Master Praktikum Bioinformatics SS19

Alexander Karollus & Julien Gagneur
Assistant Prof. for Computational Biology
Technical University of Munich
www.gagneurlab.in.tum.de
github.com/gagneurlab

To understand the genetic basis of gene regulation and its implication in diseases

Human genome: A text of 3 billion bases…

attcgagtattcgaattgatcgatgattagcatcatgcag

A single base difference can have drastic consequences

Reference attcgagtattcgaattgatcgatgattagcatcatgcag

Patient at**a**cgagtattcgaattga**g**cgatgattagca**c**catgcag

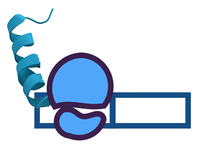
but which ones?

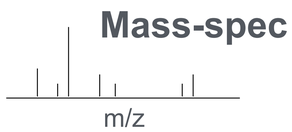
Reference attcgagtattcgaattgatcgatgattagcatcatgcag

Patient at**a**cgagtattcgaattga**g**cgatgattagca**c**catgcag

We study how gene regulation, the control of gene activity, is encoded in the genome

Translation





Splicing

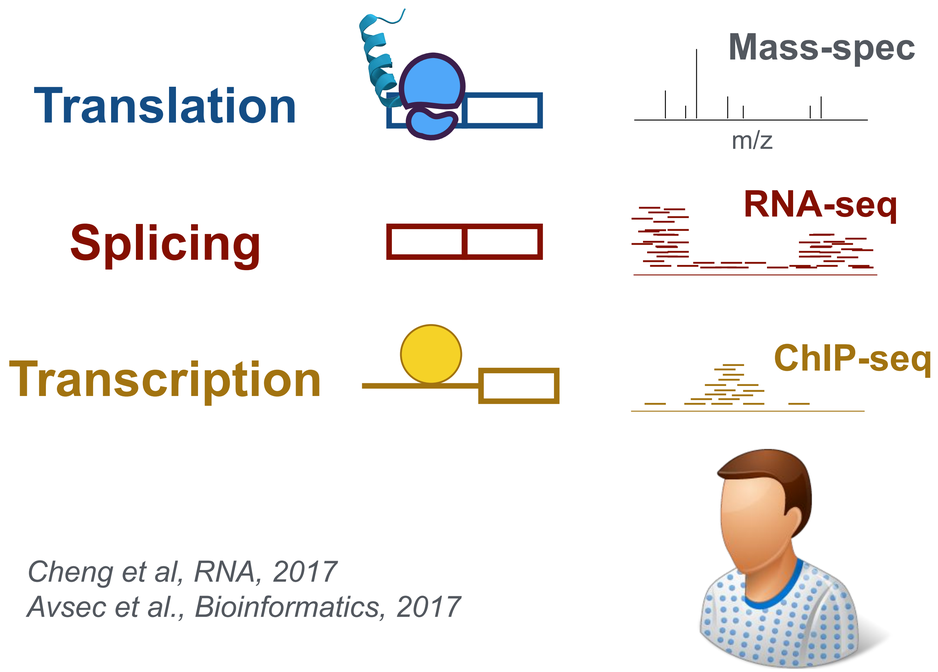
Transcription

Eser et al. Mol. Syst. Biol, 2016
Schwalb et al, Science, 2016

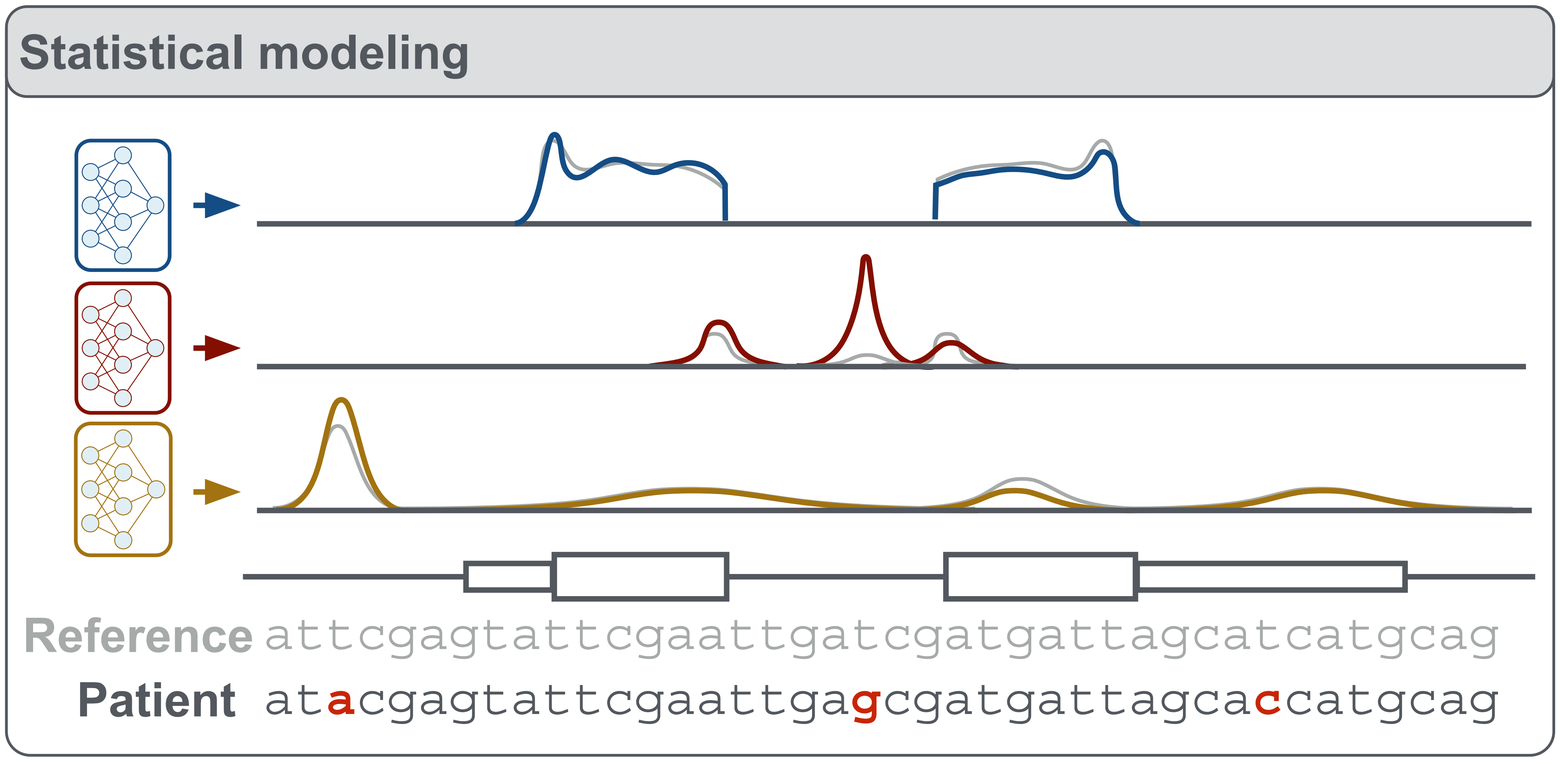
Reference attcgagtattcgaattgatcgatgattagcatcatgcag

