Machine learning to probe novel regulatory elements in the human genome

Charles Lecellier

charles.lecellier@igmm.cnrs.fr

IGMM/LIRMM/IMAG team

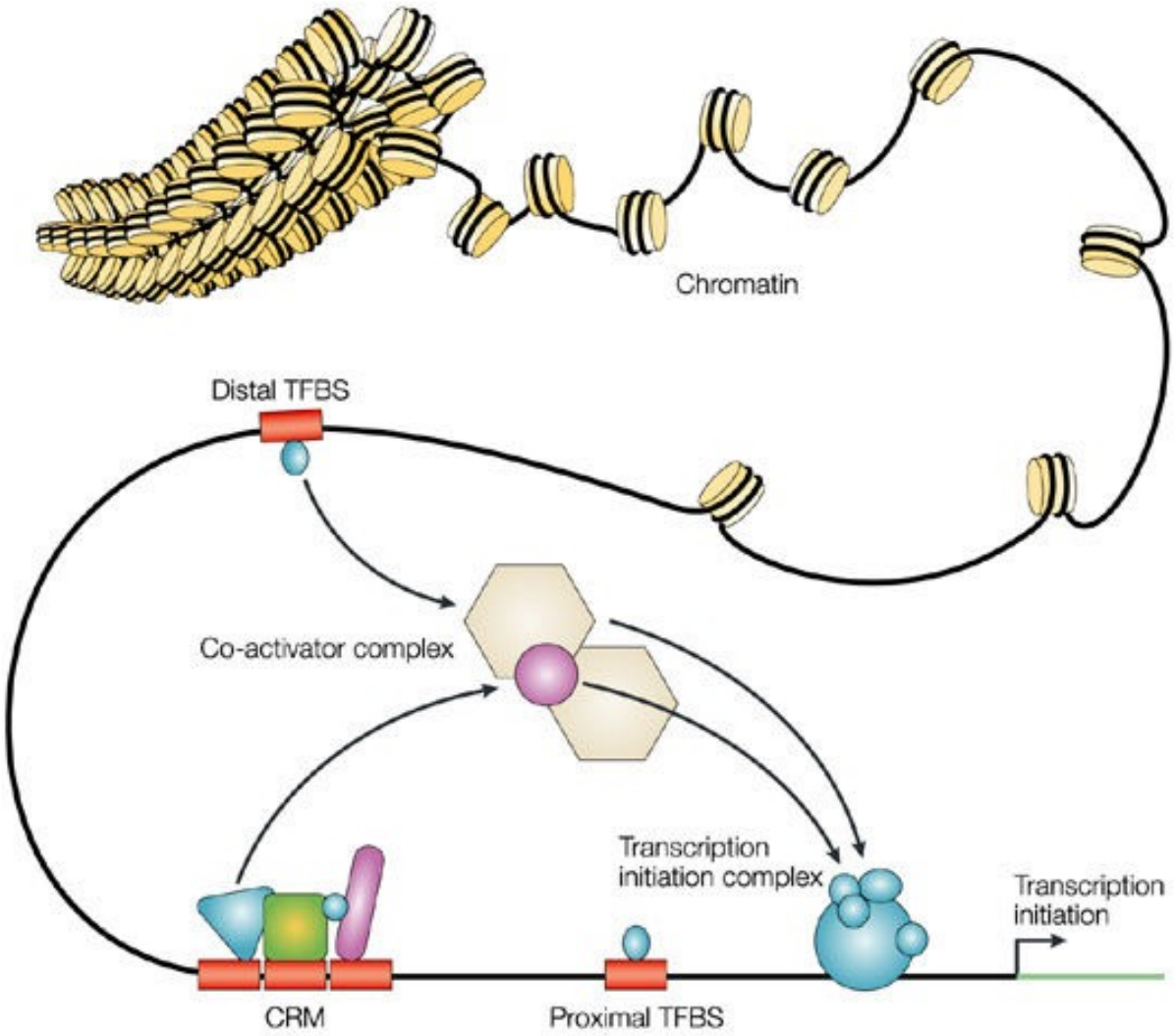
Computational Regulatory Genomics

INFORMATICS

BIOLOGY

MATHEMATICS

Probing genomic regulations



from Wasserman & Sandelin, Nature Reviews Genetics 2004

many high-throughput technologies and large amount of data

RNA-seq

ChIP-seq (TF, epigenetics marks)

Open regions (ATAC-seq)

DNA methylation

DNA sequence (motifs, k-mer, …)

formulation of the problem

predicted variable Y
(labels – supervised problem)

predictive variables X

sequence_id	class	signal	x1	x2	...	xn
seq1	0	0.83	0.87	0.91	...	0.42
seq2	1	2.52	0.72	0.54	...	0.78
...	...	...	...	...	...	...
seqn	1	3.2	0.89	0.75	...	0.95

OR

classification
(0 is not trivial)

TF binding
epigenetic marks
…

regression

RNA level
epigenetic marks
…

k-mer count
motif scores
epigenetic marks
methylation
…

$$Y \sim X$$

loss function : optimized during training to fit machine learning model parameters. In the simplest case, it measures the discrepancy between predictions and observations e.g. mean-squared error (MSE – regression), cross-entropy (classification).

objectives

. **PREDICTION**: e.g. better interpretation of genetic variations in particular in non-coding regions (80% of the variants)

what is the effect/impact of a SNP, which modifies X, on Y?

. **EXPLANATION**: e.g. characterization of molecular determinants (DNA features/instructions) and finding the molecular rules | DNA grammar | *cis*-regulatory code

what are the relevant Xs?

>>>> Need **interpretable** models

One solution among many: linear regression

$$y(s) = a + \sum_i b_i x_{i,s} + e(s)$$ **linear regression**

where y(g) is the expression of seq s, xi,s is variable i for seq s, e(s) is the residual error associated with seq s, a is the intercept and bi is the regression coefficient associated with variable i.

$$P(1|s) = S(a + \sum_i b_i x_{i,s})$$ **logistic regression**

P(1|s) is the probability that sequence s belongs to the first class, S is the sigmoid function

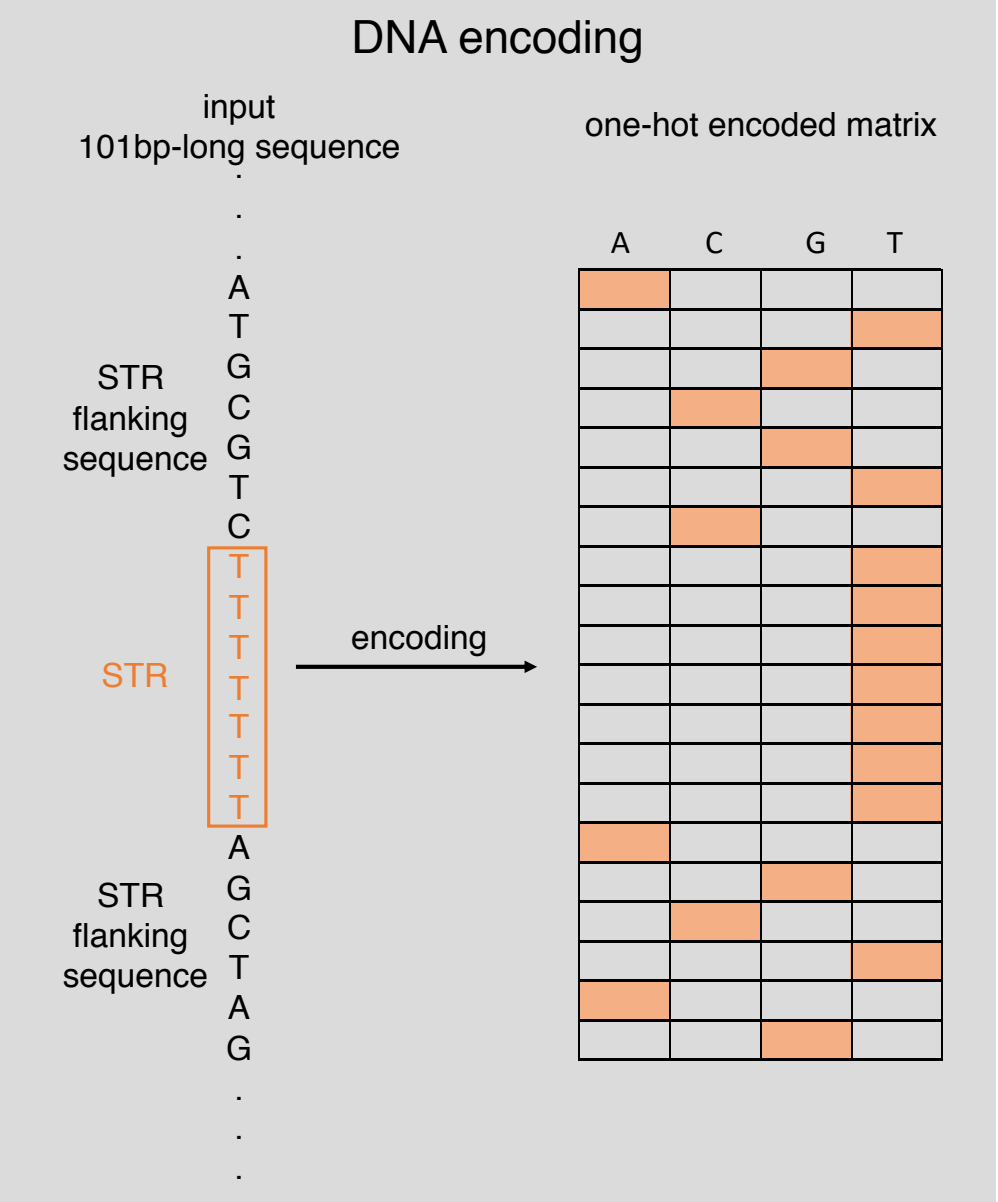
LASSO (Least Absolute Shrinkage and Selection Operator- Tibshirani R. J. of the Royal Statistical Society. 1996): by penalizing the absolute size of the regression coefficients (l1-norm), the LASSO drives the coefficients of irrelevant variables to zero, thus performing automatic variable selection.

Many publications: Li et al., PLoS CB 2014 ; Schmidt et al., NAR 2017 ; Bessière, Taha et al., PLoS CB 2018 ; Vandel et al., BMC bioinformatics 2019 ; Menichelli et al., PLoS CB 2021 ; Romero et al., bioRxiv 2022

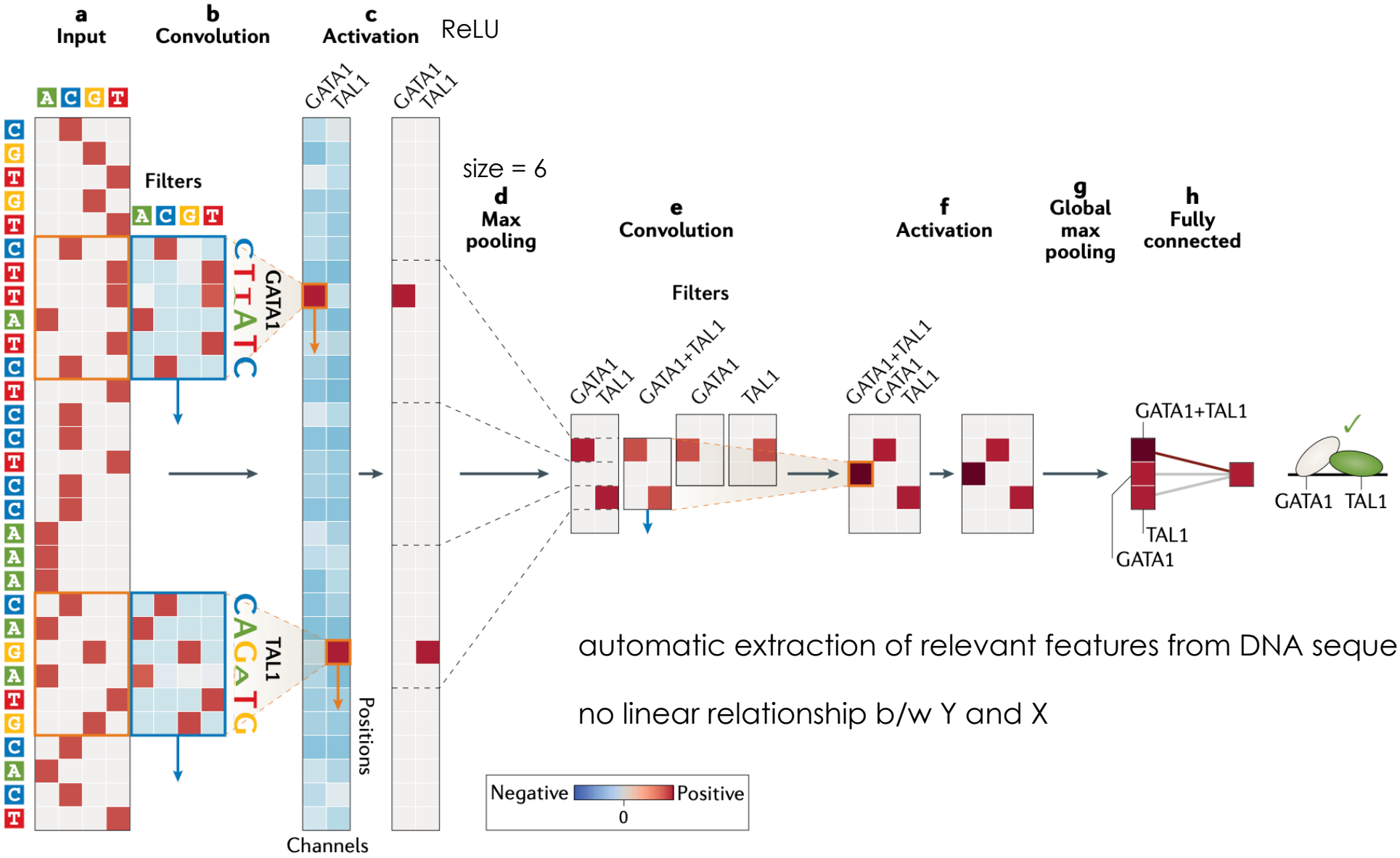
…

Another solution: deep learning (convolutional neural network, CNN)

No need of pre-defined variables as input == DNA sequence
(one hot-encoded ~ image)



Convolution Neural Network



automatic extraction of relevant features from DNA sequence

no linear relationship b/w Y and X

from Eraslan, G., Avsec, Ž., Gagneur, J. *et al. Nat Rev Genet* 2019

Finding the molecular rules | DNA grammar | *cis*-regulatory code

$$Y = f(DNA\ sequence)$$

$$RNA\ expression \mid TF\ binding \mid epigenetic\ marks = f(DNA\ sequence)$$

Whitaker et al., Nature Methods 2015
Alipanahi et al. Nat. Biotechnology 2015
Zhou et al., Nature Methods 2015
Quang and Xie, Nucleic Acids Res. 2016
Kelley et al., Genome Res. 2016
Zhou et al., Nature Genetics 2018
Bessière, Taha et al., PLoS CB 2018
Vandel et al., BMC bioinformatics 2019
Avsec et al., Nature Genetics 2021
Menichelli et al., PLoS CB 2021
Grapotte, Saraswat, Bessière et al., Nature Communications 2021
…

