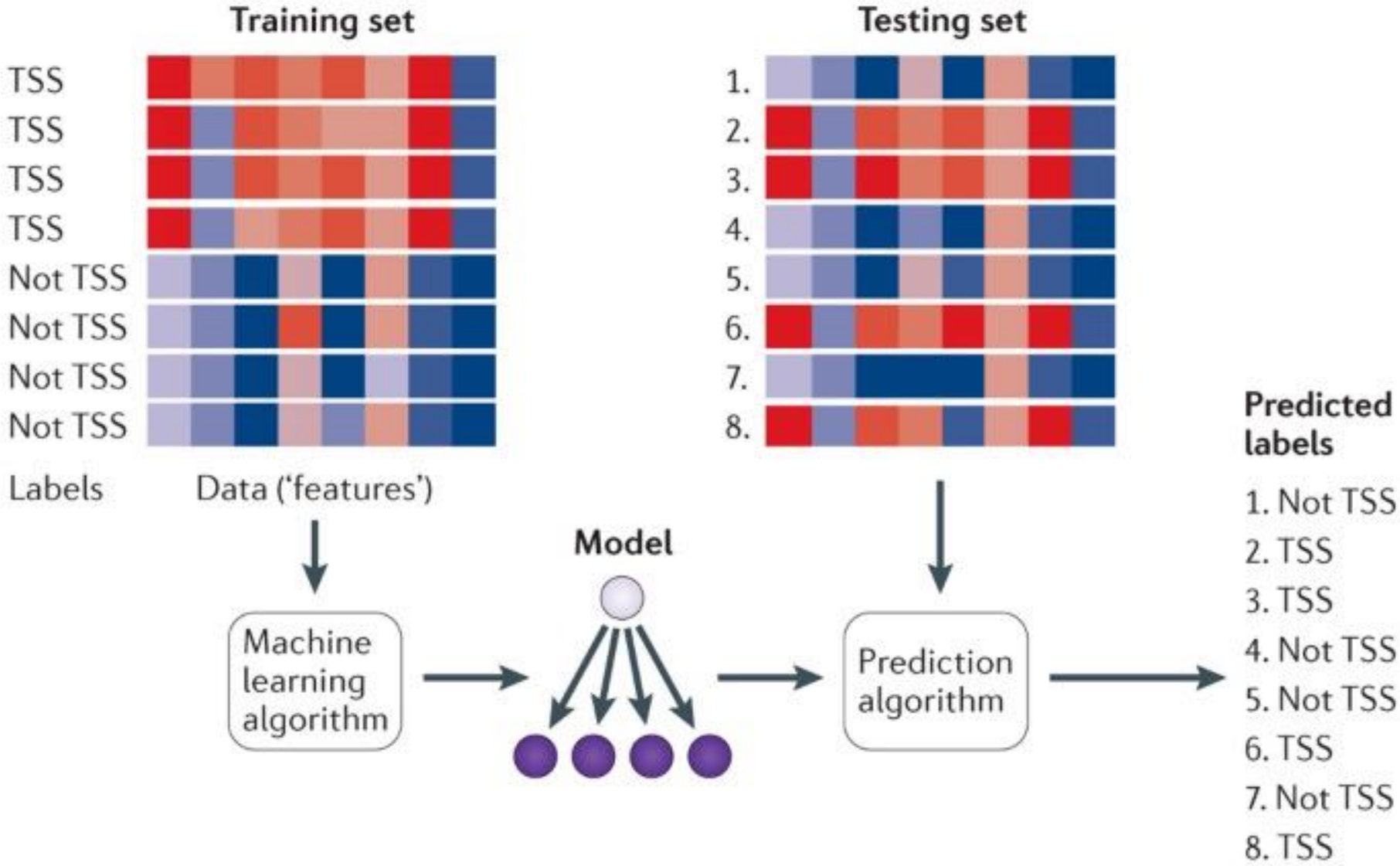
Machine Learning Practice

Tugce Bilgin Sonay, ZHAW, May 2023

Machine Learning and Bioinformatics

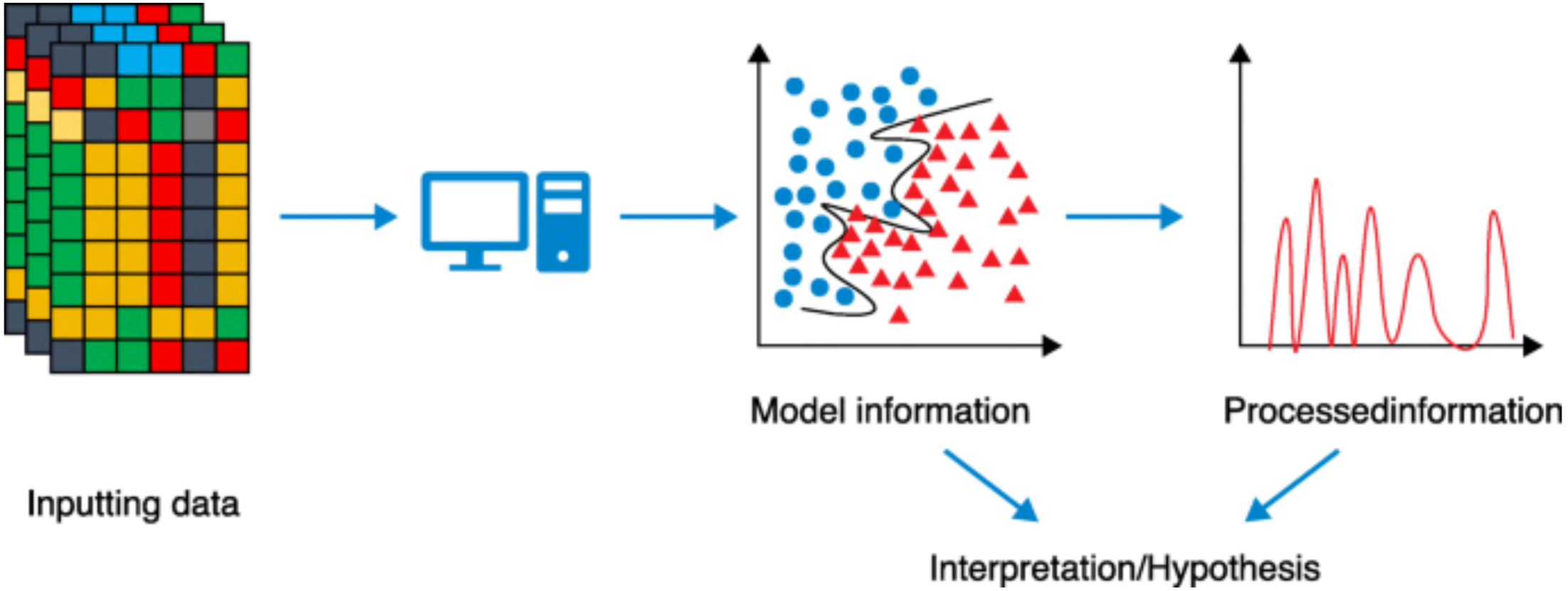