Patient at**a**cgagtattcgaattga**g**cgatgattagca**c**catgcag

build computational models that predict the implications of genetic modifications on gene regulation



Cheng et al, RNA, 2017
Avsec et al., Bioinformatics, 2017



and integrate genetic and molecular data of patients to support diagnostics of genetic disorders.

Patient board

Gene: *YFG2*
chr 2:138,928,826 t > g
✓ Translation
- Splicing 98%
✓ Transcription

Translation

Splicing

Transcription

Patient Omics

Mass-spec
m/z

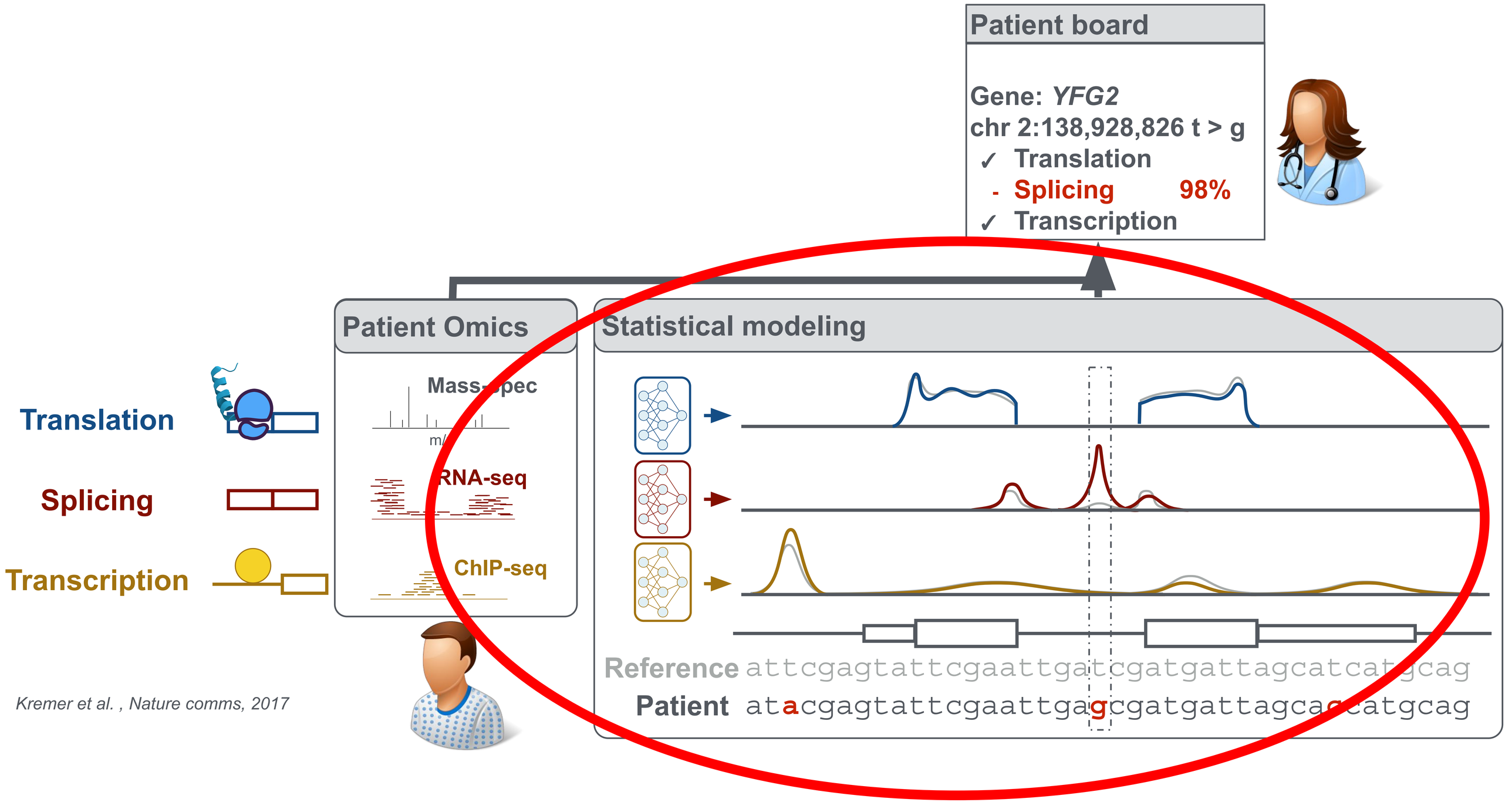
RNA-seq

ChIP-seq

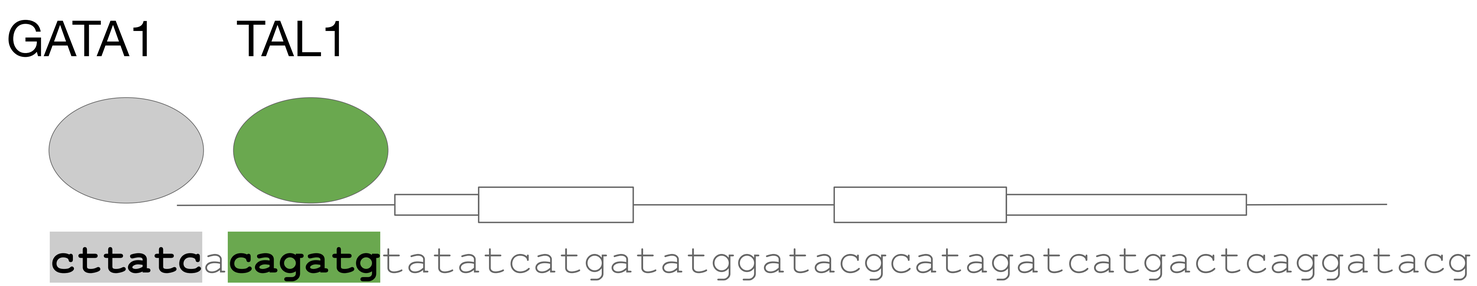