Many regulatory processes are directly predictable by DNA sequence

Decoding the DNA grammar

One example:

predicting transcription initiation at microsatellites

Grapotte, Saraswat, Bessière et al., Nature Communications 2021

Microsatellite | short tandem repeats (STRs)

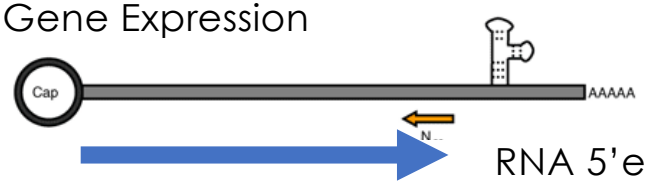
. short repetitive motifs of 1–6 nucleotides
. tandemly repeated
. flanked by unique sequences

Microsatellites	Unique sequence Repeat units Unique sequence
Mononucleotide	---GGTAGCCAA A A A A (A)n CGATCCA---
Dinucleotide	---TCGCATGCA CA CA (CA)n ATTCGCA---
Trinucleotide	---TTAGCATCAG CAG (CAG)n CCAGTGA---
Tetranucleotide	---AATGGTACCGG (CCGG)n GTCACGT---
Pentanucleotide	---CGATGATCCAAG (CCAAG)n TTACGTA---
Hexanucleotide	---GCTAAGGCCATTG (CCATTG)n ACTGTCA---

https://link.springer.com/referenceworkentry/10.1007%2F978-3-642-16483-5_3731

~3% of the human genome

CAGE: Cap Analysis of Gene Expression



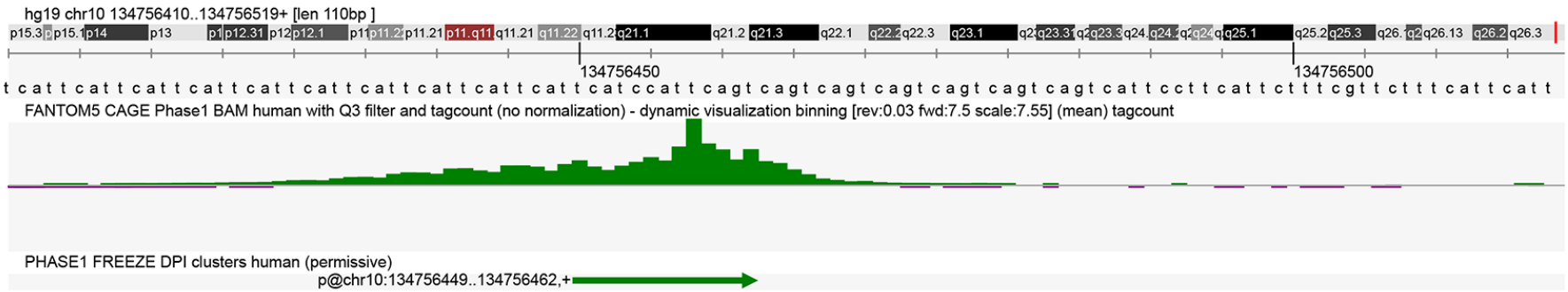
RNA 5'end (27nt)

Forrest et al., Nature 2014

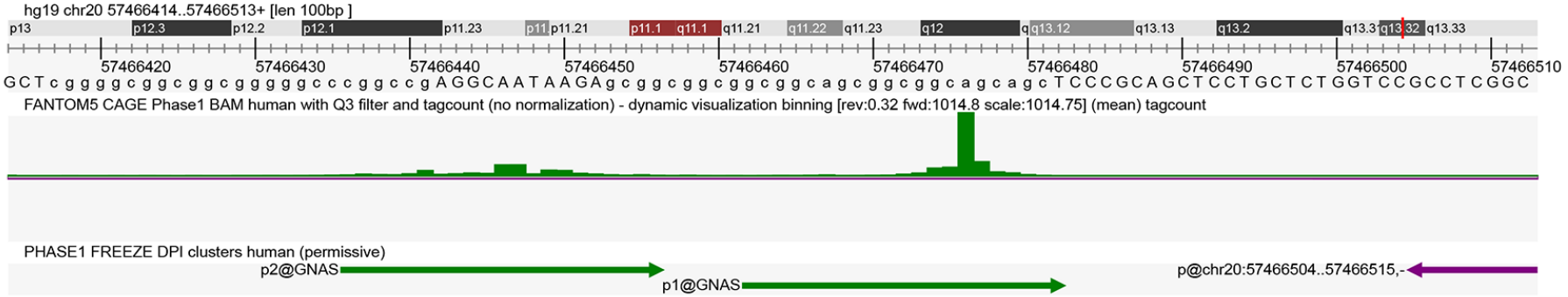
1,048,124 human CAGE peaks in **1,829** libraries
753,303 not assigned to genes (~**72**%)
89,948 CAGE peaks (~**8.6**%) initiate at STRs*

* 84,555 STRs from HipSTR catalog
Willems et al., Nat. Methods 2017

chr10:134756410-134756519;Human_STR_203176;ATTC

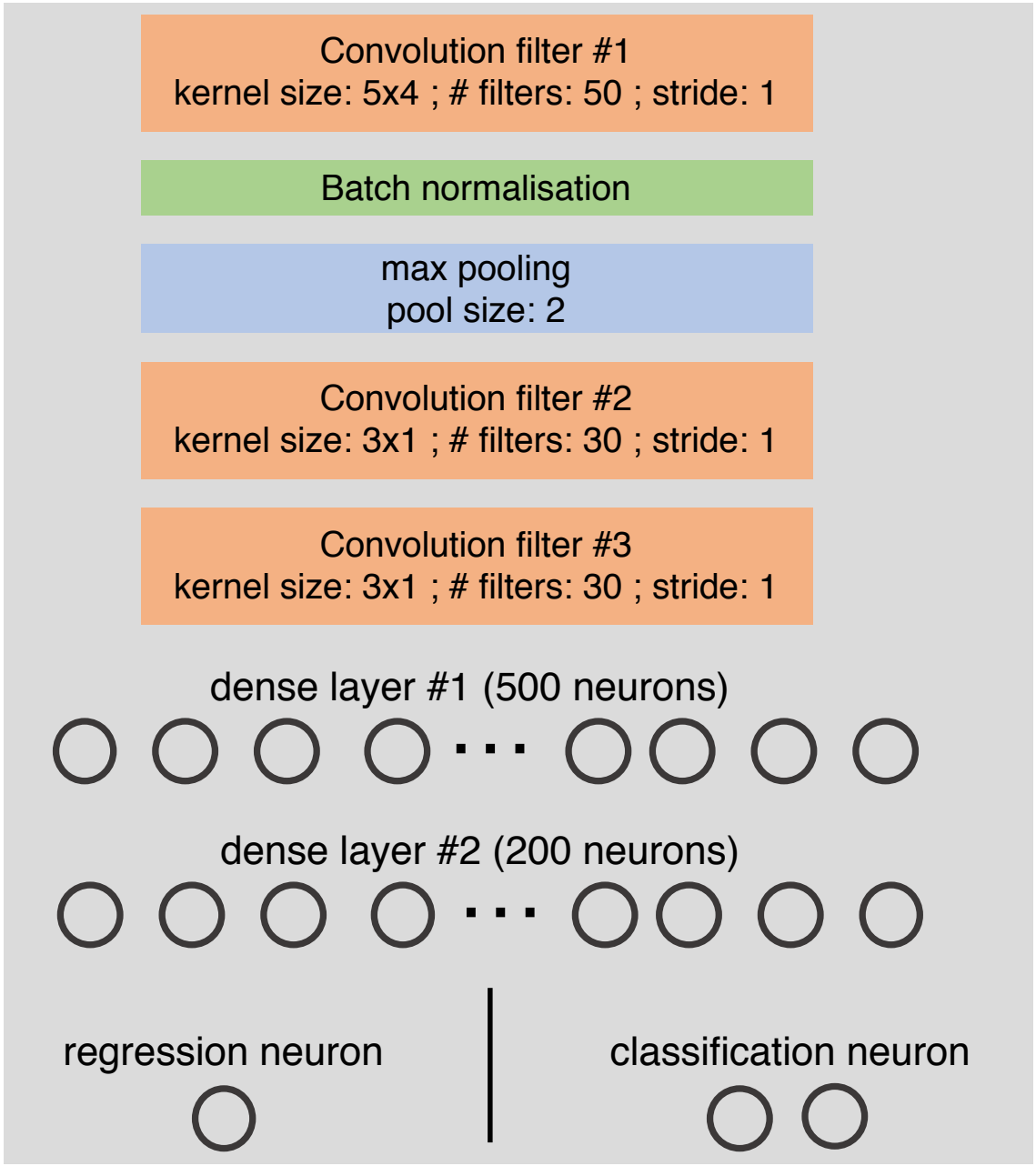


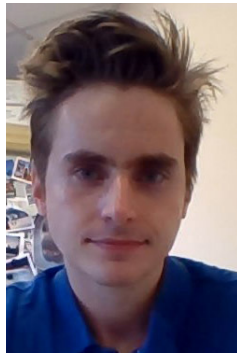
chr20:57466453-57466475;Human_STR_865890;CGG



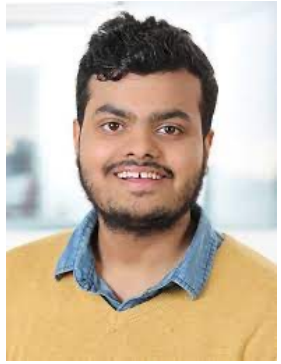
Grapotte, Saraswat, Bessière et al., Nature Communications 2021

CNN to predict transcription initiation at STRs





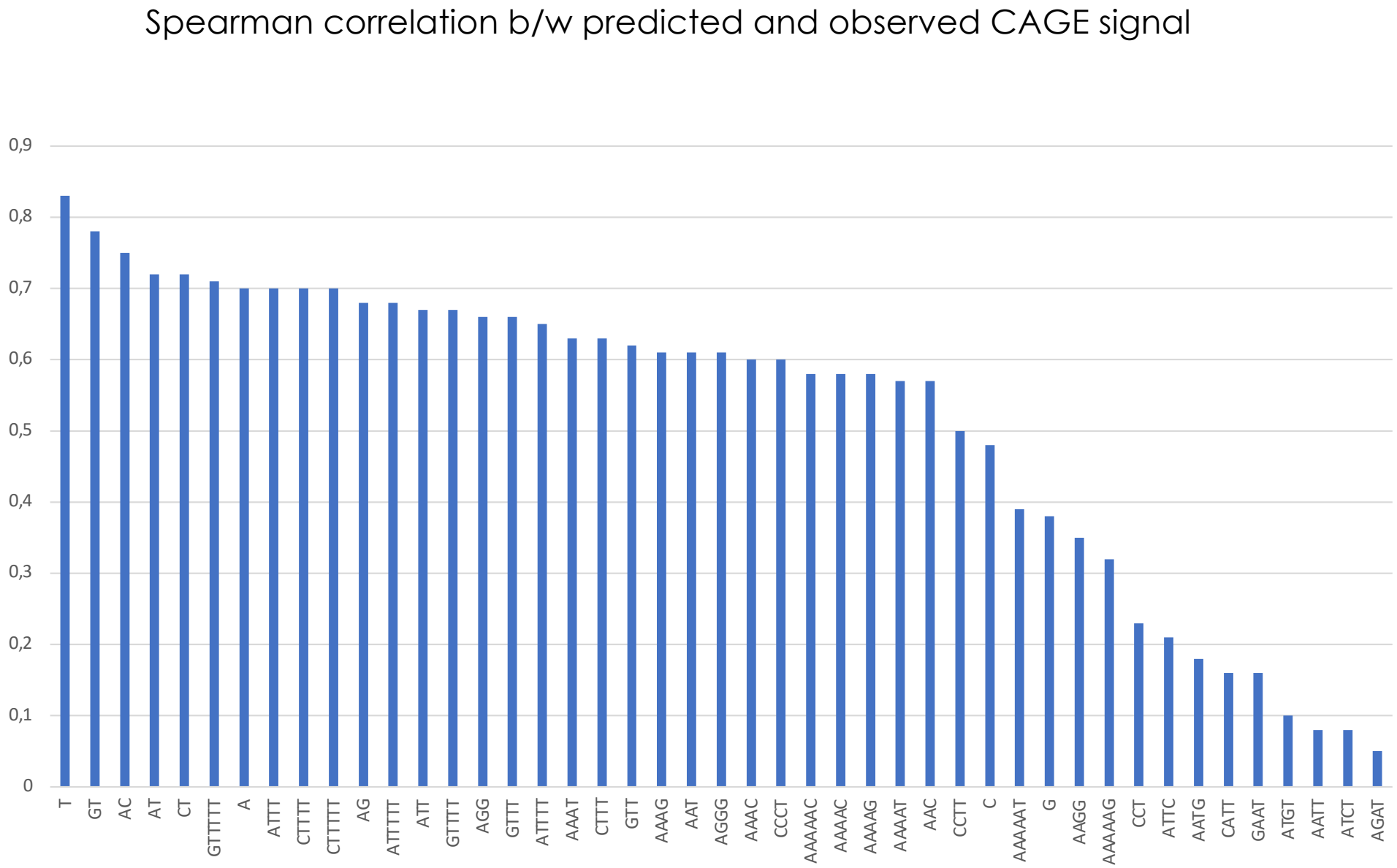
Mathys Grapotte



Manu Saraswat

https://gite.lirmm.fr/ibc/deepSTR

CNN results


Spearman correlation b/w predicted and observed CAGE signal
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
T
GT
AC
AT
CT
GTTTTT
A
ATTT
CTTT
CTTTT
AG
ATTTT
ATT
GTTTT
AGG
GTTT
ATTTT
AAAT
CTTT
GTT
AAAG
AAT
AGGG
AAAC
CCCT
AAAAAC
AAAAC
AAAAG
AAAAT
AAC
CCTT
C
AAAAAT
G
AAGG
AAAAAG
CCT
ATTC
AATG
CATT
GAAT
ATGT
AATT
ATCT
AGAT