Machine Learning and Bioinformatics

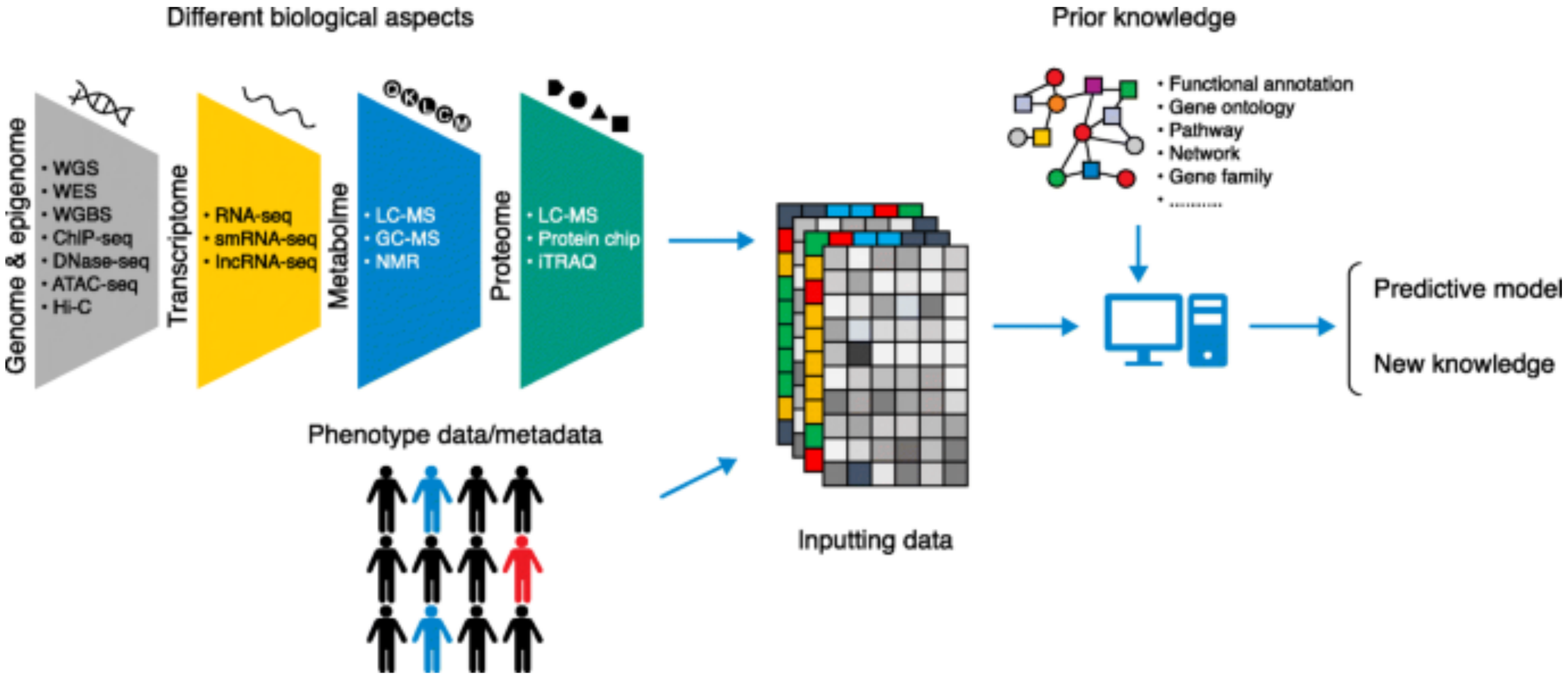


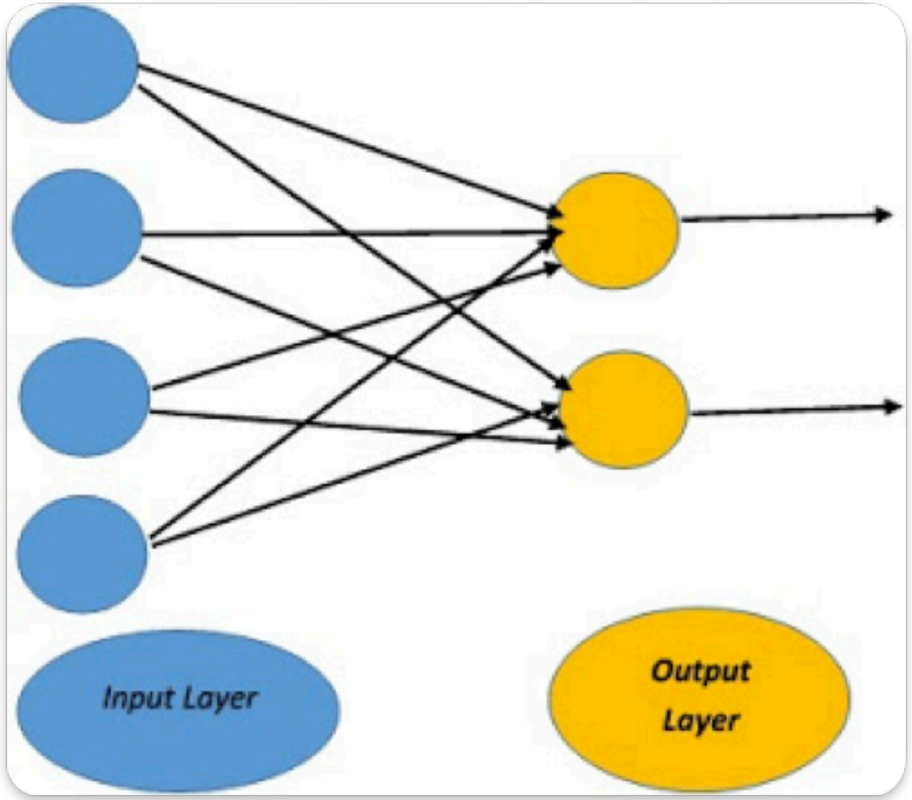
Nature Reviews | Genetics

