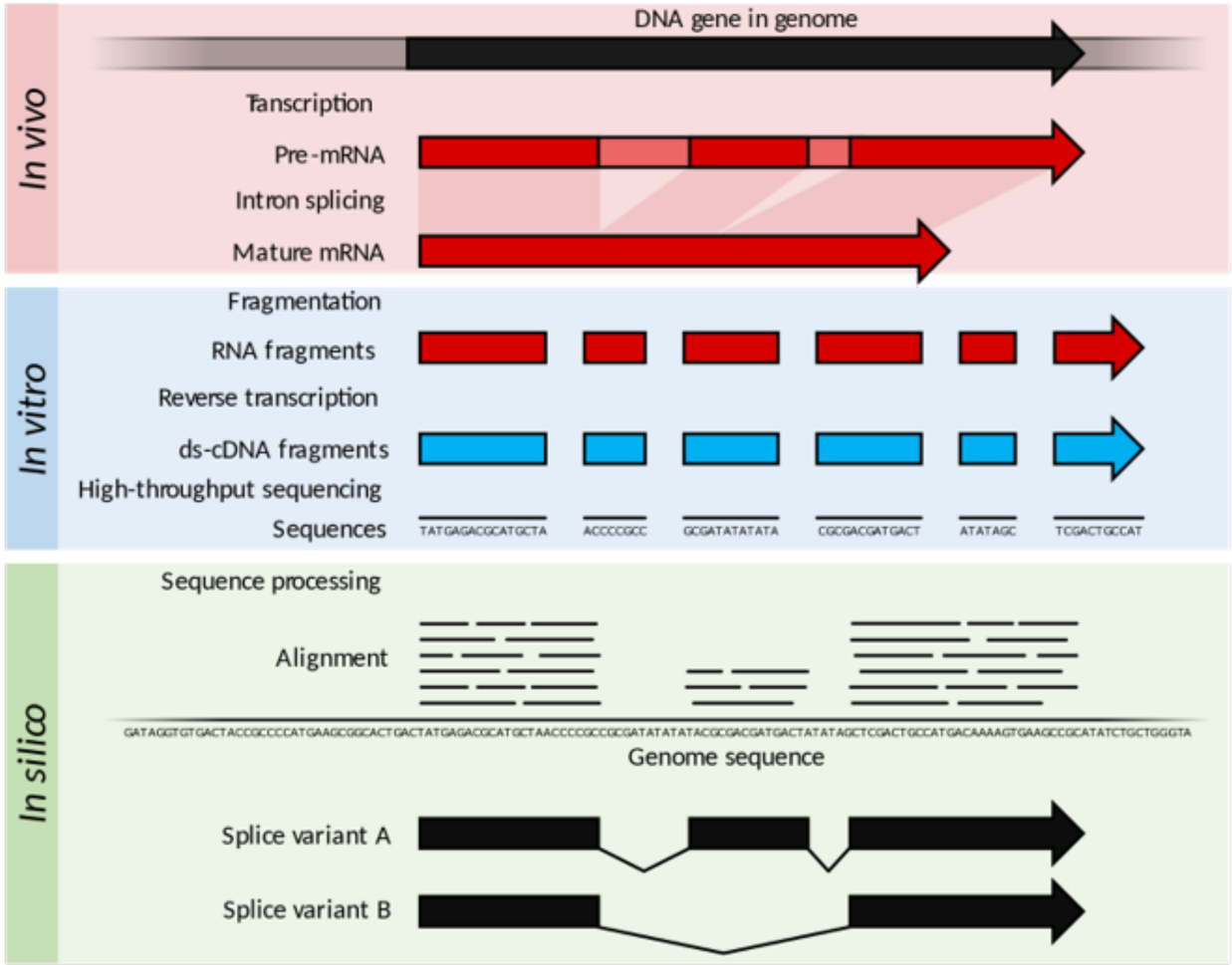
# MACHINE LEARNING APLICATIONS ON RNA-SEQ DATA