Statistical modeling

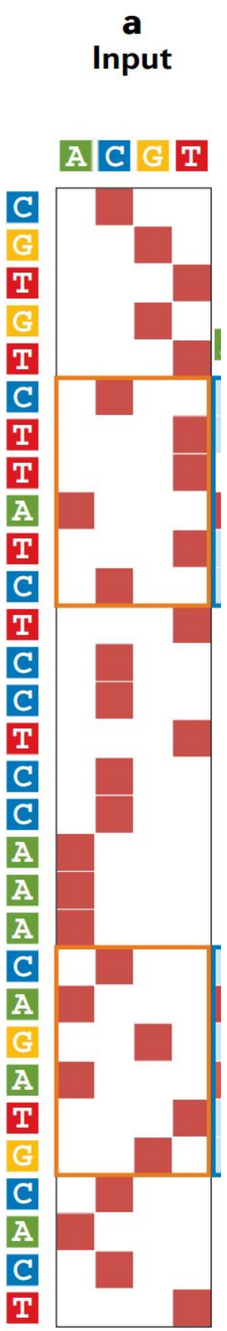
Reference attcgagtattcgaattgatcgatgattagcatcatgcag
Patient atacgagtattcgaattgagcgatgattagcaccatgcag

Kremer et al. , Nature comms, 2017

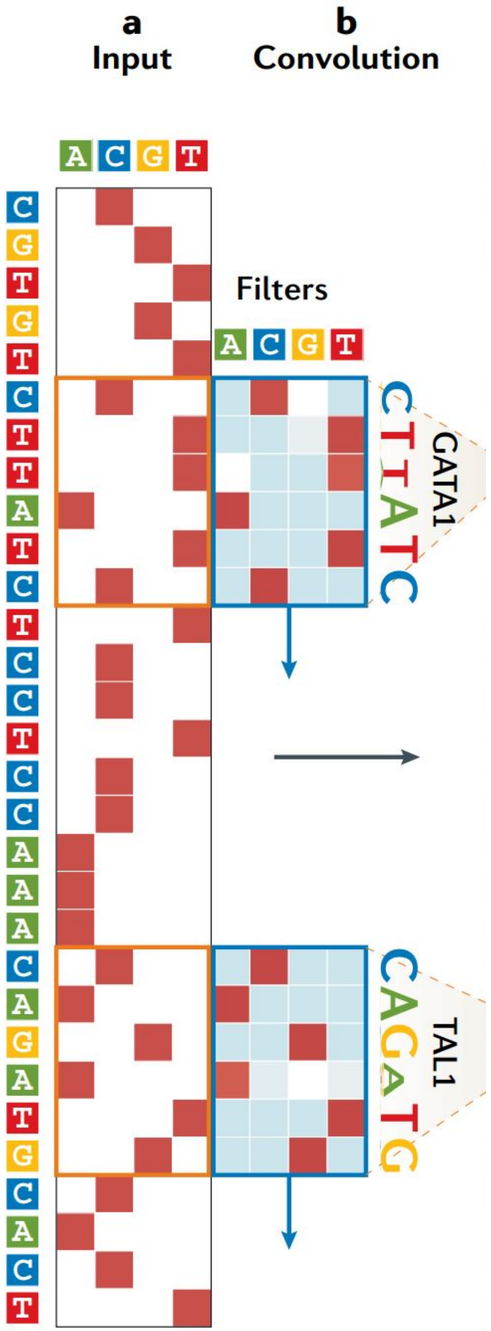

Patient board
Gene: YFG2
chr 2:138,928,826 t > g
✓ Translation
- Splicing 98%
✓ Transcription
Patient Omics
Mass-spec
m/
RNA-seq
ChIP-seq
Statistical modeling
Translation
Splicing
Transcription
Reference attcgagtattcgaattgatcgatgattagcatcatgcag
Patient atacgagtattcgaattgagcgatgattagcagcatgcag
Kremer et al. , Nature comms, 2017