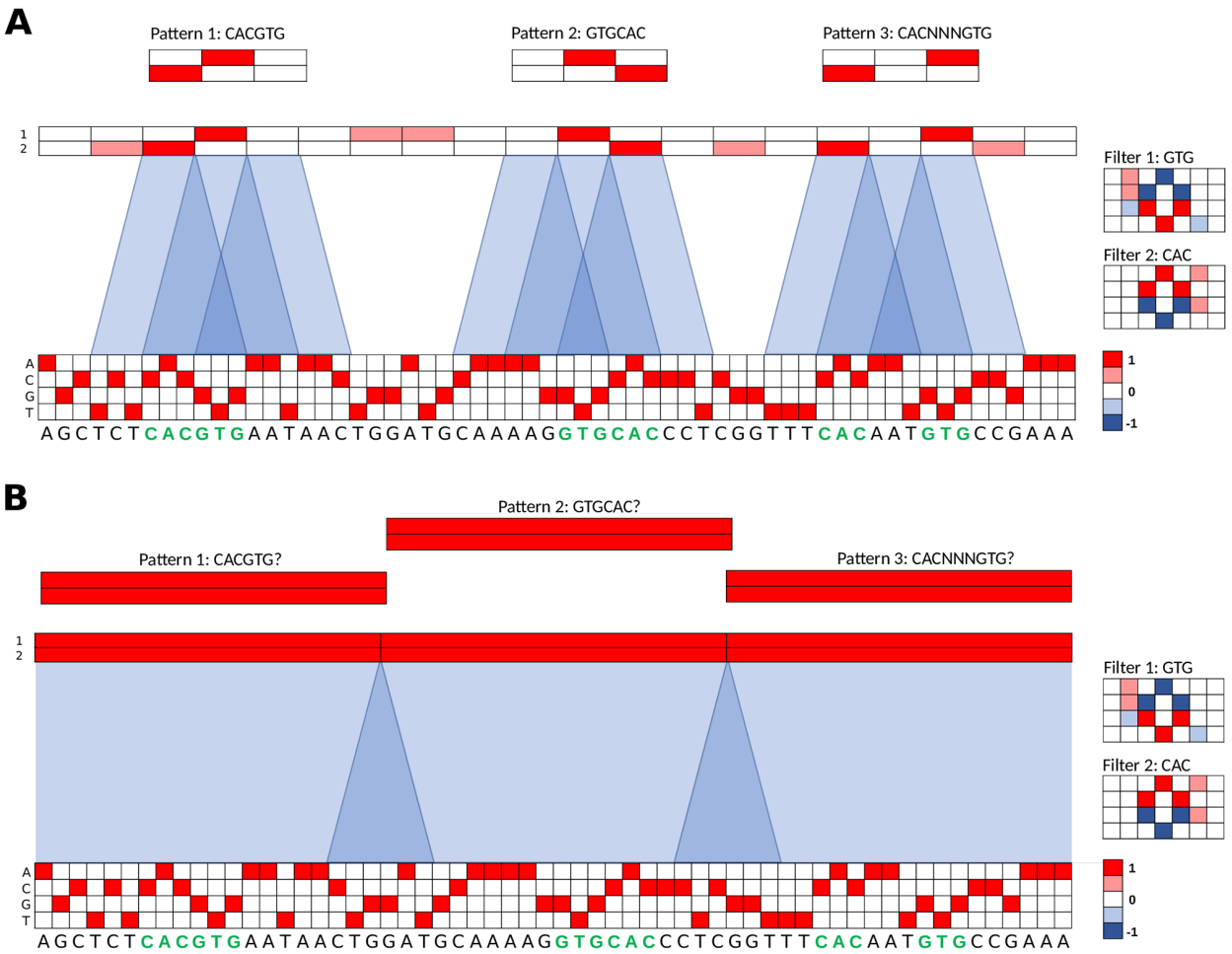
Importance of CNN architecture

CNN architecture, which is optimized for prediction, influences the representations of sequence motifs captured by first layer filters.

Prediction and interpretation require distinct CNN architectures, in particular adapting max-pooling and convolutional filter size

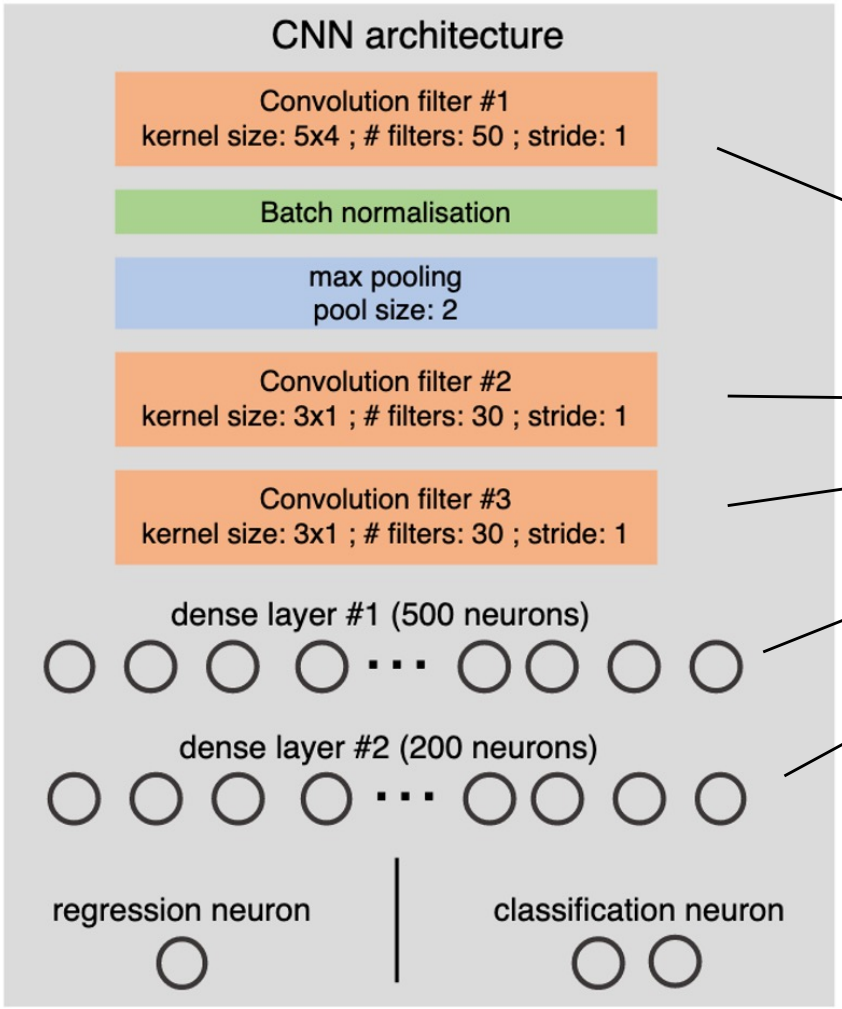
max-pooling size = 3

max-pooling size = 20



from Koo & Eddy, PLoS CB 2019

The black box issue



#parameters == 776,580
(small model)

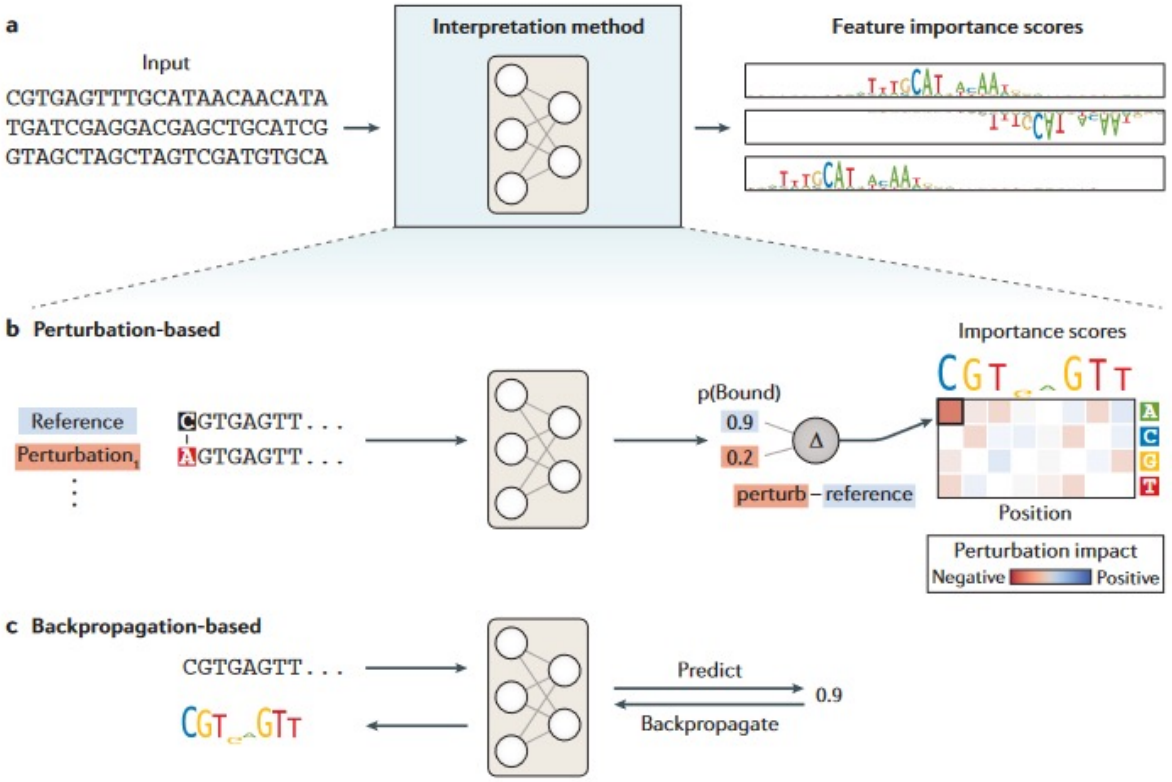
features used by the model
not directly interpretable

Black Box issue: a model is accurate but we do not know **how and which information** it uses to make the prediction.

>>> need **interpretation methods**

Interpreting CNNs: determining feature importance scores

GOAL: attribute feature importance scores to highlight the parts of a given input that are most influential for the model prediction and thereby help to explain why such a prediction was made. When DNA is used as input, importance scores highlight sequence motifs.



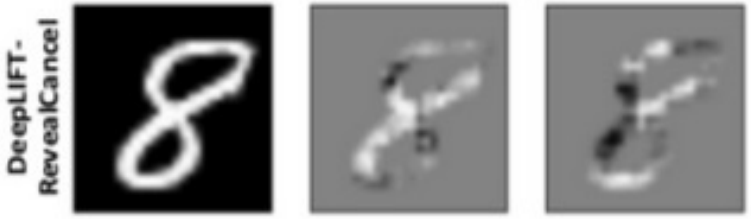
from Eraslan, G., Avsec, Ž., Gagneur, J. *et al. Nat Rev Genet* 2019

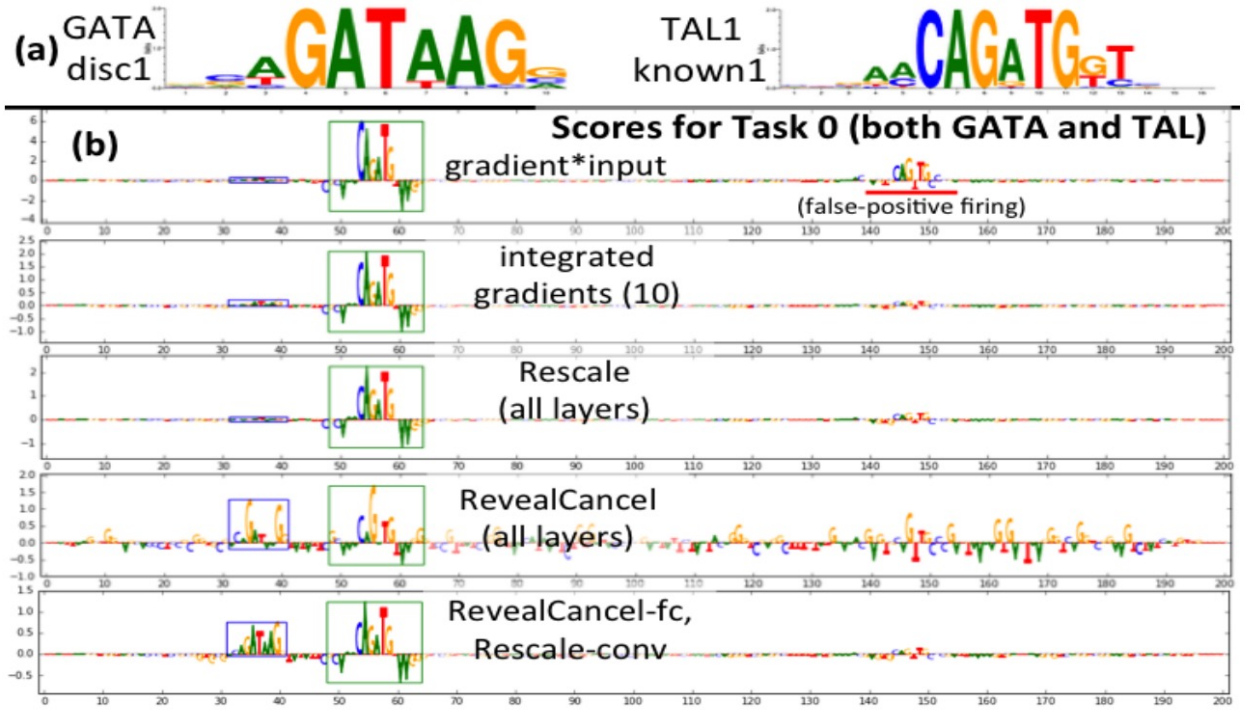
Interpretation methods use a **black box** model and a **sequence of interest** as the input and, then (i) mutate the input and evaluate the consequences on the output (**perturbation-based**) or (ii) backpropagate the output relative to the input (**backpropagation-based**)

Example of backpropagation-based method: DeepLIFT

backpropagation approaches propagate an importance signal from an output neuron backwards through the layers to the input in one pass and generate a **saliency map**.

DeepLIFT explains the difference in output from some 'reference' output in terms of the difference of the input from some 'reference' input.

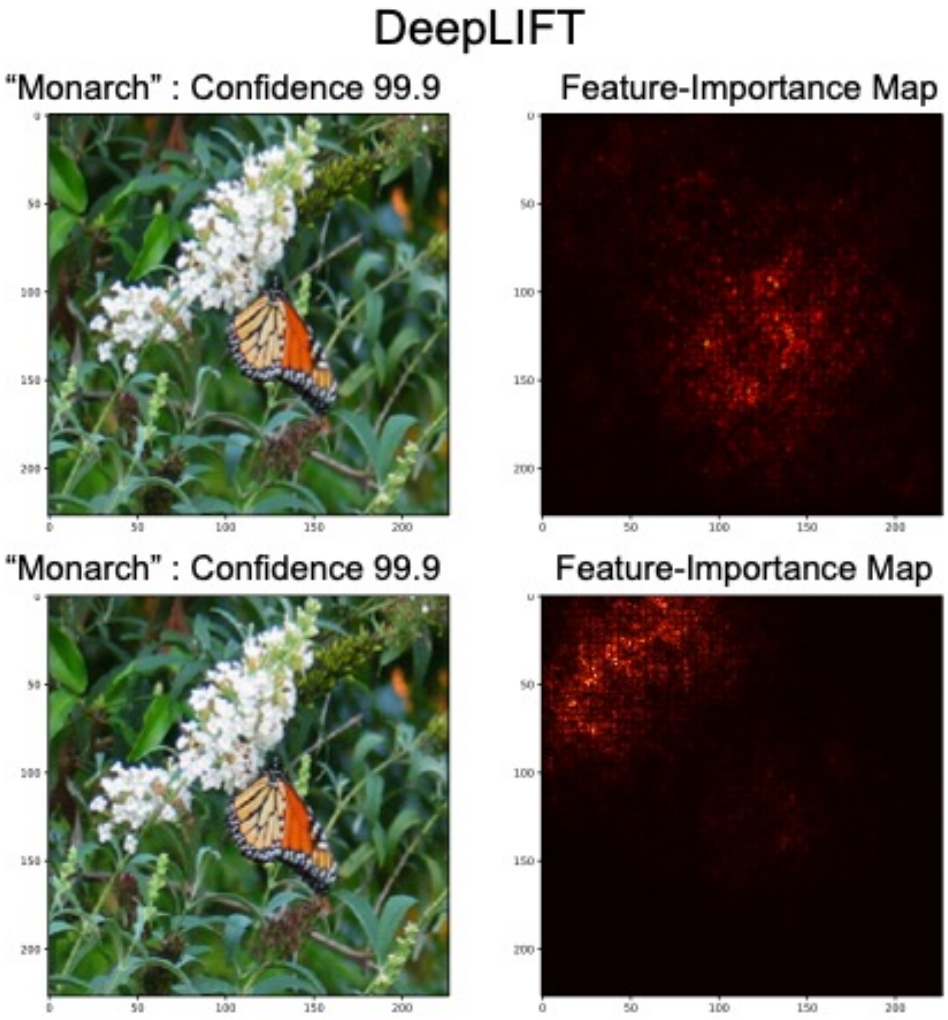




from Shrikumar et al. Arxiv, oct. 2019

Limits of interpretation methods (ii)

Saliency maps depend on the chosen input



which features to choose?

from Ghorbani, Abid & Zou, The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

top row shows the the original images and their saliency maps and the bottom row shows the perturbed images

Limits of interpretation methods (i)

. Explainable model - two steps :

(i) Training step and

(ii) Interpretation step

. In the interpretation step, a subset of sequences is used to explain the model, but ***are they truly representative?***

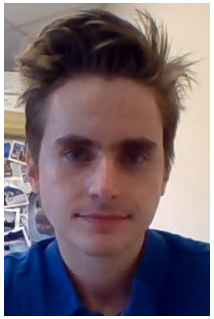
. Besides, how do we know ***which features*** are truly important for the prediction?