BORIS JURIČ

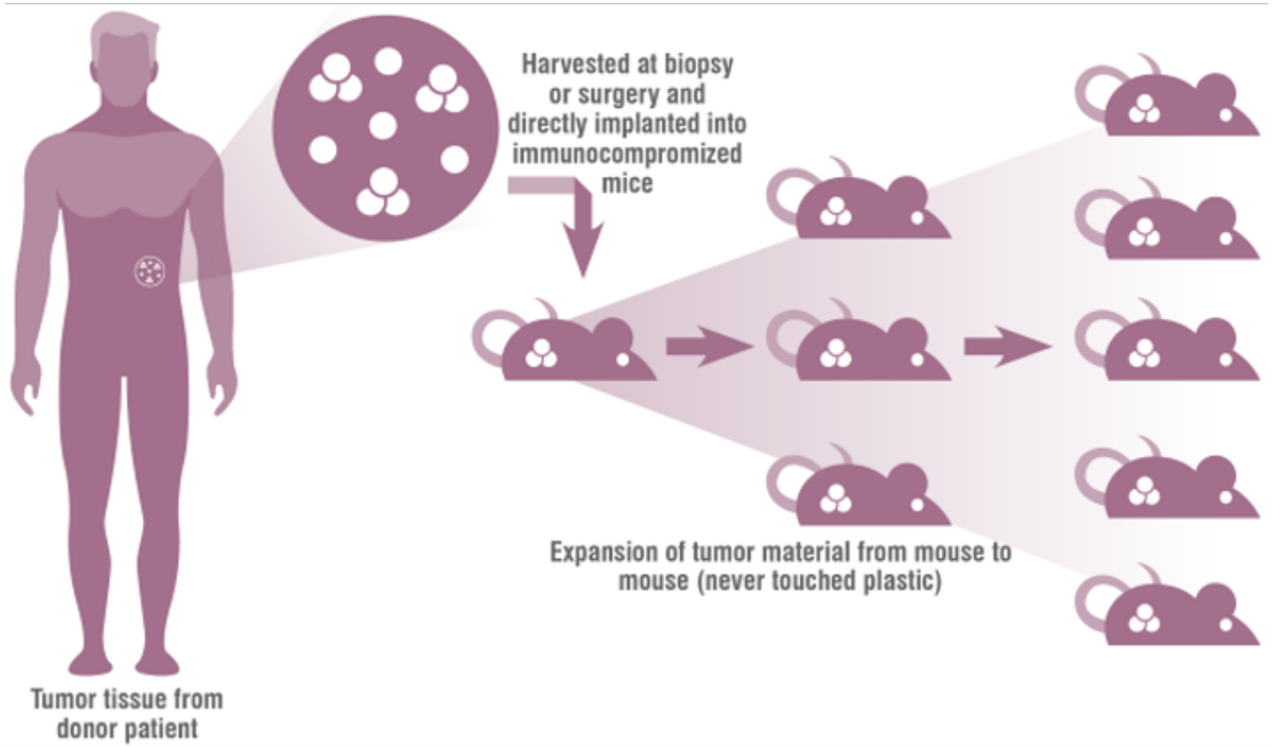
# RNA-sequencing

- Presence and quantity
- Gene expression, transcriptome assembly, differential analysis
- Disease biomarkers, diagnostics