Detecting regulatory elements with deep neural networks

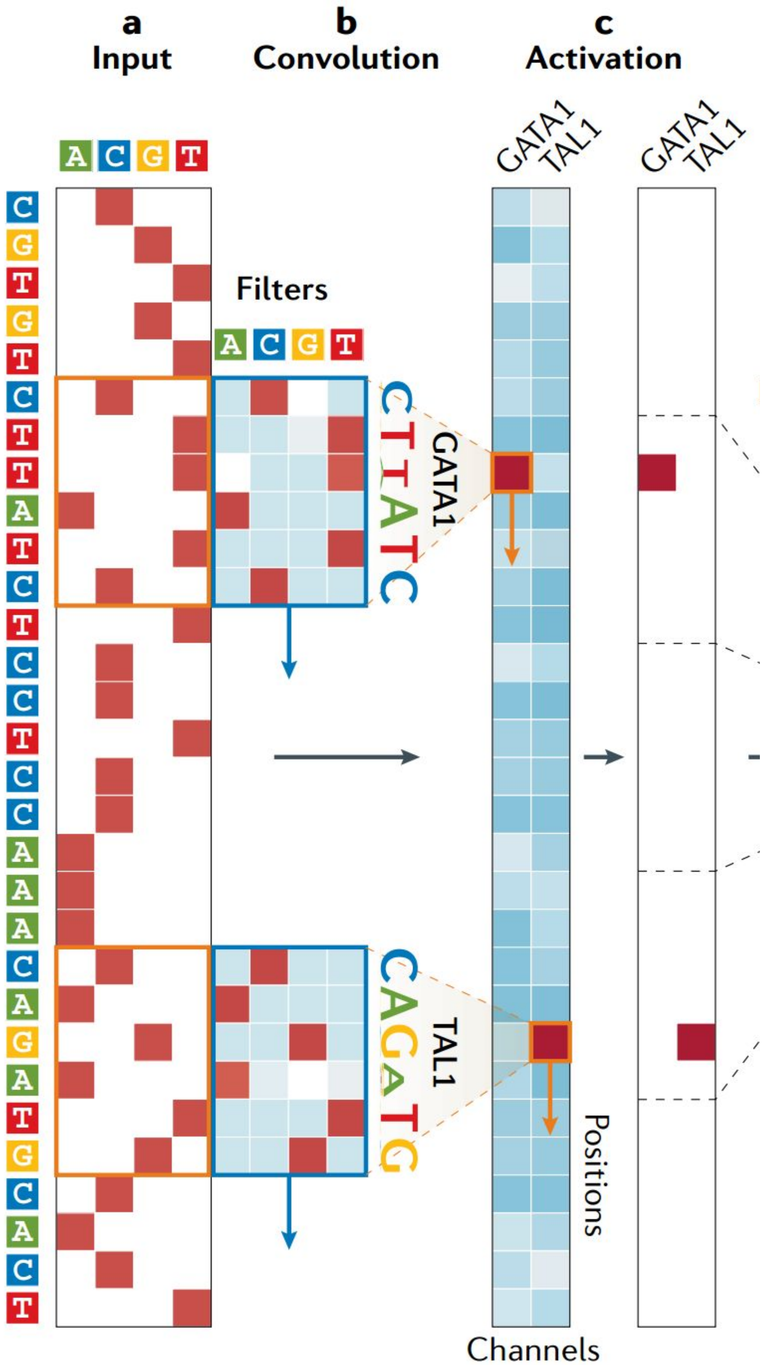




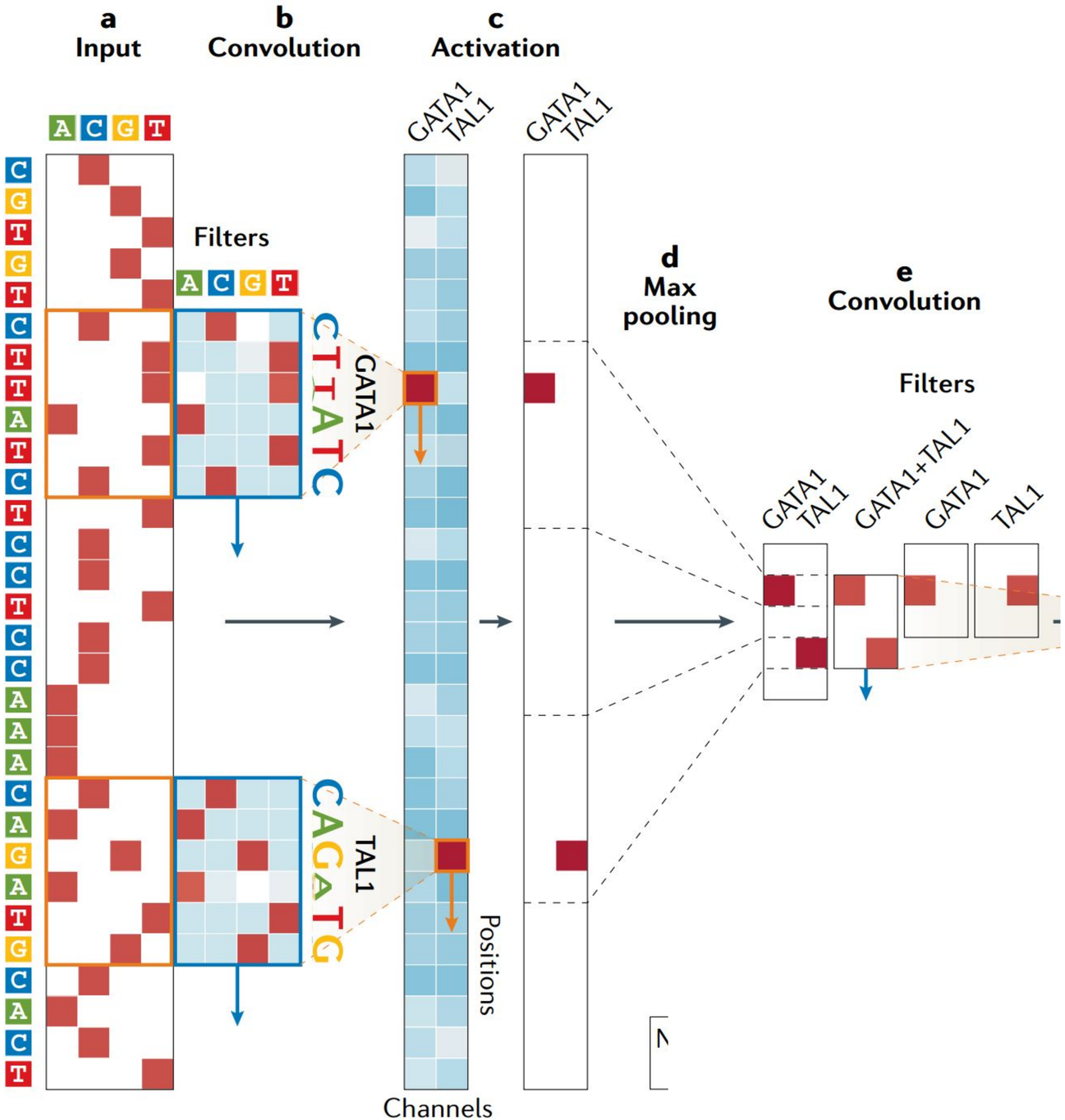
Eraslan*, Avsec* et al Nature Review Genetics 2019