. Need models that are **interpretable in the first place**
(Rudin, Nature Machine Intelligence 2019)

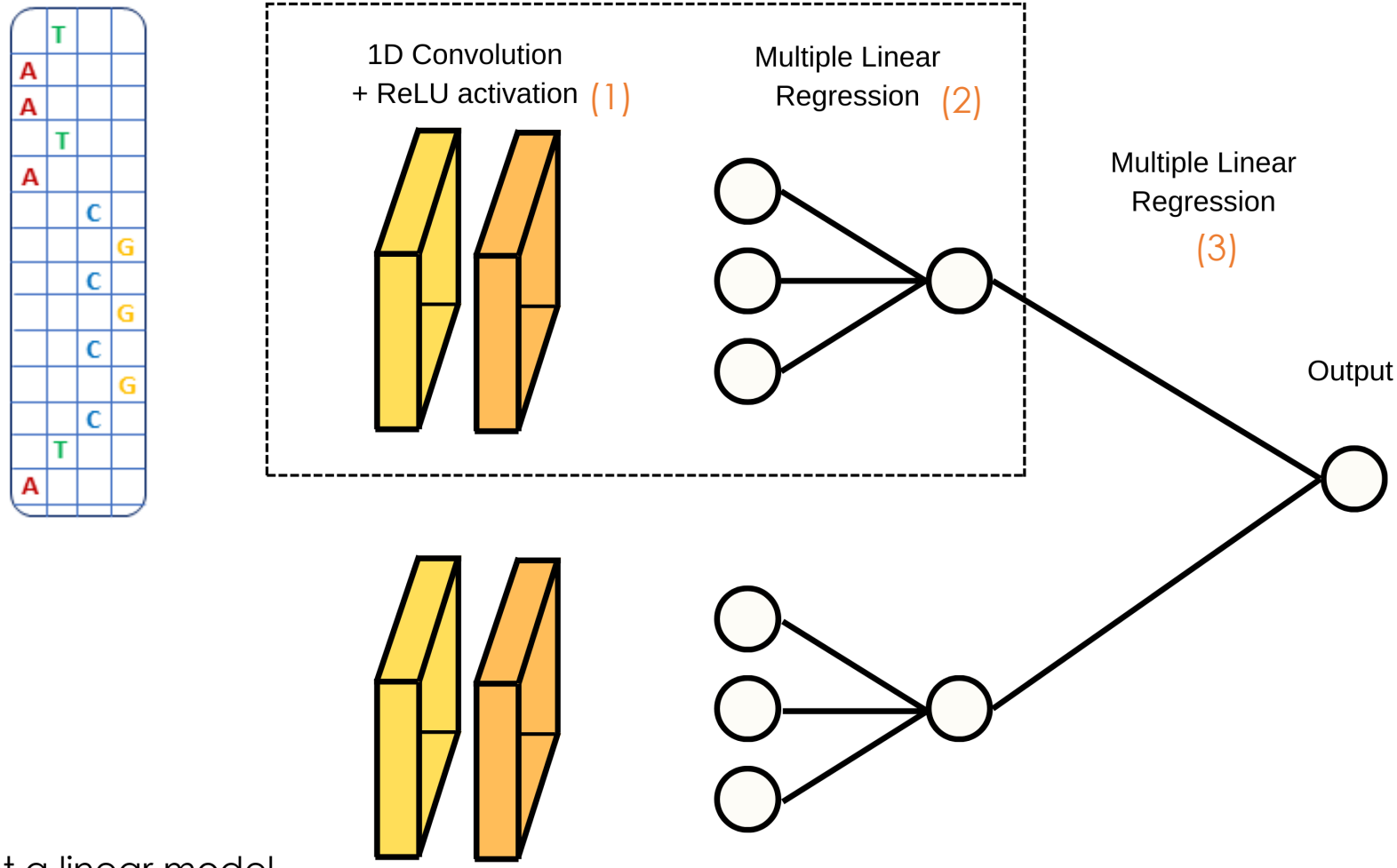
i.e. ***combine training & interpretation*** and therefore do not rely on any input for interpretation

. Need also to infer features **directly from DNA sequence**

Modular Neural Networks



Mathys Grapotte

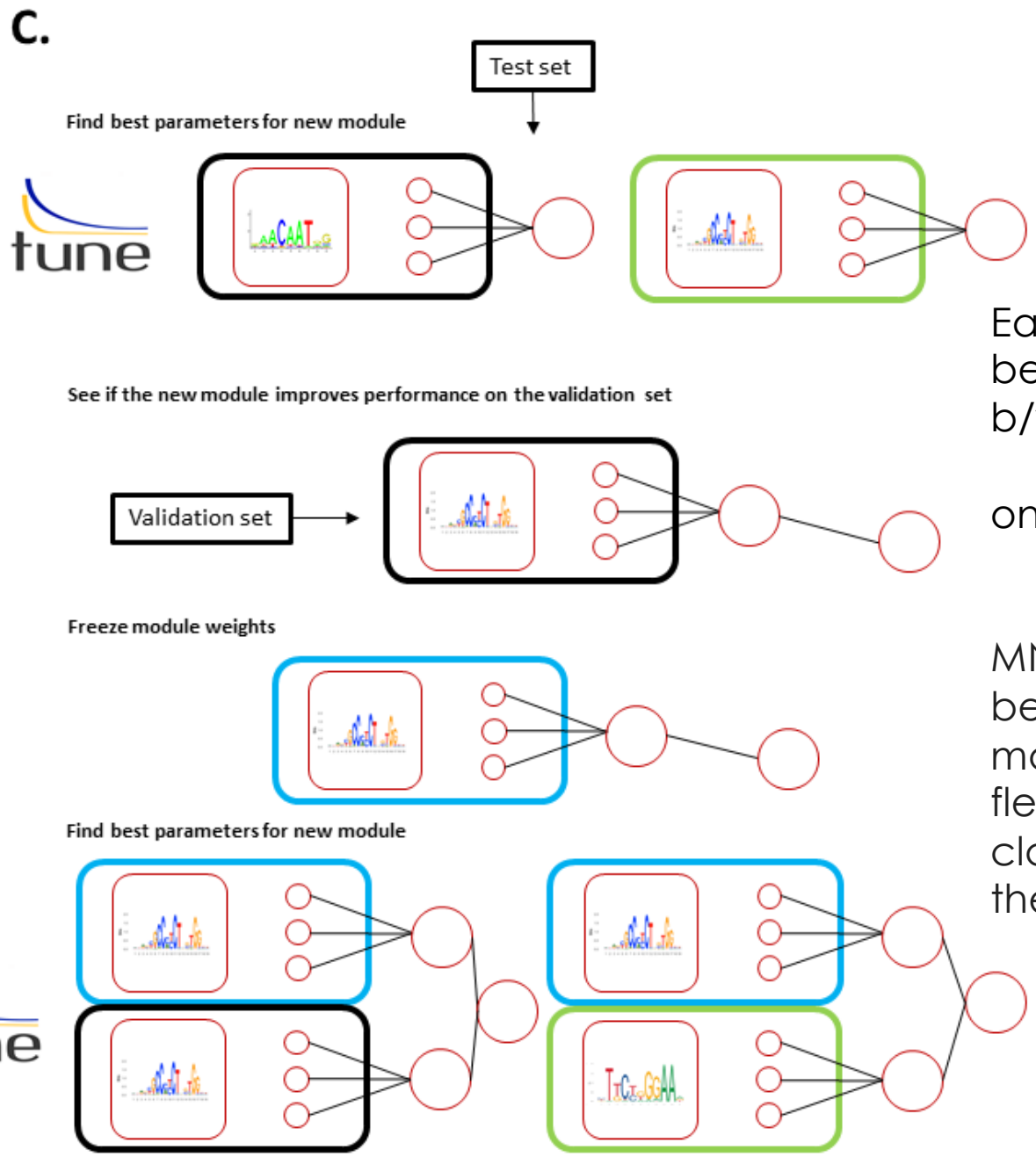


(1) not a linear model

(2) the coefficients of the regression indicate **where** the model is expected to find the filter/motif

(3) the coefficients indicate **which** module/filter/motif is important

learning an MNN



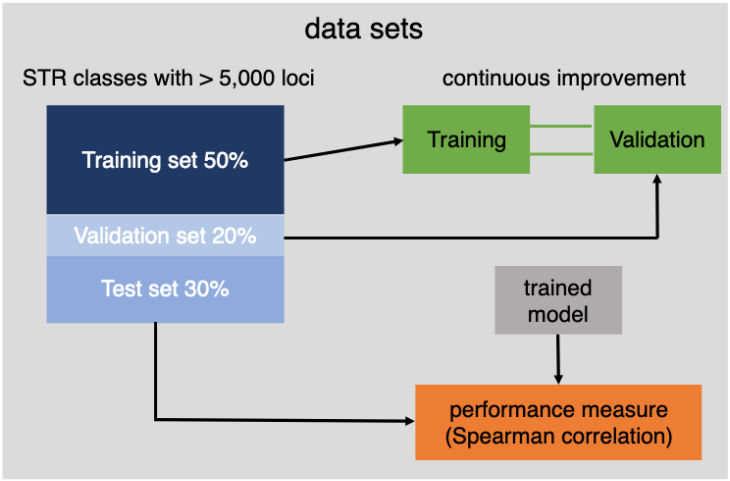
Each filter of each module cannot be combined with another (no link b/w filters)

one module > one filter > one motif

MNN is not a mere 1-layer CNN because filters from different modules can have various sizes, a flexibility that is not allowed in a classical CNN wherein all filters from the same layer have the same size.

CNN vs. MNN

CNN 776,580 parameters
MNN ~1,000 parameters



STR Class	CNN-Model * (Spearman)	Modular Network (Spearman)	Number of filters
T	0.87	0.79	9
AC	0.78	0.72	9
GT	0.78	0.72	8
A	0.73	0.69	7
AG	0.72	0.64	4
AGG	0.57	0.51	5
AAT	0.62	0.61	10

** The CNN-Model contains 110 filters*

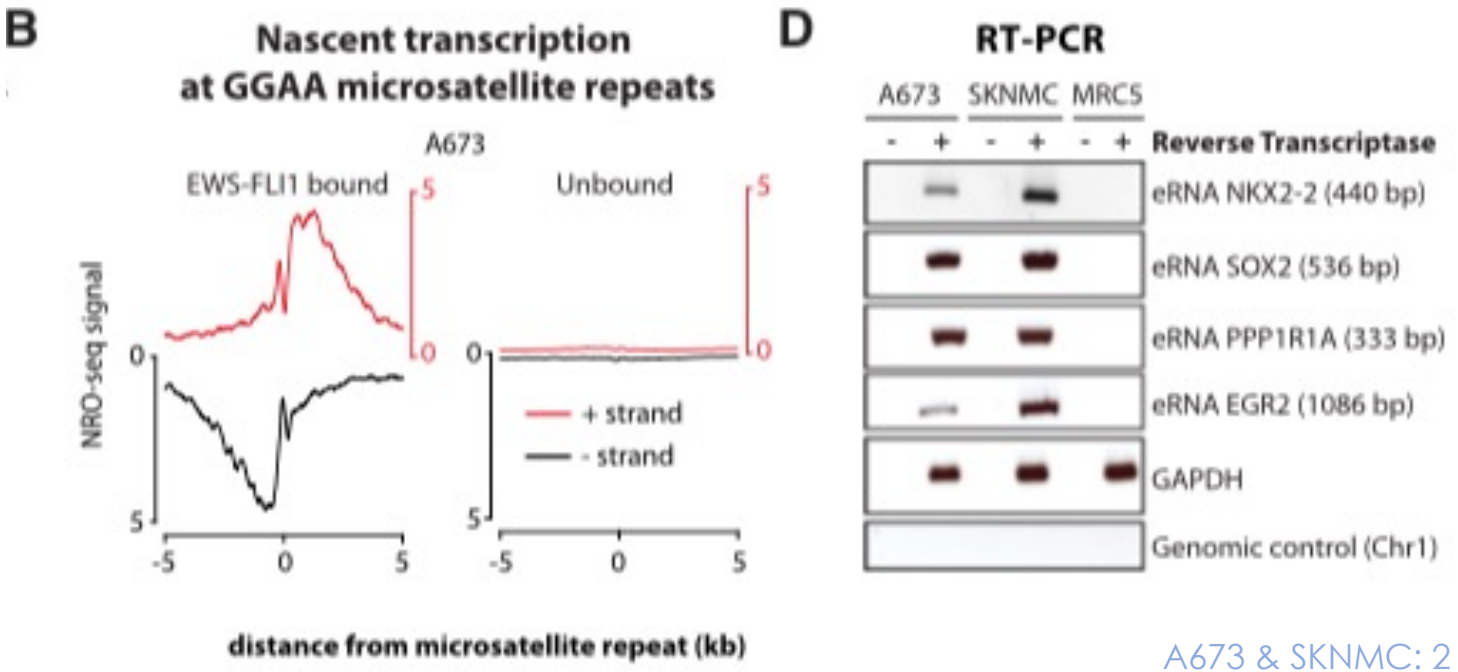
MNN to probe the functional consequences of transcription initiating at STRs