Machine Learning and Bioinformatics

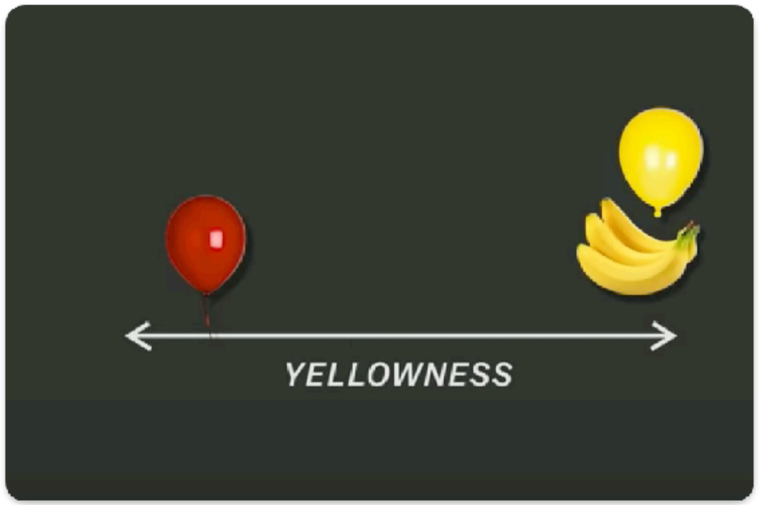


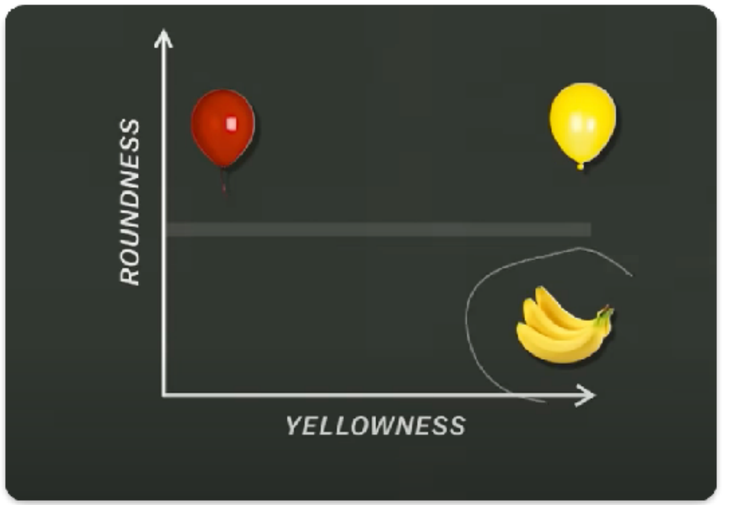
Machine Learning and Bioinformatics