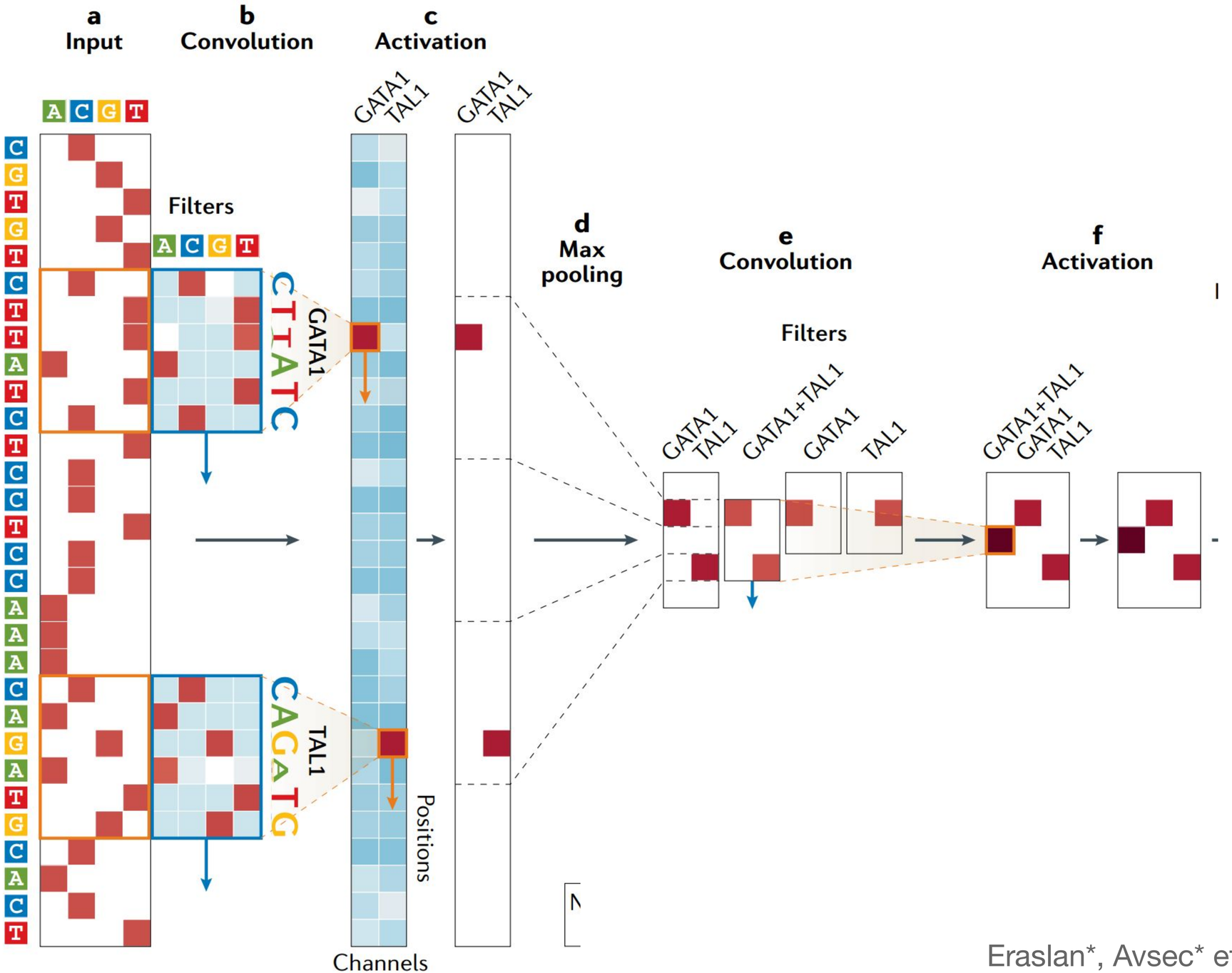
C

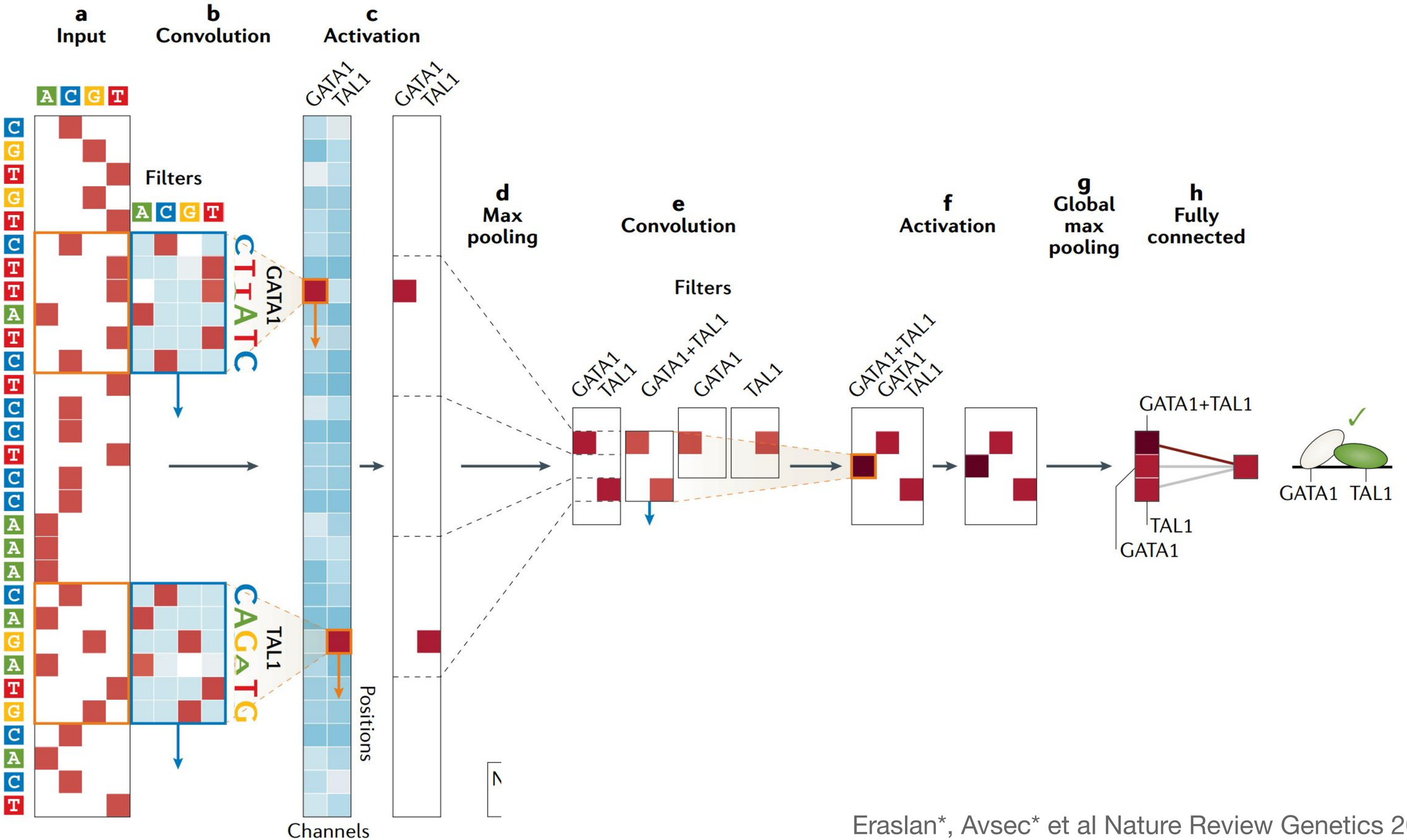


Eraslan*, Avsec* et al Nature Review Genetics 2019

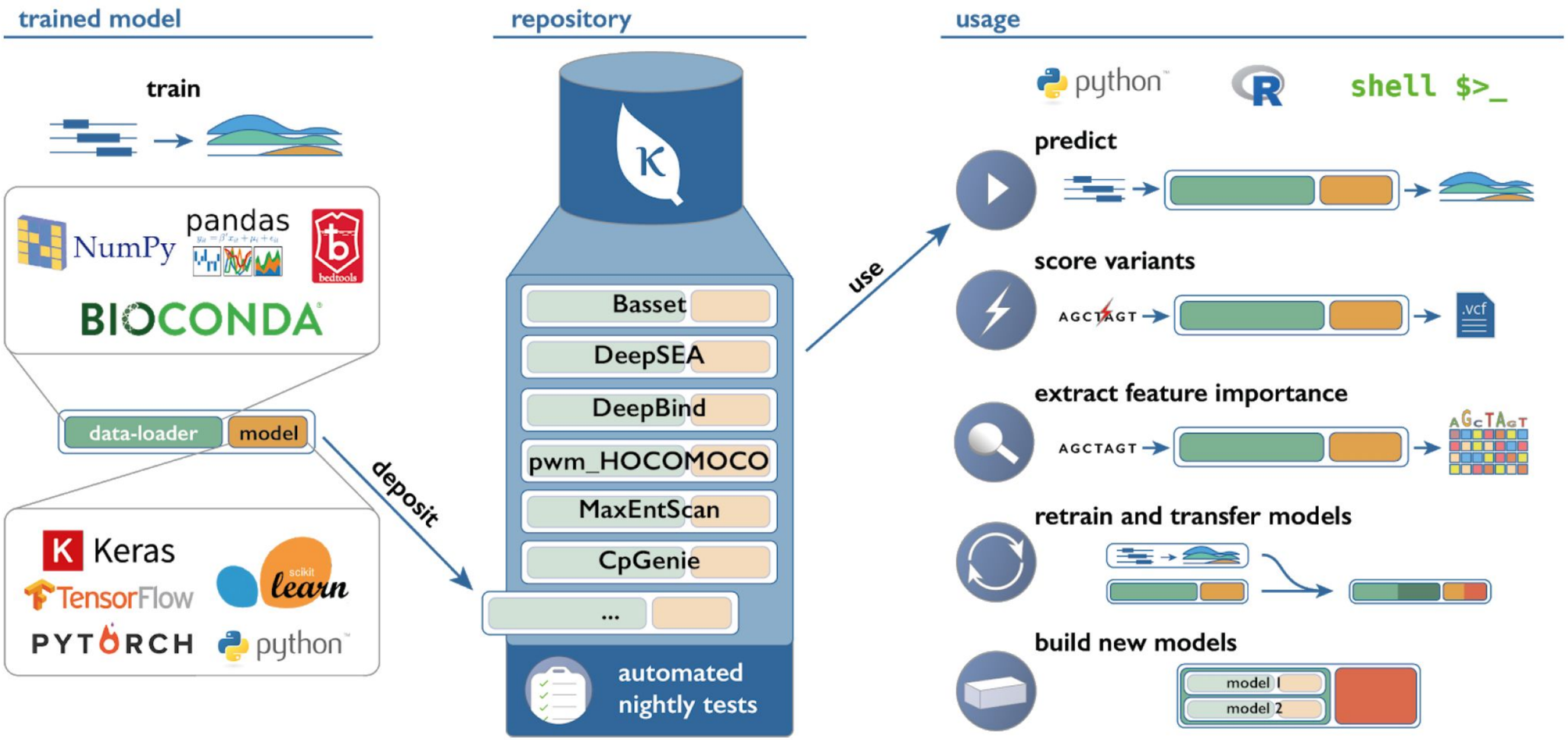


Eraslan*, Avsec* et al Nature Review Genetics 2019


a
Input
b
Convolution
c
Activation
d
Max
pooling
e
Convolution
f
Activation
Filters
GATA1
TAL1
GATA1
TAL1
GATA1
TAL1
GATA1+TAL1
GATA1
TAL1
GATA1+TAL1
GATA1
TAL1