STRs are implicated in various regulatory functions

Process	Gene (Organism)	Repeat Motif	References
Binding of transcription factors to microsatellite DNA	SLC11A1 (human)	GT (imperfect)	Bayele etal. 2007
	ECE-1c (human)	CA (imperfect)	Taka etal. 2013
	TH (human)	TACT	Li etal. 2012
	PIG3 (human)	TGYCC	Albanese etal. 2001
	nadA (N. meningitidis)	TAAA	Contente etal. 2002, Martin etal. 2005
Spacing between promoter elements	GP91-PHOX (human)	CA	Uhlemann etal. 2004
	IGF1 (human)	CA	Chen etal. 2016
Long-range interactions	Intergenic (Drosophila & huma	GATA	Kumar etal. 2013
Transcription start site selection	HO-1 (human)	AC	Kramer etal. 2013
	ECE-1c (human)	CA (imperfect)	Li etal. 2012
Transcription end site selection	ASS1 (human)	GT	Tseng etal. 2013
RNA half-life	FGF9 (human)	TG (imperfect)	Chen etal. 2007
Alternative splicing	APOA2 (human)	GT	Cuppens etal. 1998
	CFTR (human)	TG	Hefferon etal. 2004
	eNOS (human)	CA	Hui etal. 2003
	Various (human)	CA	Hui etal. 2005
Nucleosome packaging	HIS3 (S. cerevisiae)	A	Iyer and Struhl 1995
	CSF1 (human)	TG	Liu etal. 2001, Liu etal. 2006
	CYC1 (S. cerevsiae)	CG	Wong etal. 2007
	Genomic (human)	BAA	Zhao etal. 2015
Histone modification	Genomic (human)	Various	Gymrek etal. 2016
Methylation	Genomic (human & chimpanze	CG	Fukuda etal. 2013
	Genomic (human)	CG	Quilez etal. 2016
Noncoding RNA function	Genomic (Drosophila)	AAGAG	Pathak etal. 2013
	Genomic (mammals)	GAA	Zheng etal. 2010
Meiotic recombination	ARG4 (S. cerevisiae)HIS4 (S. ce	TGCCGNN	Gendrel etal. 2000, Kirkpatrick etal. 1999
	Genomic (A. thaliana)	CCT & CCN	Choi etal. 2013, Shilo etal. 2015

Bagshaw, Genome Biol Evol 2017

EWS-FLI1-bound GGAA microsatellite repeats are transcriptionally active



from Boulay et al., Genes & Dev. 2018

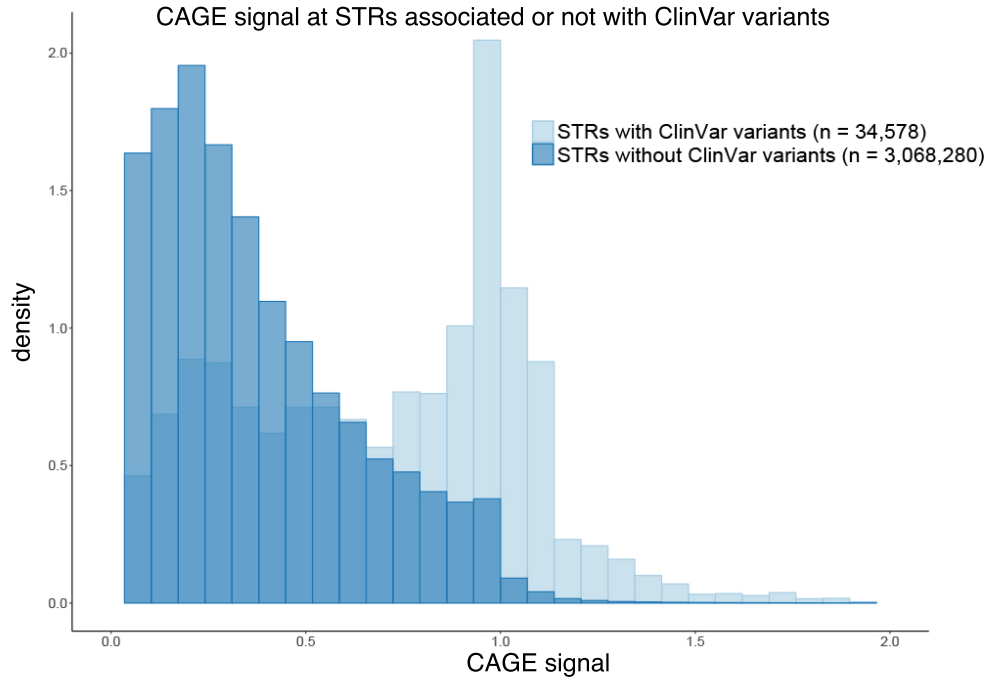
A673 & SKNMC: 2 Ewing Sarcoma cells*
MRC5: fibroblasts – negative control (no EWS_FLI)

*not tested in FANTOM5
but transcription initiation detected at several GGAA STRs in non-sarcoma cells

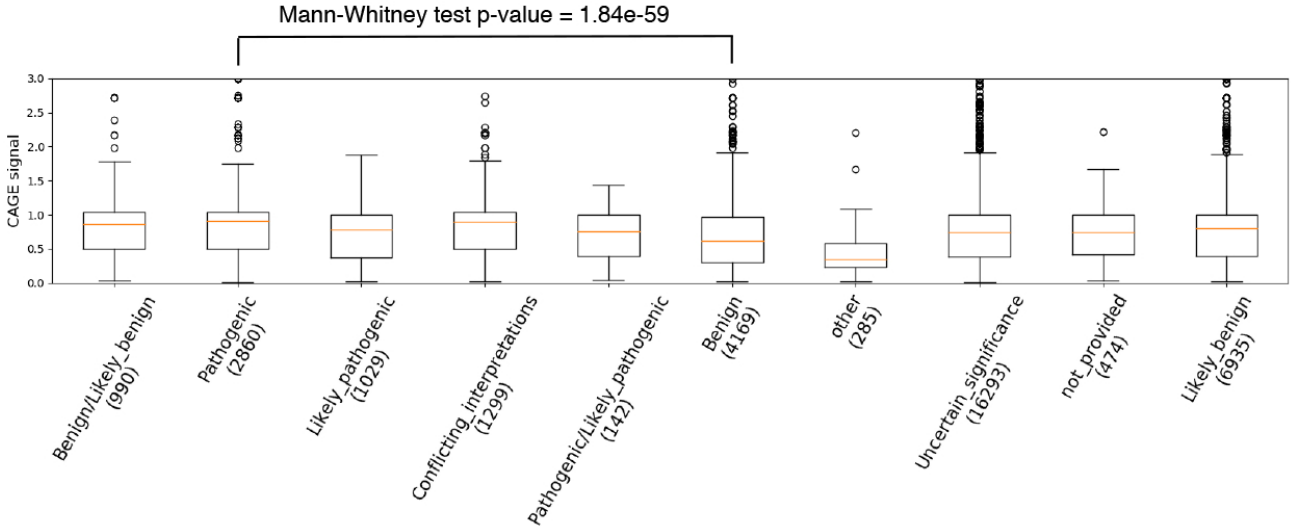
. Transcription inhibition is feasible by targeting immediately adjacent unique genomic sequences with CRISPR/dCas9-KRAB

. Silencing of a single EWS-FLI1-bound GGAA repeat enhancer 470 kb from the SOX2 locus is sufficient to impair the growth of Ewing sarcoma xenografts.

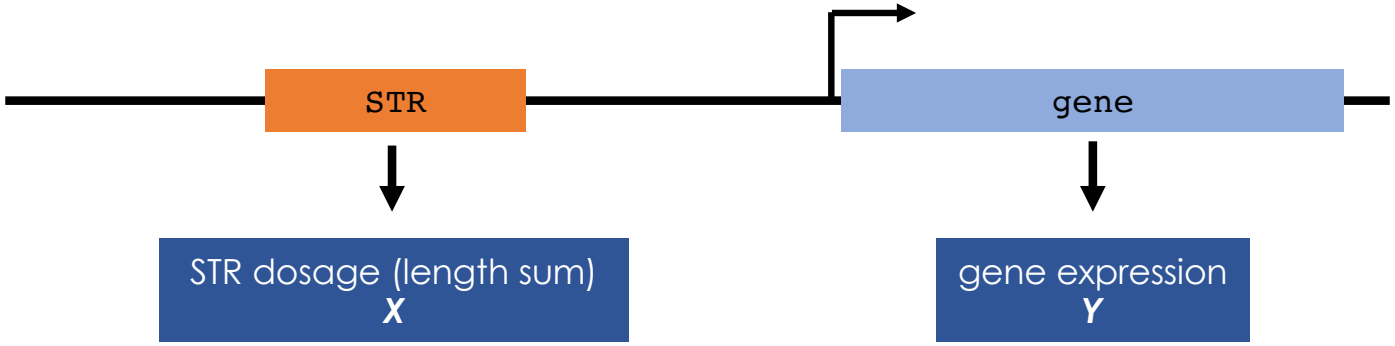
pathogenic variants are associated with STRs initiating transcription

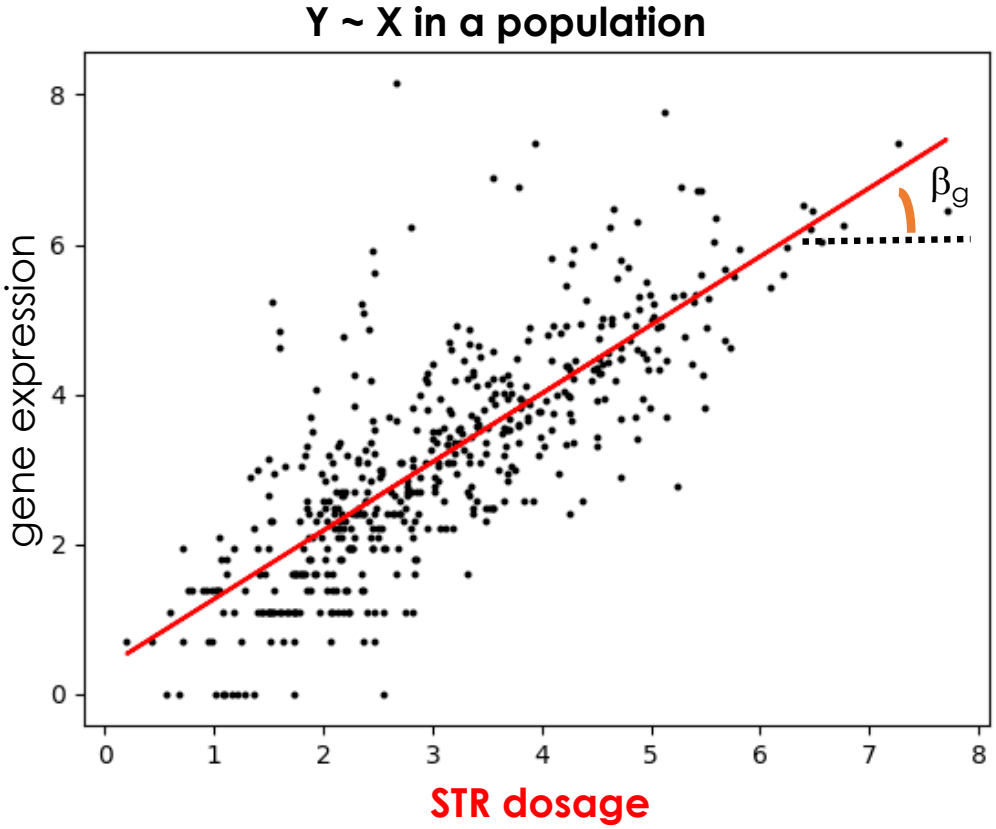


CAGE signal at STRs associated with ClinVar variants - one-way ANOVA test p-value p-value = 2.5e-27



STR-mediated long-range regulations
expression STR (eSTR) Gymrek et al., Nature Genetics 2016

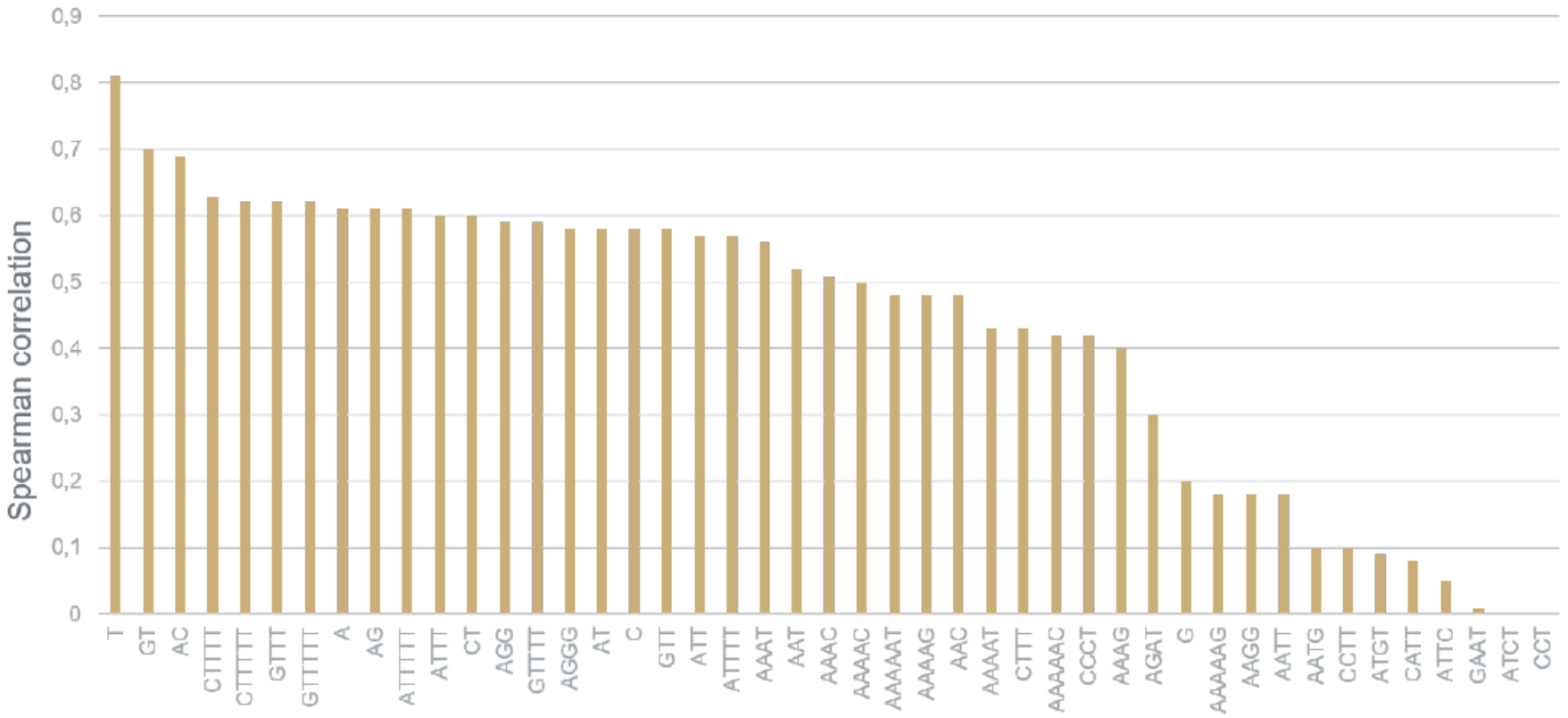




For all gene/STR pairs:

$$Y = \beta_0 + \beta_g X + \varepsilon \quad \begin{cases} H_0: \beta_g = 0 \\ H_1: \beta_g \neq 0 \end{cases}$$