# PDX

- Cancer research
- Data contaminated
- High sequence similarity ~85%



# Standard methods

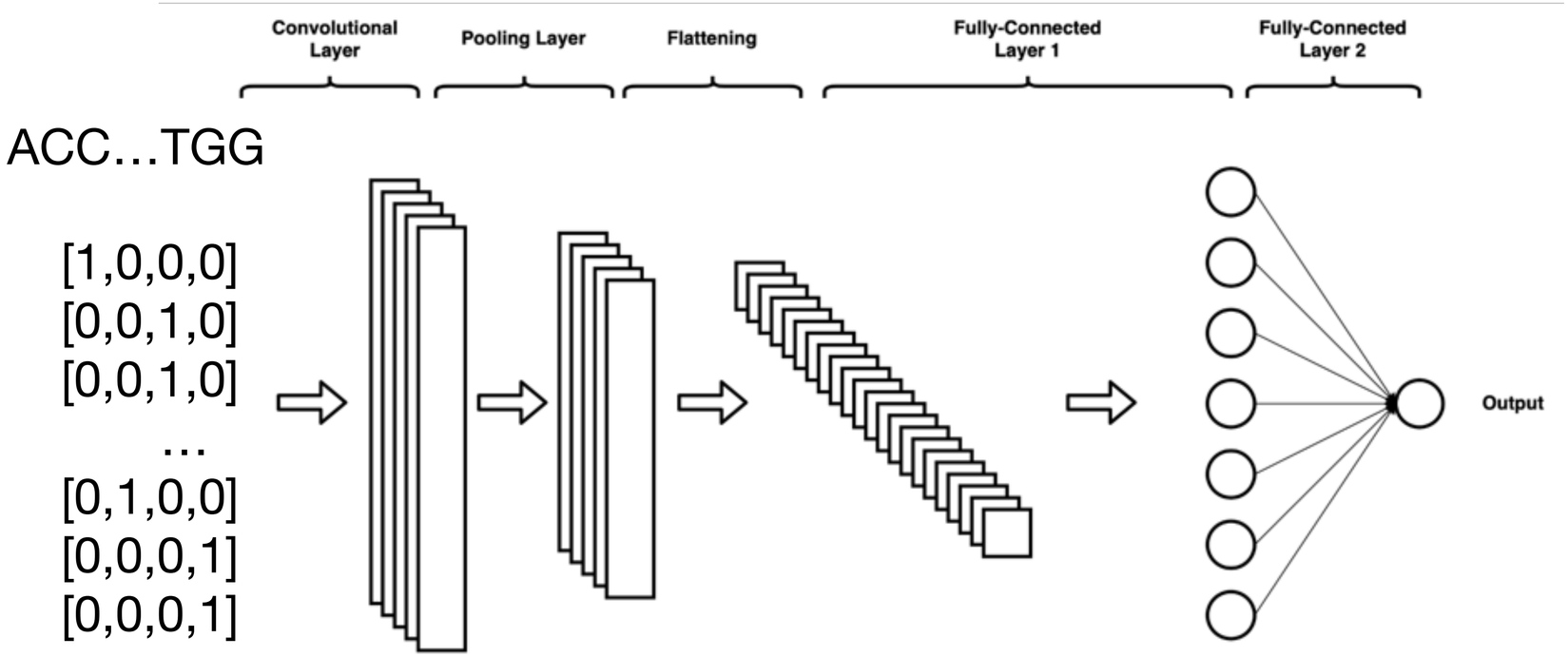
- Xenome - kmer index table
- NGS Disambiguate - two alignments
- Pure alignments methods

# Our approach

- Convolutional neural network

- encoding

A = [1,0,0,0]
T = [0,1,0,0]
C = [0,0,1,0]
G = [0,0,0,1]



# Model assisted alignment

- Goal is gene expression quantification
- Model predicts gene, alignment only local
- Output is n-dimensional space

# Search space and metric

- Kmer content defines distance
- Similar genes closer together

| | AAAAAAAA | AAAAAAAT | ... | GGGGGGGG |
|---|---|---|---|---|
| DDX11L1 | 1 | 0 | ... | 1 |
| WASH7P | 0 | 1 | ... | 0 |
| MIR6859-1 | 0 | 0 | ... | 1 |
| MIR1302-2HG | 1 | 0 | ... | 0 |
| … | … | … | … | … |

- Random projection to reduce number of dimensions
- Relative distance preserved

# High dimensional regression

- Model predicts coordinates
- Tree search the matrix
- Alignment on N closest neighbours

# So far

- Works on small amount of genes
- Uneven gene size, insufficient metric
- Reference data structure needs improvement

# Current work

- Clustering kmers from the whole genome
- Generating search space in similar fashion
- Model predicts cluster, best alignment is chosen

# Thank you for attention