Positions
Channels


a
Input
b
Convolution
c
Activation
d
Max
pooling
e
Convolution
f
Activation
g
Global
max
pooling
h
Fully
connected
Filters
GATA1
TAL1
GATA1+TAL1
Positions
Channels
GATA1
TAL1

Kipoi.org [Kípi]: A repository of trained predictive models for genomics

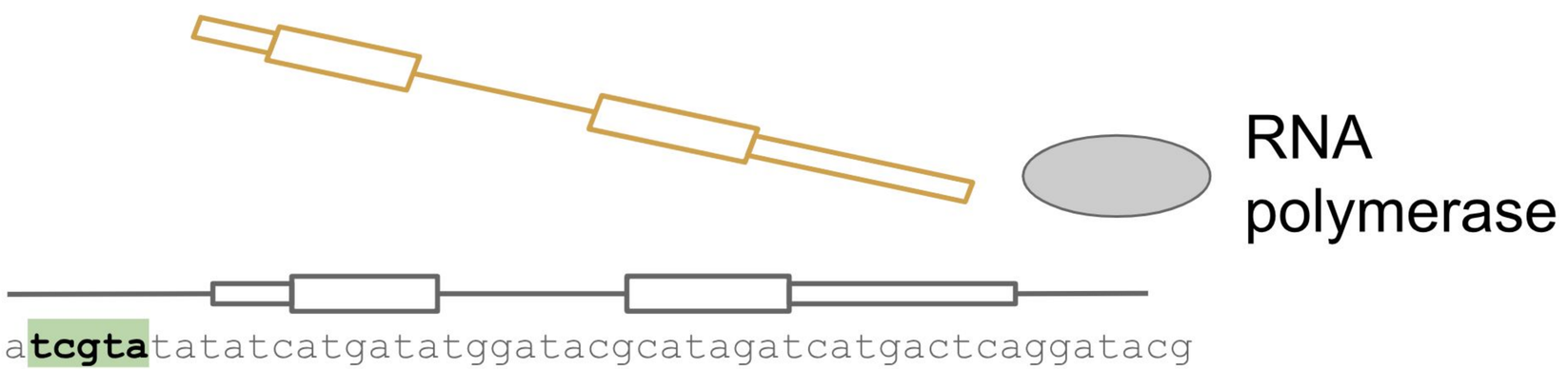


with A. Kundaje, Stanford, and O. Stegle, DKFZ

Avsec et al, Nature biotechnol. (accepted)