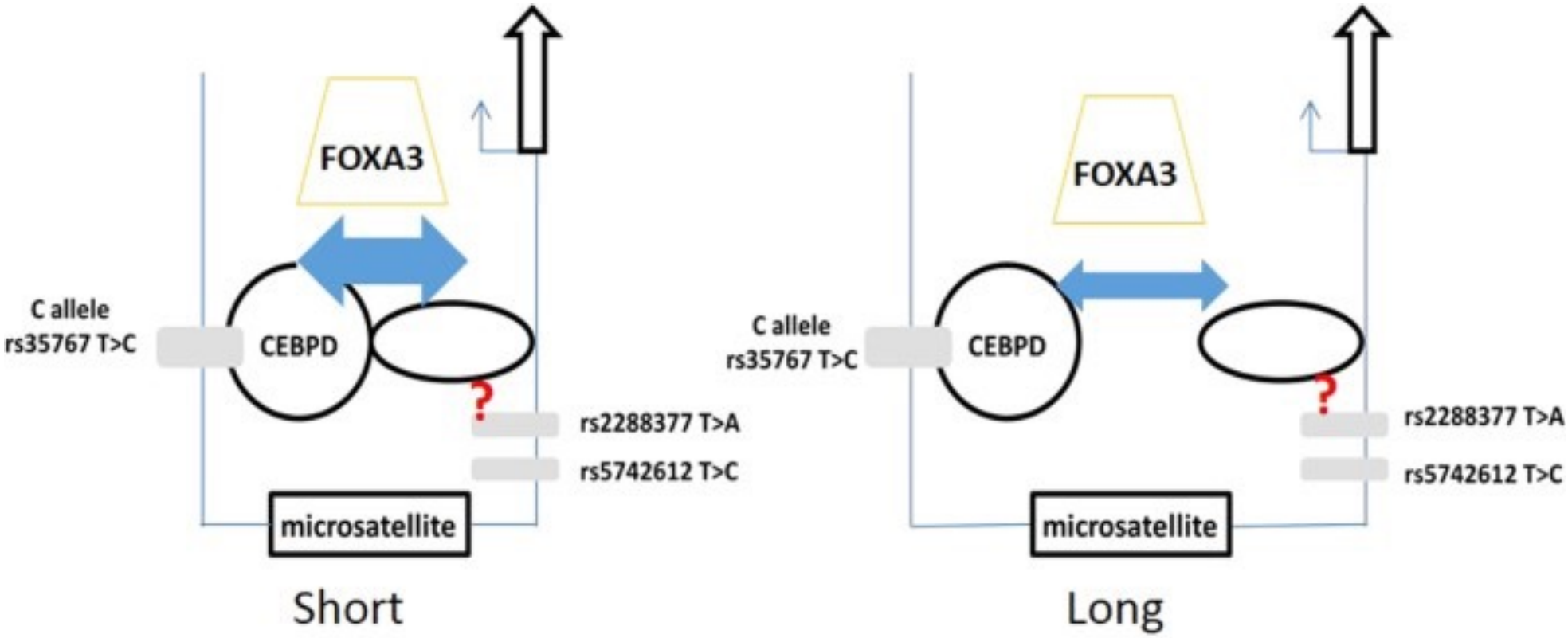
Model accuracies using as input only flanking sequences



Example of sequence used as input in D:

```
>Human_STR_873642;T;+
CAGTCAAAGGTAATCATGGGTCCTTACCTGGGGTCTAGTAGAGCTTACAG-NNNNNNNNN-CTTAacaaaaatatctgtacatgtttcatgtacacaatttagtgcatatg
```

eSTR require adjacent SNPs



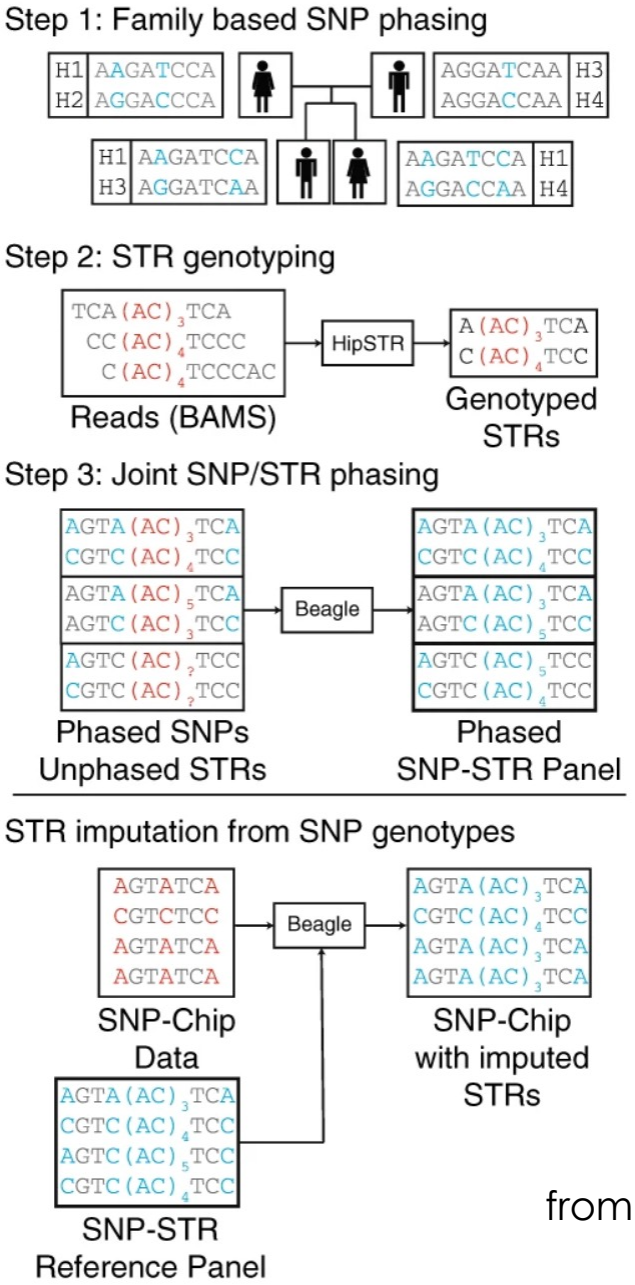
from Chen et al., Scientific Reports, 2016

'adjacent SNPs located outside the STR were required for the STR to function as eSTR.'

Functional link b/w STR and surrounding sequences

SNPSTR *(Mountain et al., Genome Res 2002 ; Agrafioti et al., Nucleic Acids Res. 2007)*

STR length imputation with surrounding SNPs (Saini et al., Nat. Comm 2018)

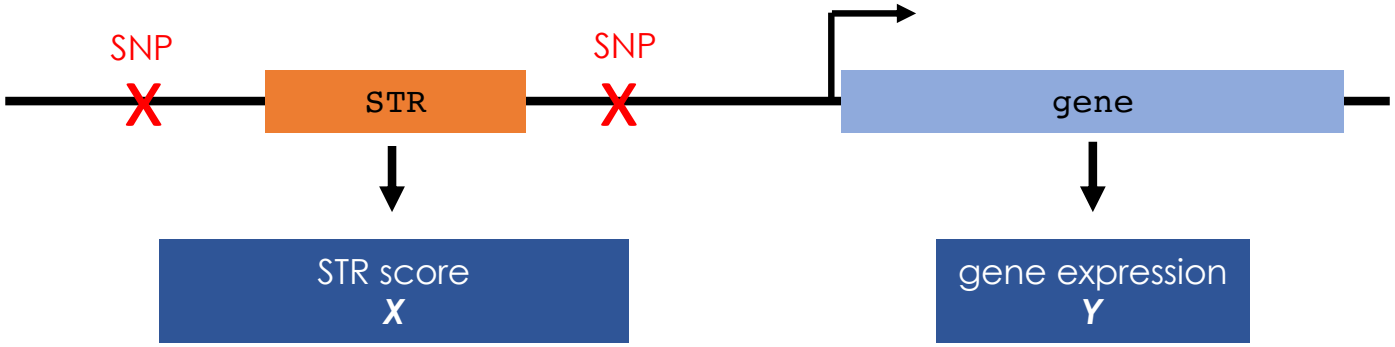


50kb around STRs

from Saini et al., Nat. Comm 2018

Taking into account STR surrounding sequences and SNPs located within these sequences appears instrumental to better understand regulations orchestrated by STRs

STR-surrounding SNPs (i.e. SNPSTR) should be taken into account for eSTR computation

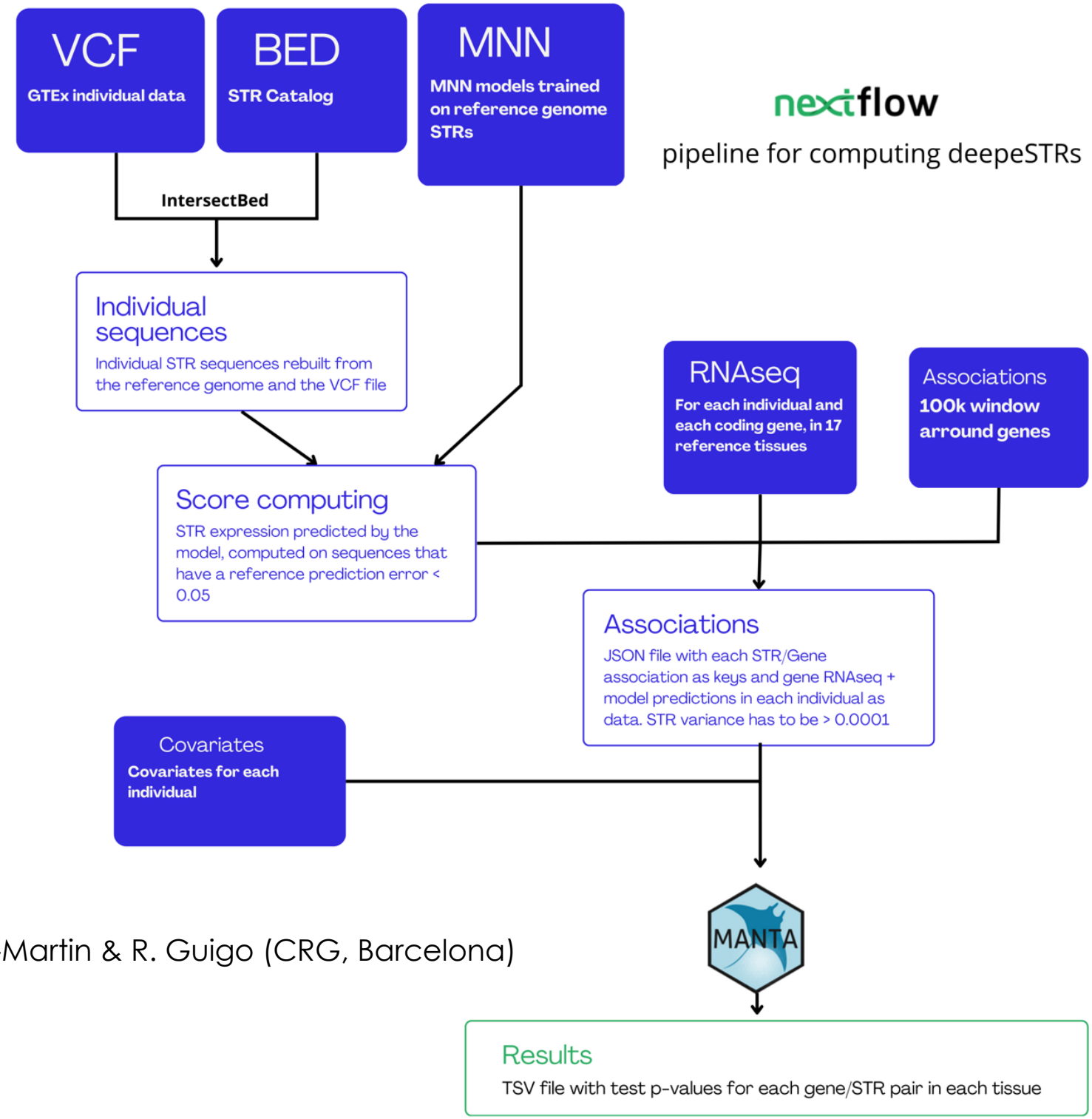


Combination of SNPs and haplotypes

For each STR, 2 scores (2 alleles)
=> Multivariate regression

GTEx

no CAGE data
(i.e. no information on Tx at STRs)



D. Garrido-Martin & R. Guigo (CRG, Barcelona)

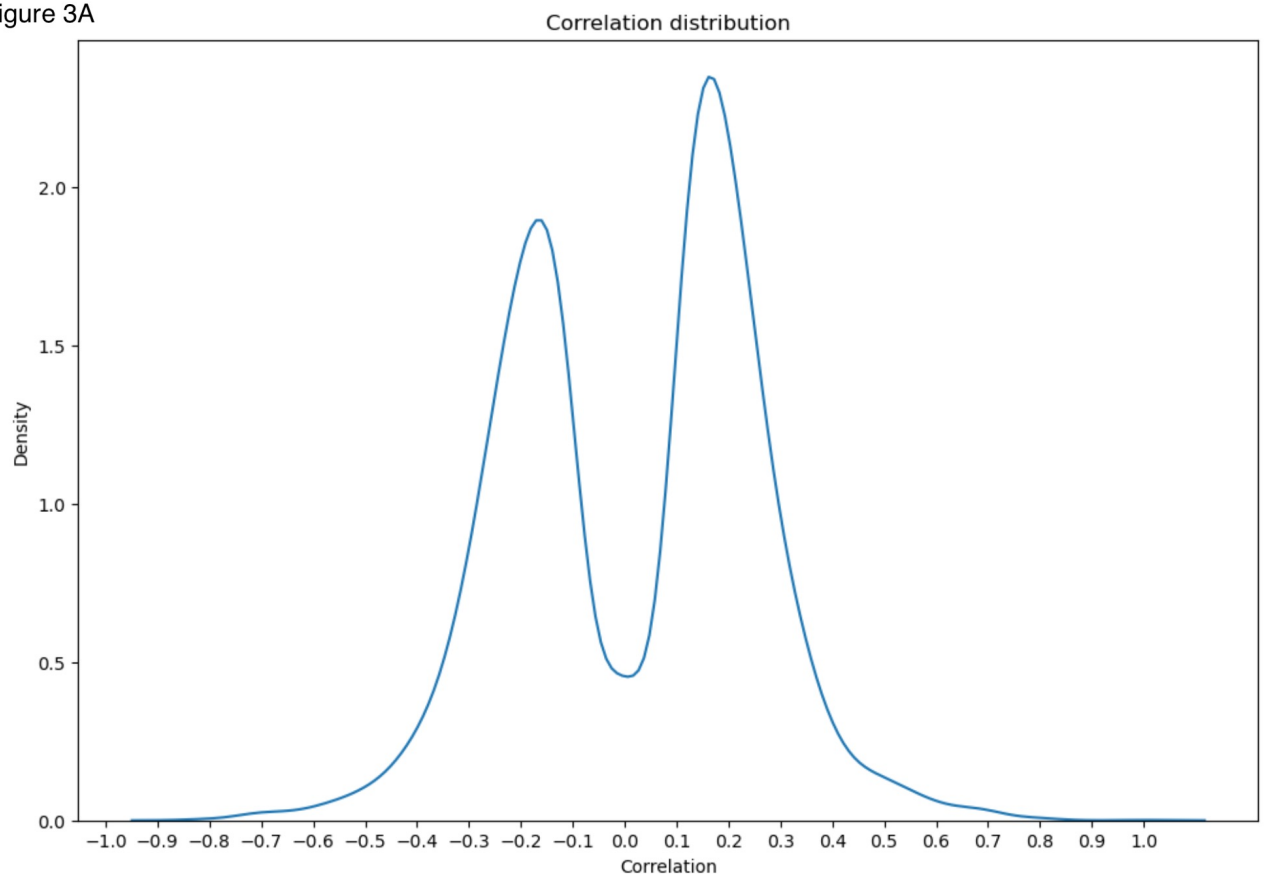
115,608 (STR, gene) pairs tested – 17 tissues as in Fotsing et al., Nature Genetics 2019