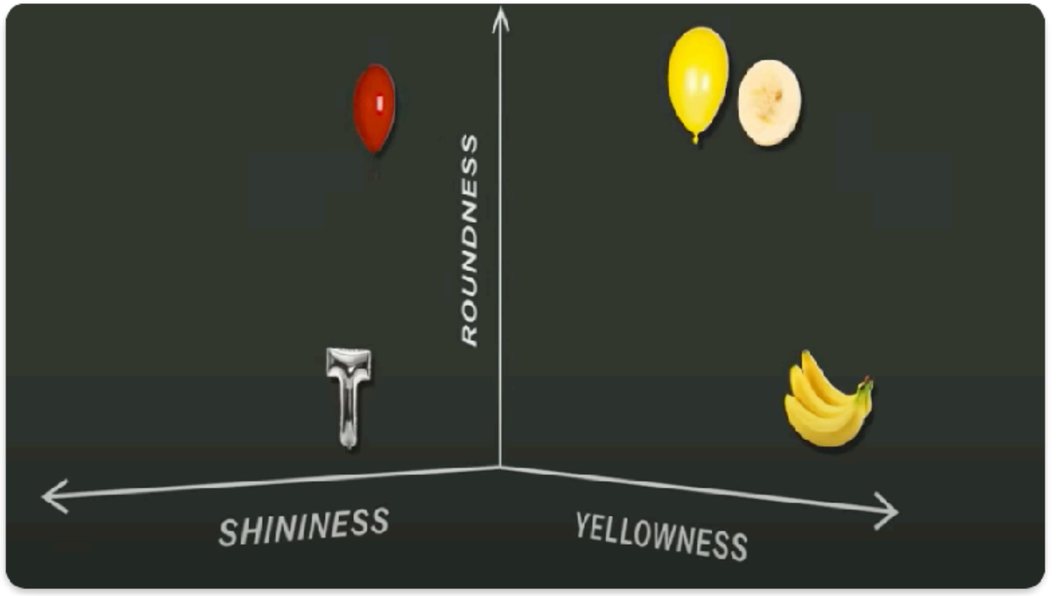


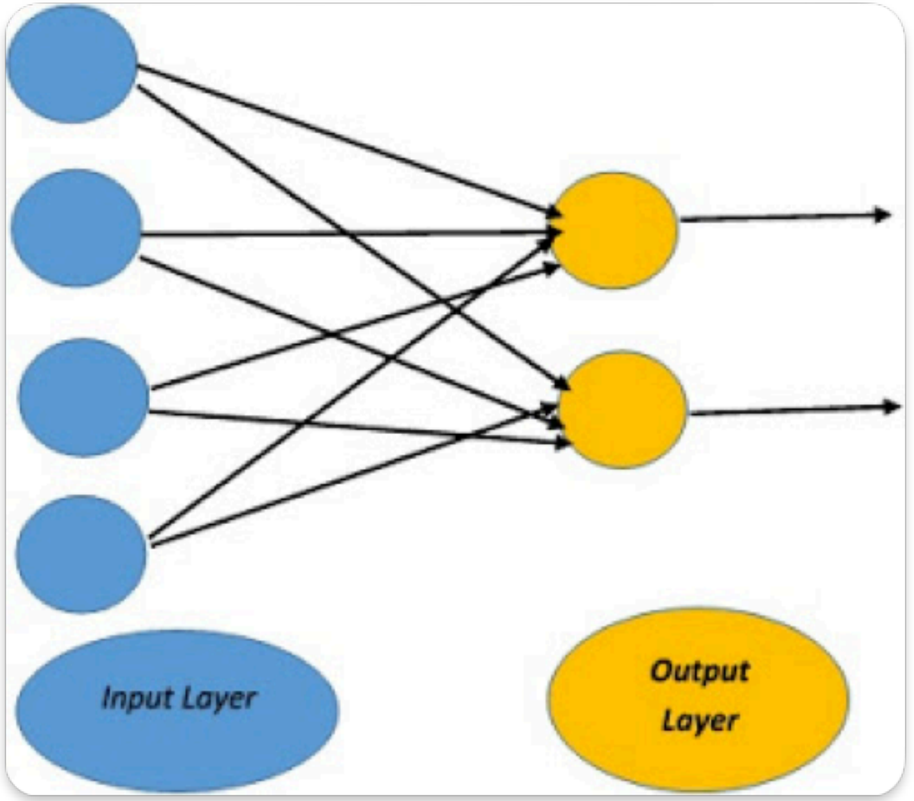

Input Layer
Output
Layer

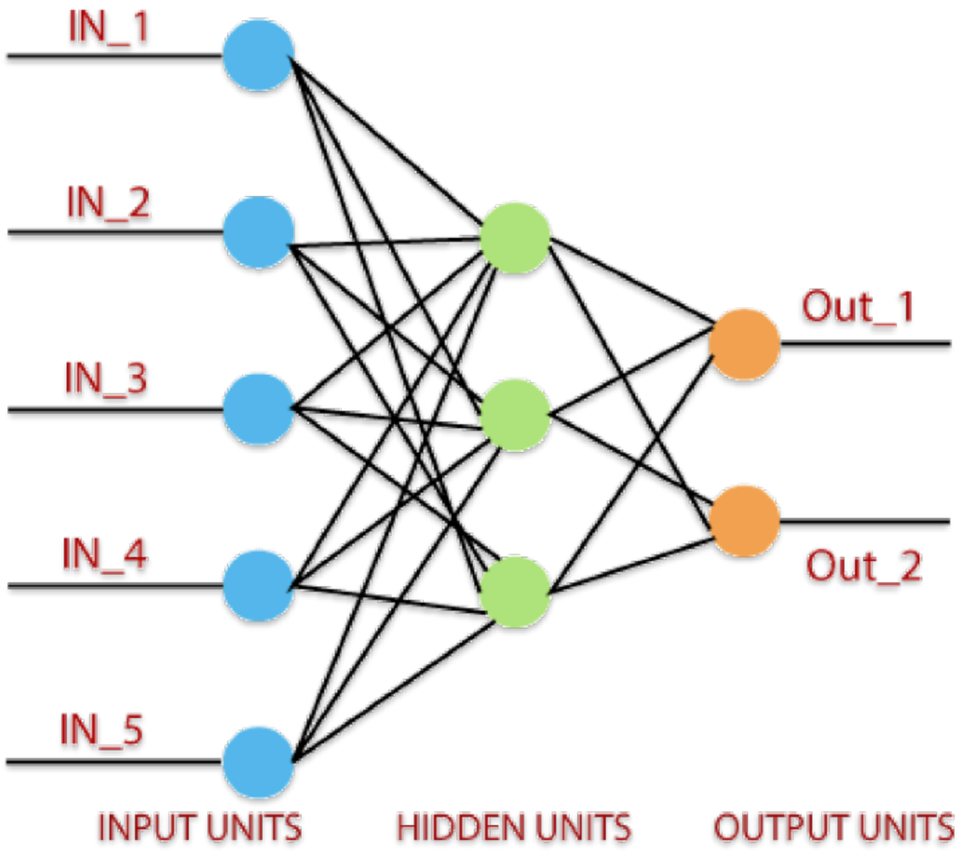

Red party balloon
Banana bunch