The projects

1. **Regulatory code of Transcription Start Site (TSS)**
2. **Regulatory code of polyA sites**

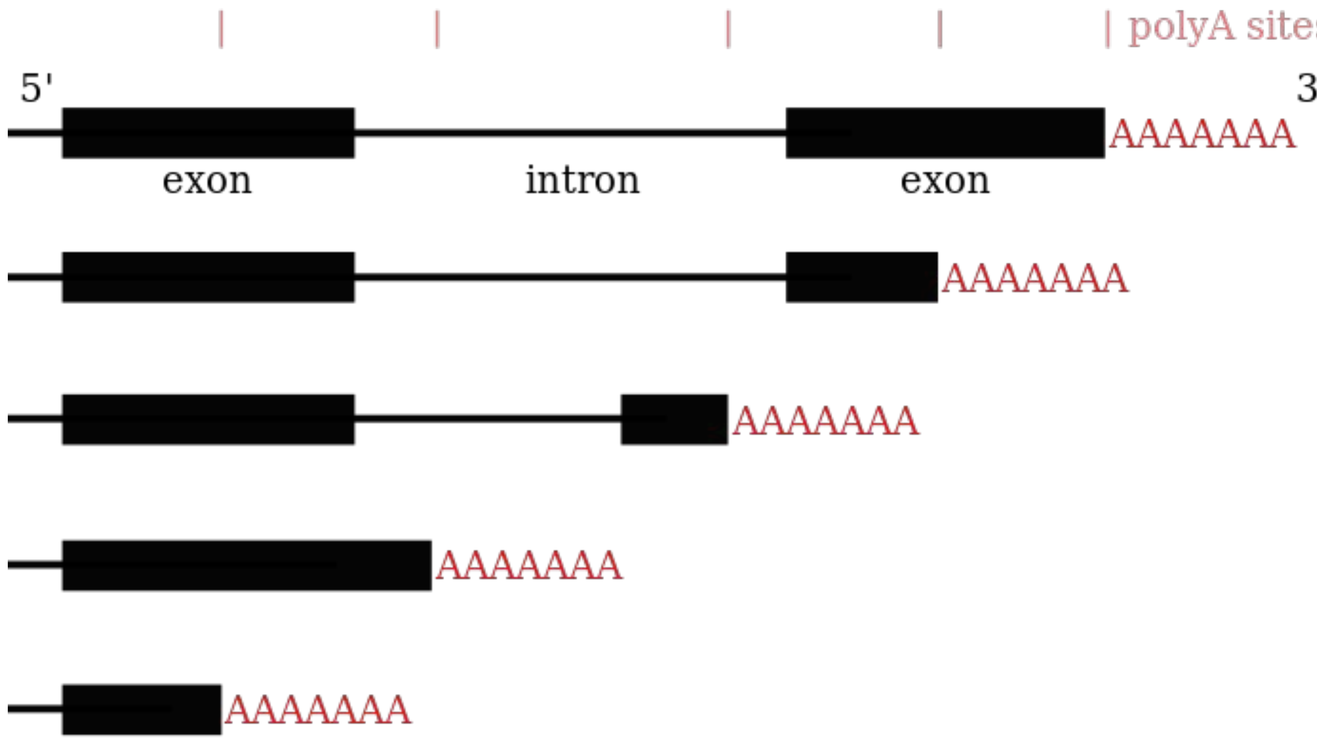


Motivation

Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues

They can lead to changes in the coding sequence and in unstranslated regions

They are an important regulatory mechanism with considerable roles in many biological processes and diseases, such as cell differentiation, proliferation and cancer.



Goal

- Build deep neural networks predicting TSS (or polyA site choice) from DNA sequence trained on:
  i) Endogeneous genes and/or Massively parallel reporter assays
  ii) (if time allows) endogeneous genes in multiple conditions (eg: tumour vs healthy control tissues)

- Benchmark different network architectures

- Share the model with the community in the model repository Kipoi. It will be applicable to predict effects of genetic variants from human genotype data.

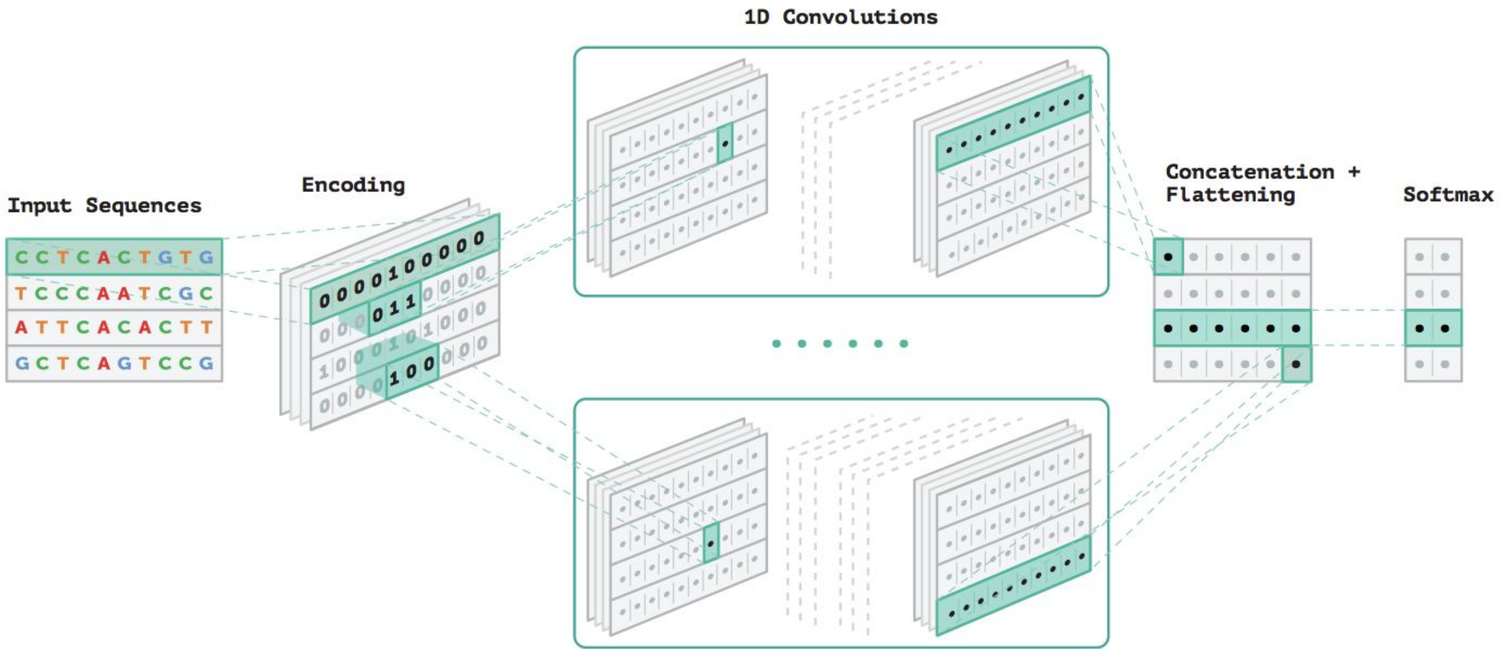
- (if time allows) Apply it to patient genotype data (rare disease).