14,340 significant associations
between 8,458 STRs and 11,060 genes in at least one tissue (FDR = 0.05)

enrichment in previously published
eQTLs (GTEx, Nature 2017)
sQTLs (Garrido-Martin et al, Nature Communications, 2021)
eSTRs (Fotsing et al., Nature Genetics 2019)

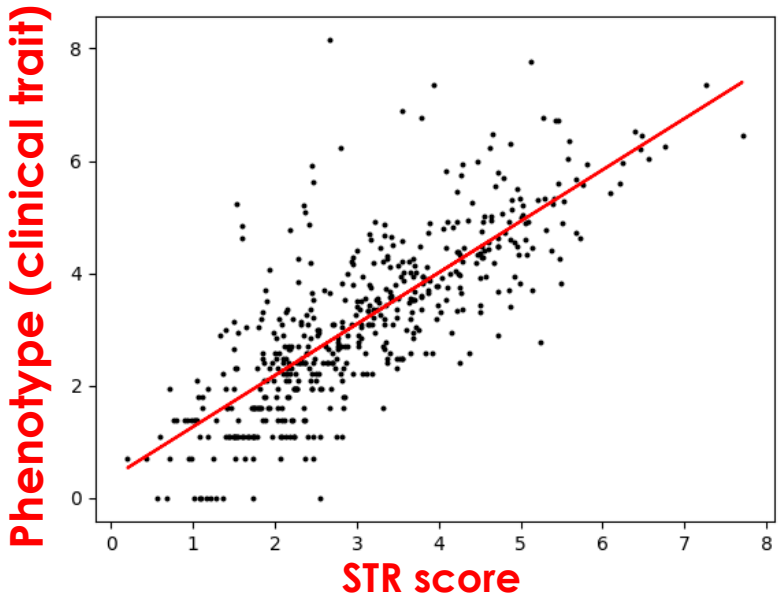
positive and negative impact on gene expression

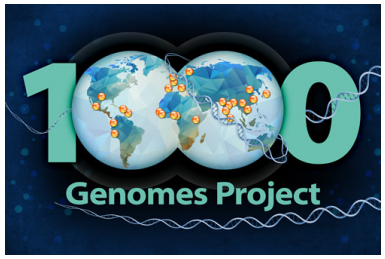
Figure 3A



Likewise in Genome Wide Association Study (GWAS/PheWAS)

Clinical trait ~ STR score





~2.5k participants

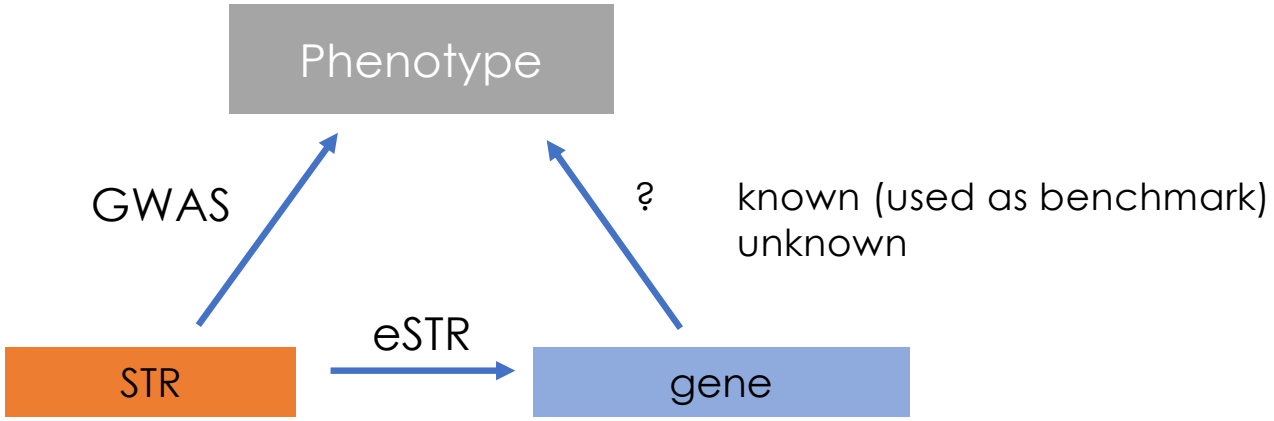
FINNGEN

~500k participants

SANOFI

Collaboration with Sanofi

combine both approaches to verify that STR associated to specific traits regulate genes implicated in those traits.



better interpret > 30,000 variants located around STRs

and often in non-coding regions

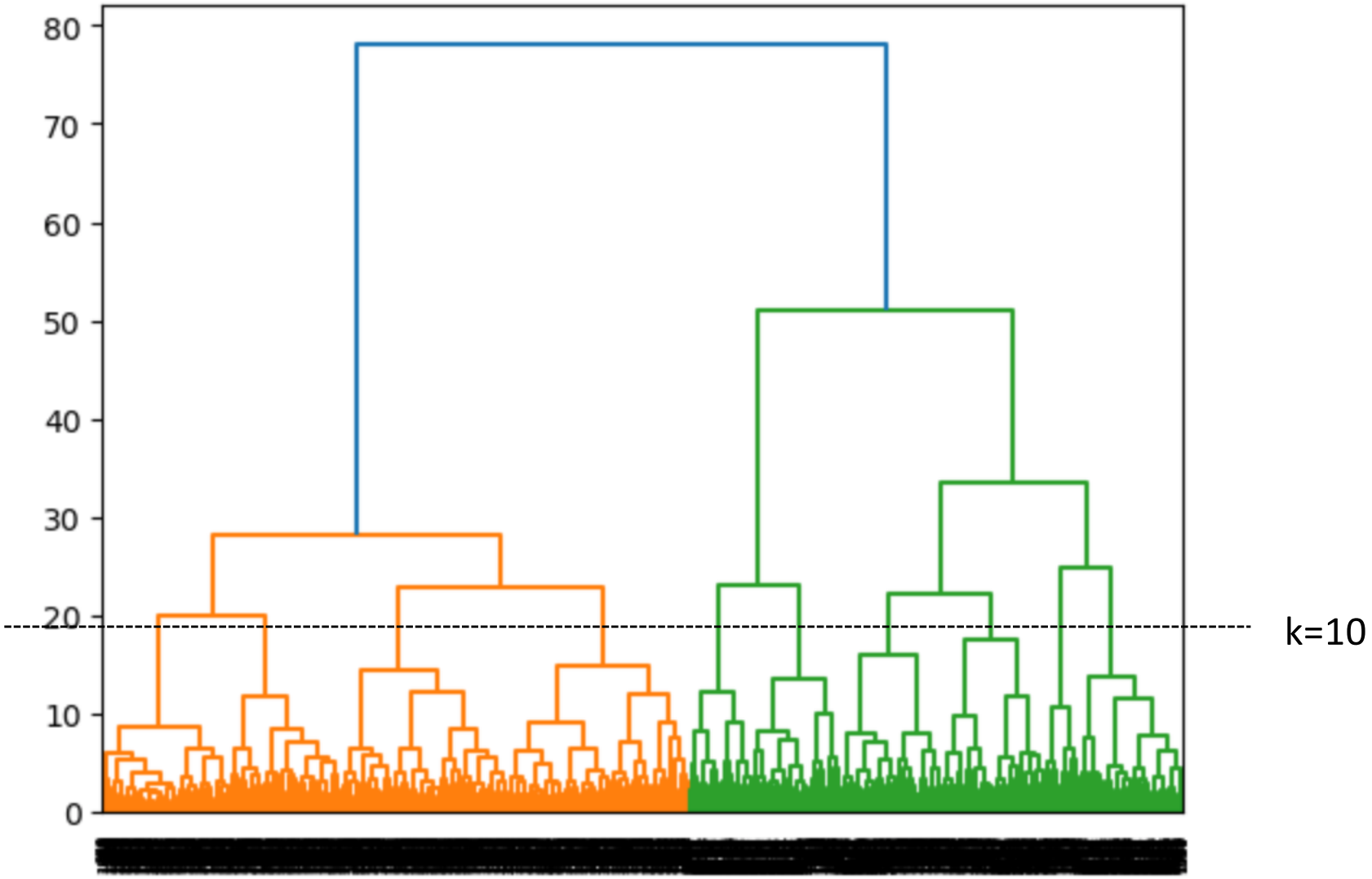
At the molecular DNA level:
exploiting the interpretability nature of MNNs

Filter interpretation

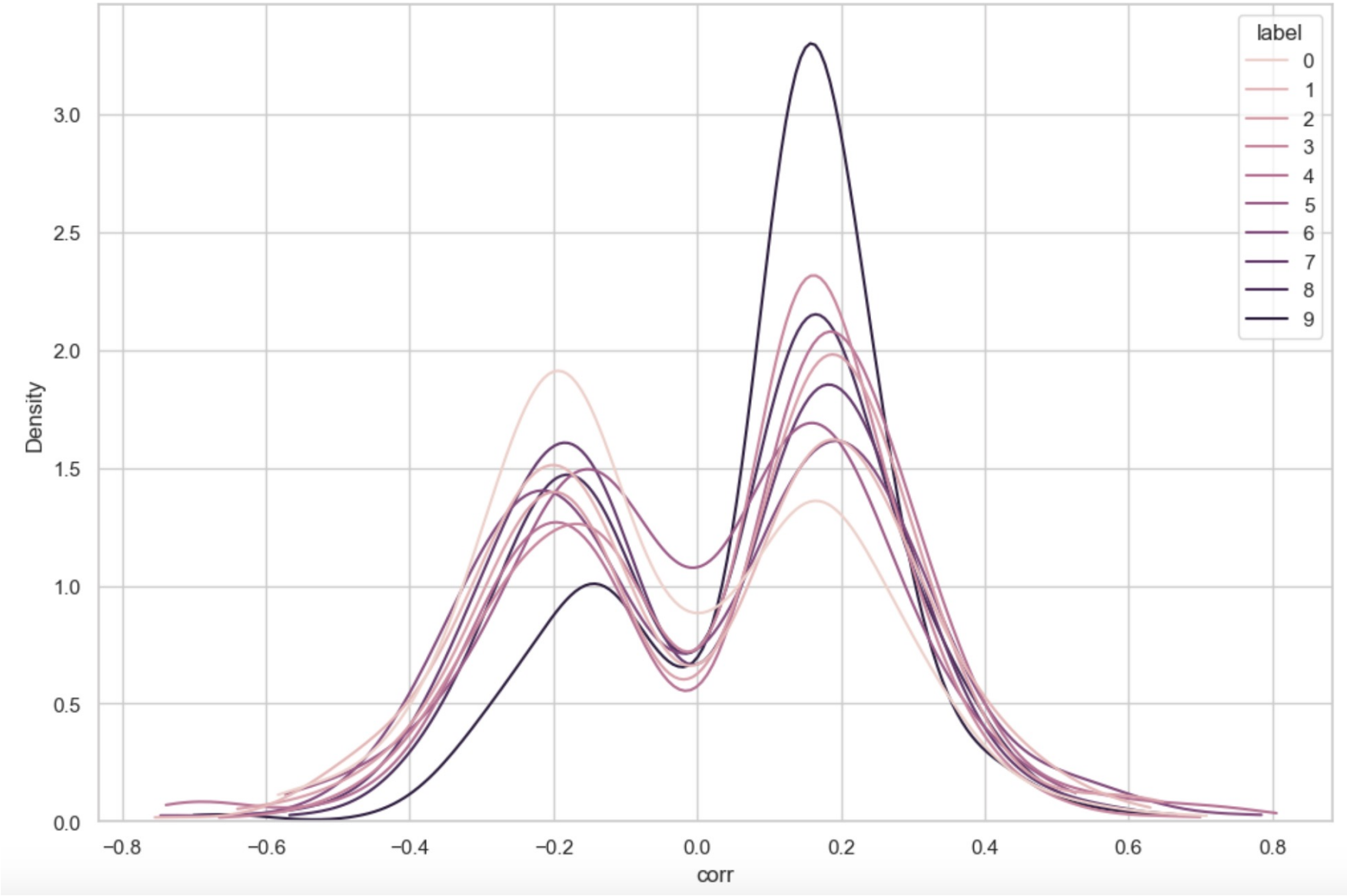
Consider $(T)_n$ model (9 modules/filters – combined through a linear regression)

Each $(T)_n$ eSTR is associated with a vector of 9 values corresponding to the regression coefficients

Hierarchical clustering



Correlation between MNN scores and expression of the gene target



For each cluster, find relevant motifs/filters and compare w/ existing motifs (JASPAR)

Experimental validation (S. Spicuglia, Marseille)

Take-home messages

. ML is able to model gene expression and various genomic regulations

. DNA sequence *per se* is key and a lot of information/instructions remain to be discovered – we have not finished reading the book…

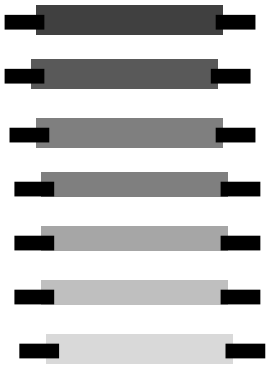
. Need fully interpretable models in order to provide hypotheses than can be experimentally validated and to generate novel biological knowledge

. Need of specific architecture and/or specific benchmarks (i.e. not MNIST*) not necessarily derived from image or natural language processing

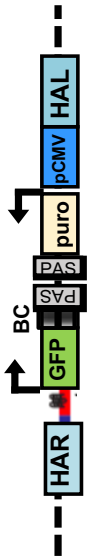
* We know what's an '8', we don't know the DNA grammar yet
(synthetic sequences and MPRA - ?)