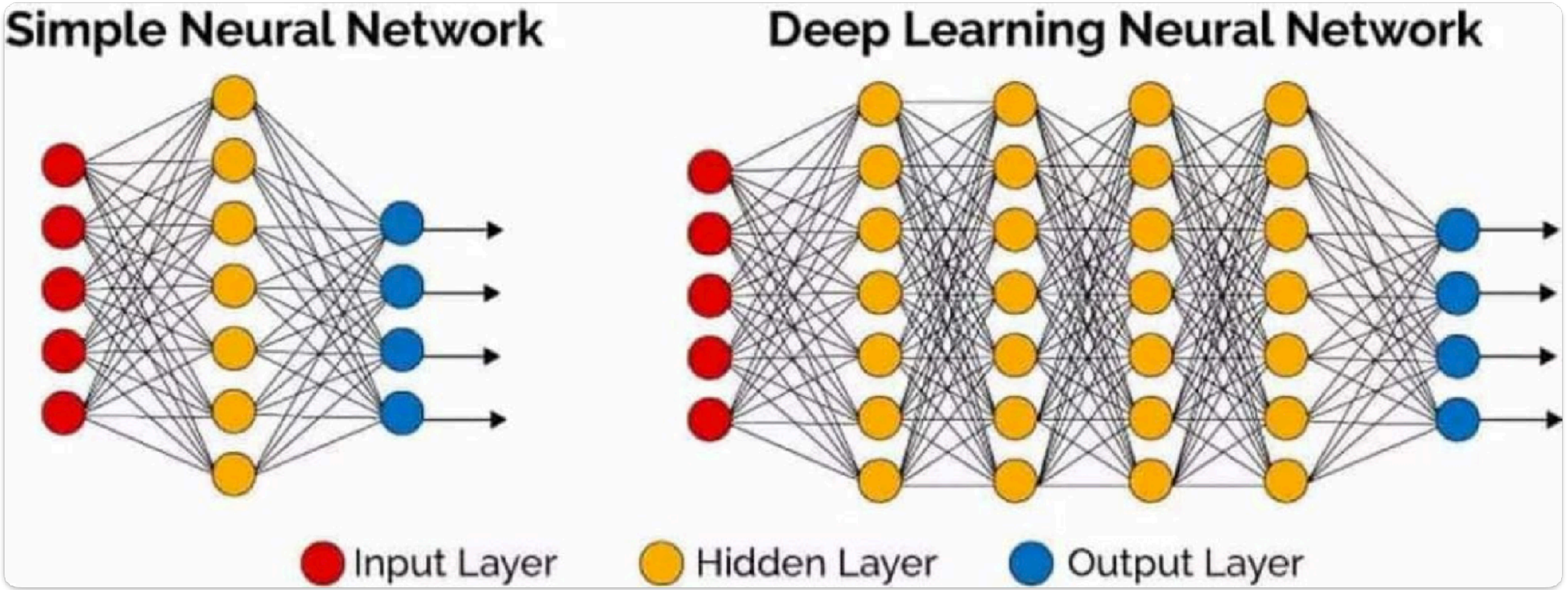

YELLOWNESS


ROUNDNESS
YELLOWNESS

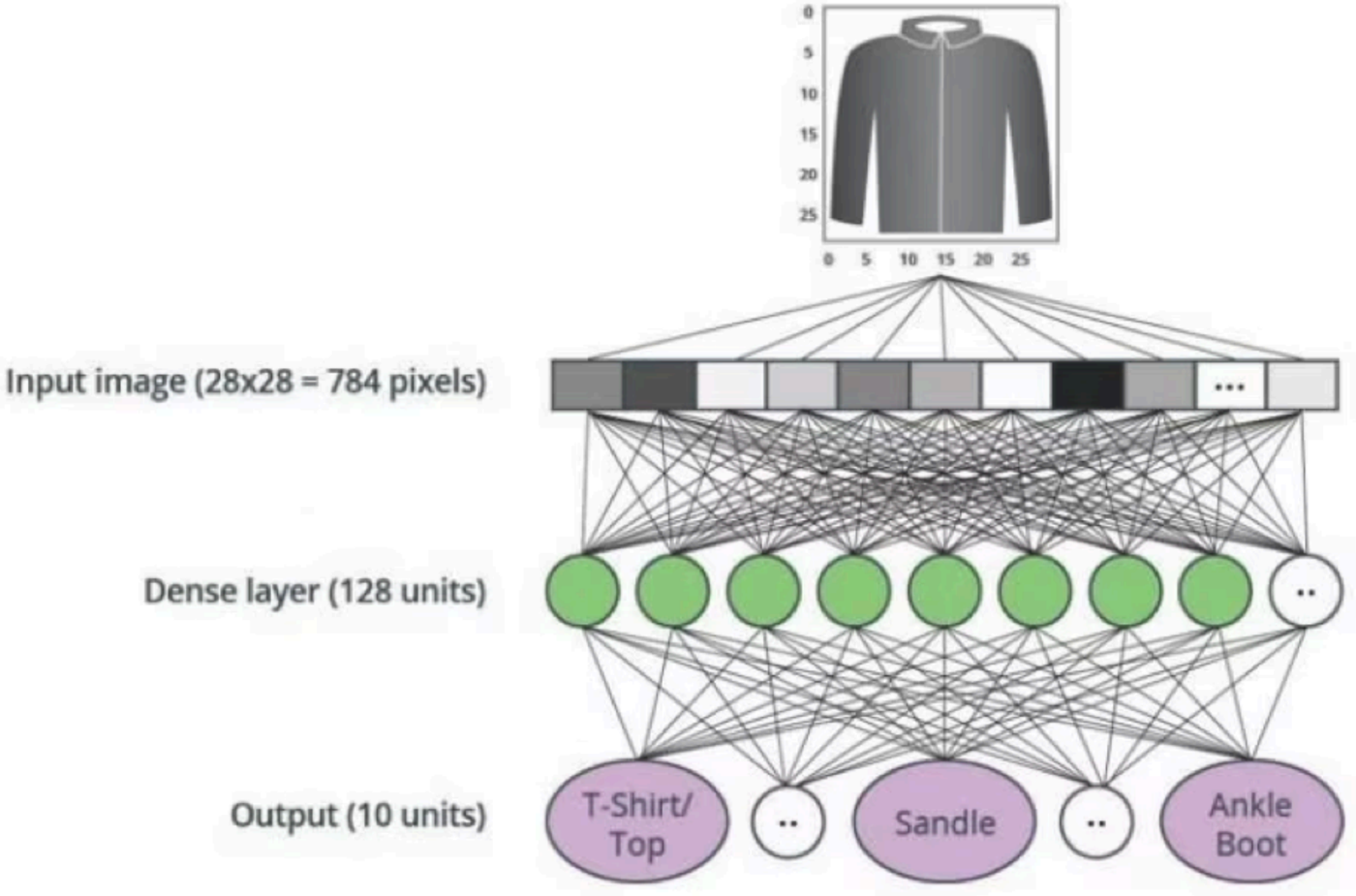

ROUNDNESS
SHININESS
YELLOWNESS


Input Layer
Output Layer


IN_1
IN_2
IN_3
IN_4
IN_5
Out_1
Out_2
INPUT UNITS
HIDDEN UNITS
OUTPUT UNITS


Simple Neural Network
Deep Learning Neural Network
Input Layer
Hidden Layer
Output Layer

Fashion MNIST



Classification
Model output is a set of numbers that represent probabilities (sum is 1)
Loss function: sparse_categorical_crossentropy