TSS baseline model and data

- PROMID Umarov et al, arxiv.

- Curated human promoters
  - 29,598 promoters

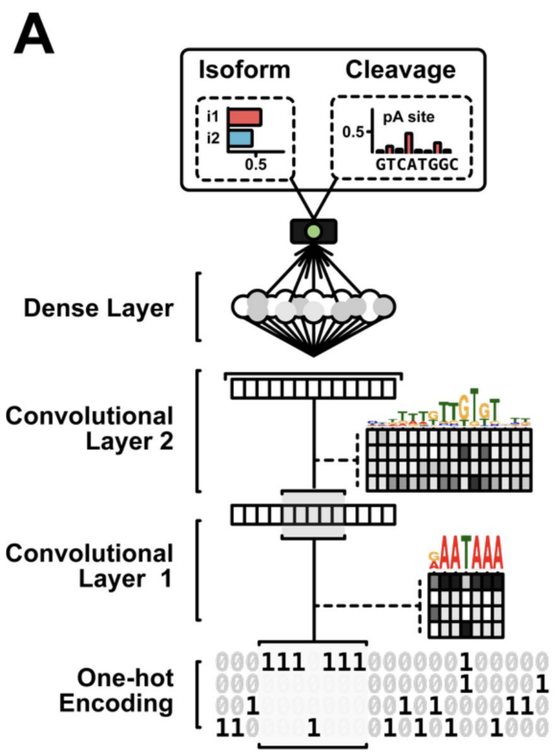
- FANTOM5 mapped TSS
  - 184,827 robust TSSs

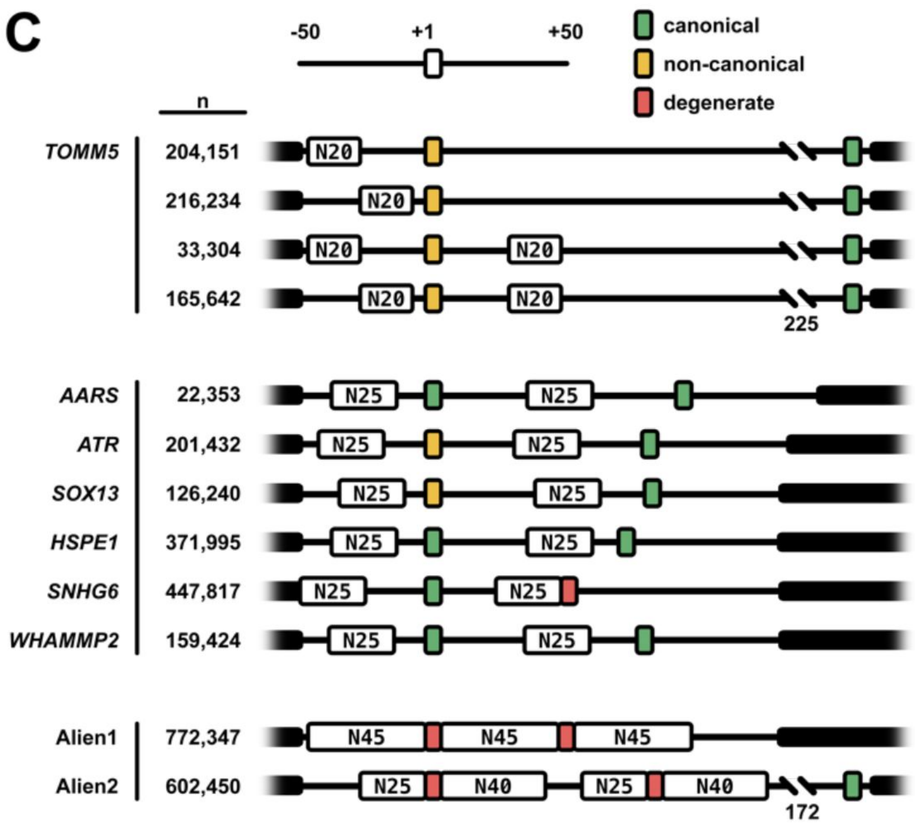


SIB EPD EUKARYOTIC PROMOTER DATABASE

pA site baseline model and data

- APPARENT Bogard et al, biorxiv.
- Massively parallel reporter assay data (Bogard et al)
  - 2.4 Millions artificial sequences across 12 UTR contexts
- Endogeneous genes
  - GENCODE v. 30 human annotation (ca 83,000 protein-coding transcripts)





References

- Eraslan, Avsec, et al. Deep learning: New computational modeling techniques for genomics, Nature Reviews Genetics, (2019) http://rdcu.be/bvNef

- Avsec et al. Kipoi: accelerating the community exchange and reuse of predictive models for genomics (2018), bioRxiv

- Reyes & Huber Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues (2018), NAR

- Umarov et al, PromID: human promoter prediction by deep learning (2018) arXiv

- Bogard et al, Predicting the Impact of cis-Regulatory Variation on Alternative Polyadenylation (2018) biorxiv