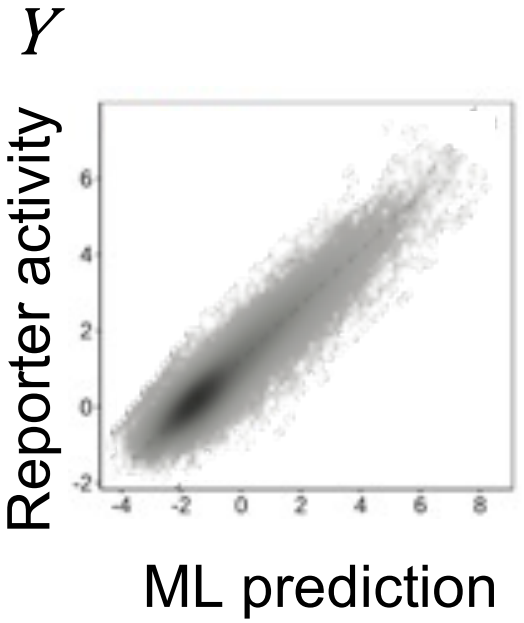
genomic-specific benchmark for ML

Design synthetic library



X





S. Spicuglia

$$Y \sim X$$

Y: reporter gene expression ; binding of a TF ; …

Acknowledgments

Computational Regulatory Genomics

Laurent Bréhélin	CR CNRS, LIRMM
Quentin Bouvier	PhD student, IGMM
Mathys Grapotte	PhD student, IGMM
Sophie Lèbre	MCF, IMAG & LIRMM
Charles Lecellier	DR CNRS, IGMM & LIRMM
Mathilde Robin	Engineer, LIRMM & IRCM
Océane Cassan	Post-doc, LIRMM

We are hiring! post-doc (ANR 3 years)

Alumni

Chloé Bessière	*PhD student, IGMM*
Lisa Calero	*grad. Student, IGMM*
Manu Saraswat	*grad. Student, IGMM*
Christophe Menichelli	*PhD student, LIRMM*
Mateo Meynier	*grad. Student, IGMM*
Yulia Rodina	*Post-doc, LIRMM*
Raphael Romero	*PhD student, IMAG & LIRMM*
Elodie Simphor	*grad. Student, IGMM*

Collaborators

Wyeth W. Wasserman	Vancouver, Canada
FANTOM consortium, Piero Carninci	Yokohama, Japan
Garrido-Martin Diego and Roderic Guigo	Barcelona, Spain
Sanofi - Bioinformatics Unit	Boston, USA
Salvatore Spicuglia	Marseille, France
Jose-Juan Lopez-Rubio	Montpellier, France

One possible solution: TF-MoDISco

Shrikumar et al. Arxiv, oct. 2020

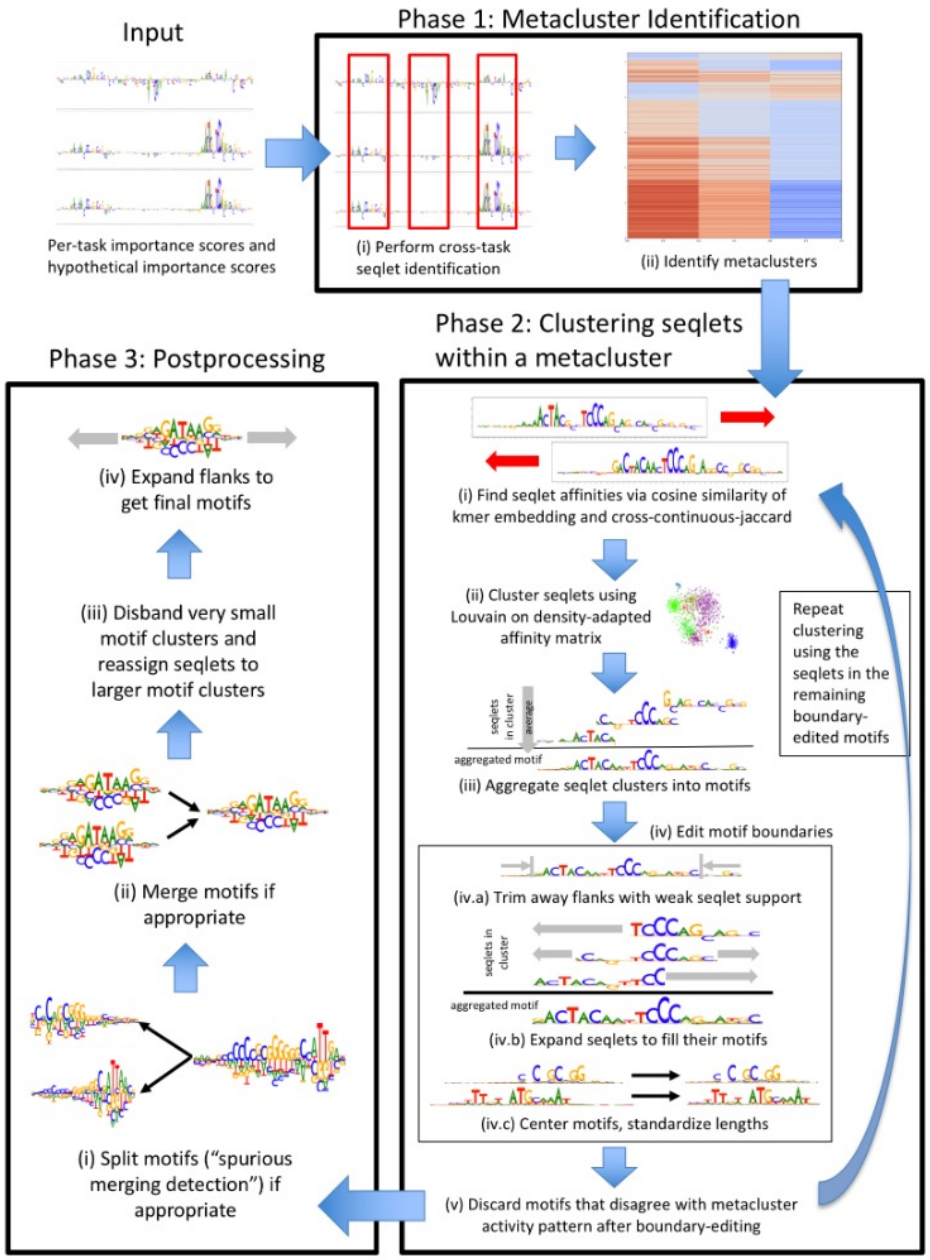
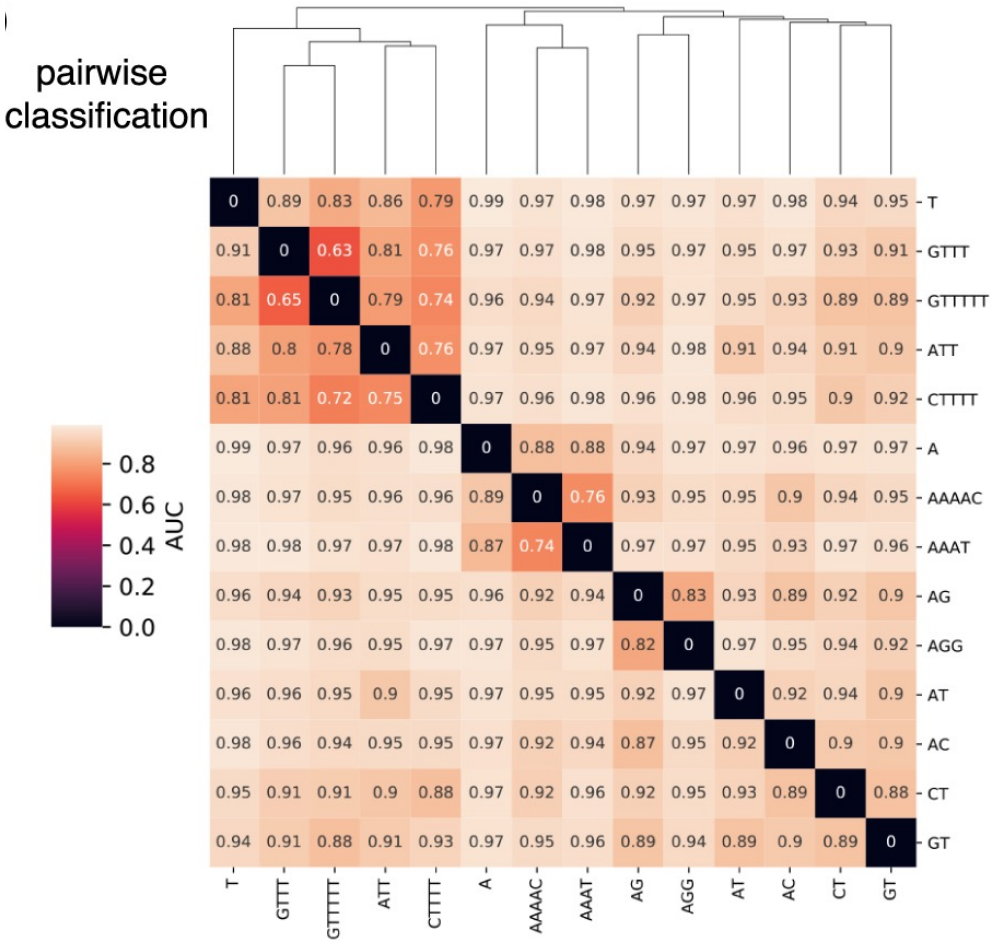


Figure 1: **Summary of TF-MoDISco**

In TF-MoDISco (which first uses DeepLift), a **clustering method** on a large dataset used as the input is used.

While being more robust, TF-MoDISco still uses only a subset of the sequence used for learning

CNN-based pairwise classification of STRs using only STR flanking sequences




```
>Human_STR_873642;T;+
CAGTCAAAGGTAATCATGGGTCCTTACCTGGGGTCTAGTAGAGCTTACAG-NNNNNNNNNN-NNNNNNNaaaatatctgtacatgtttcatgtacacaatttagtgcatatg
```

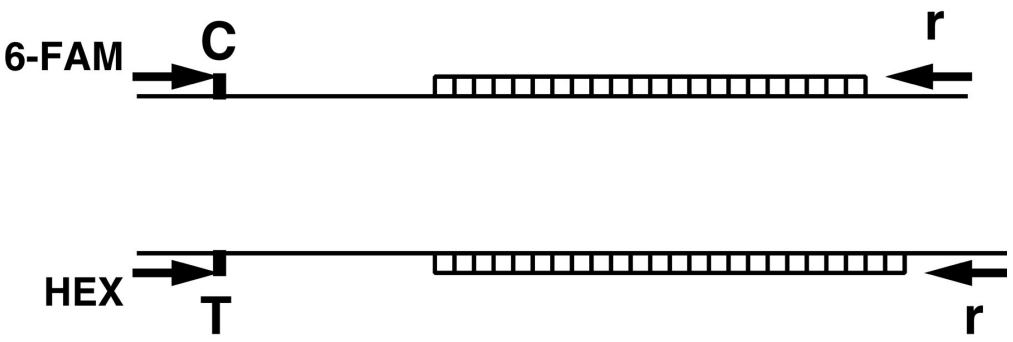
Functional link b/w STR and surrounding sequences

SNPSTR *(Mountain et al., Genome Res 2002 ; Agrafioti et al., Nucleic Acids Res. 2007)*

SNPSTR (Mountain et al., Genome Res 2002)

SNPs linked tightly to STR polymorphisms.

Such combinations satisfy the following three requirements:
(1) close physical linkage of two or more polymorphisms;
(2) significant difference in mutation rate between polymorphisms; and,
(3) potential for a large number of independent compound haplotypic systems.

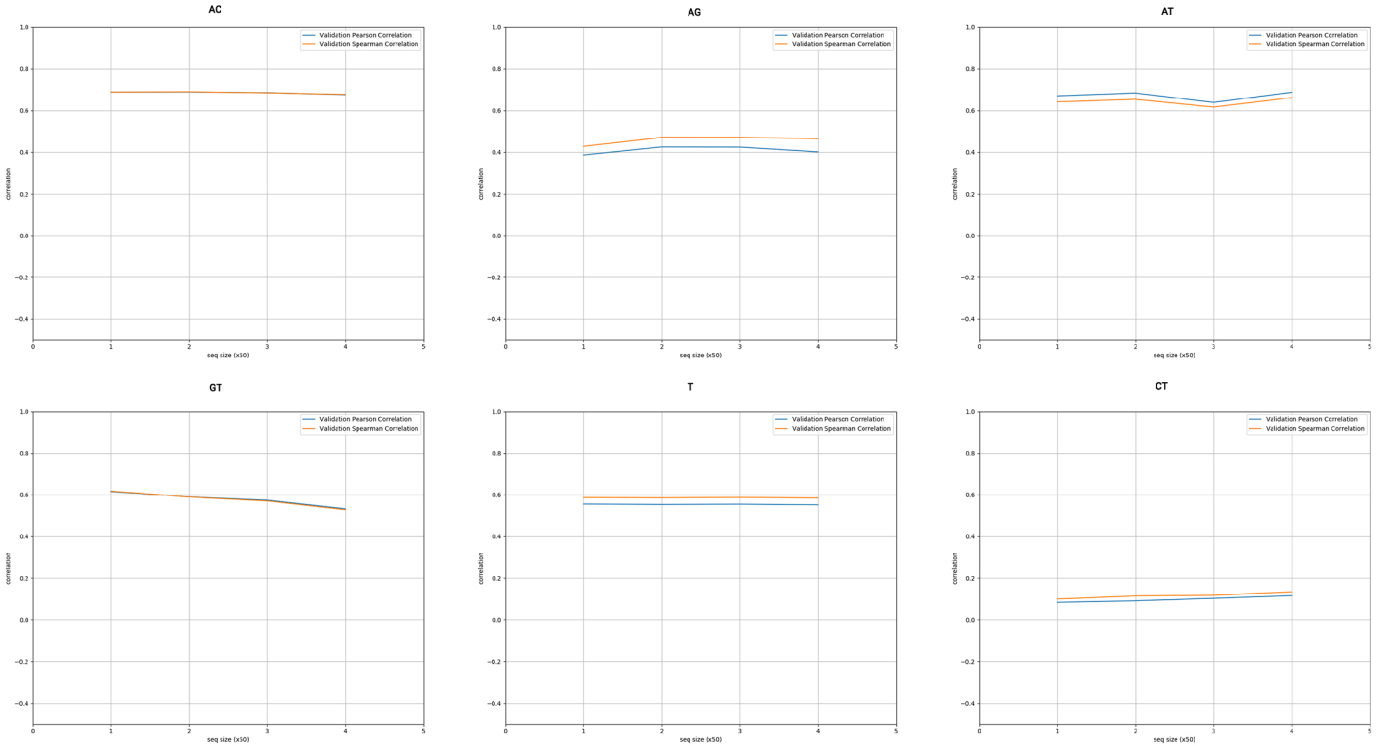


SNP mutation rate ~ 2.0–2.5 × 10^{-8} mutations per nucleotide position per generation
STR mutation rate ~1.5 × 10^{-3} per STR per generation

'This combination of co-inherited markers evolving at different rates may offer the possibility of gaining better resolved insights into population genetic processes compared to when these different marker types are used separately'
(Agrafioti et al., Nucleic Acids Res. 2007)

SNPSTR DB [Agrafioti et al., Nucleic Acids Res. 2007]
arbitrarily consider SNPs located 250bp apart of STRs (so each SNPSTR can be considered a small haplotype with no recombination occurring between the two individual markers)

Remember that, to predict transcription initiation, 50bp around STR 3'end (as defined by the tx strand) are sufficient - Grapotte et al., Nat. Comm. 2021



The complete sequence of a human genome

T2T consortium, Science 376, 44-53 1 April 2022

	GRCH38	T2T-CHM13	difference
Percentage of repeats (%)	51.89	53.94	
Repeat bases (Mbp)	1,516.37	1,647.81	+8.7
Long interspersed nuclear elements	626.33	631.64	+0.8
Short interspersed nuclear elements	386.48	390.27	+1.0
Long terminal repeats	267.52	269.91	+0.9
Satellite	76.51	150.42	+96.6
DNA	108.53	109.35	+0.8
Simple repeat	36.5	77.69	+112.9
Low complexity	6.16	6.44	+4.6
Retroposon	4.51	4.65	+3.3
rRNA	0.21	1.71	+730.4

https://www.ncbi.nlm.nih.gov/assembly/GCA_009914755.4