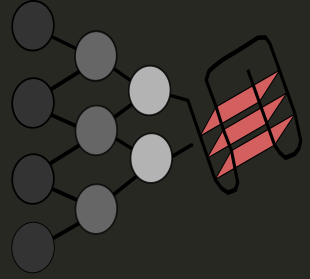


DeepG4

E. Nassereddine

DeepG4: A deep learning approach to predict active G-quadruplexes

Séminaire MIAT

Vincent ROCHER, Matthieu Genais, Elissar Nassereddine and Raphaël Mourad

CBI-Toulouse | Chromatin and DNA Repair | 19/03/2021

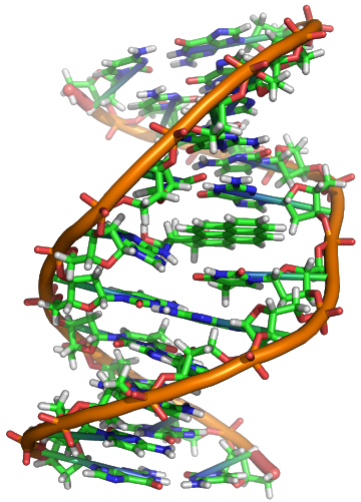
CBI Toulouse
Centre de Biologie Intégrative

UNIVERSITÉ TOULOUSE III PAUL SABATIER

BSB

DNA: The secret of life

B-DNA *(1953)*

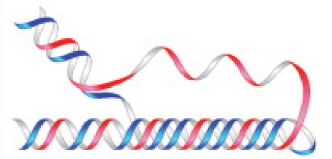


> The B- DNA (double helix structure) is the most stable structure.

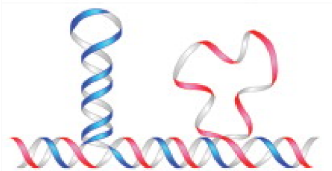
Non B-DNA *(1954)*

Cruciform

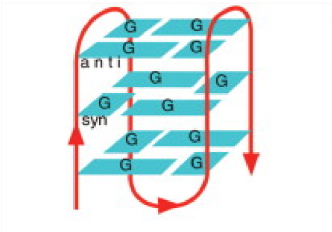
Triplex



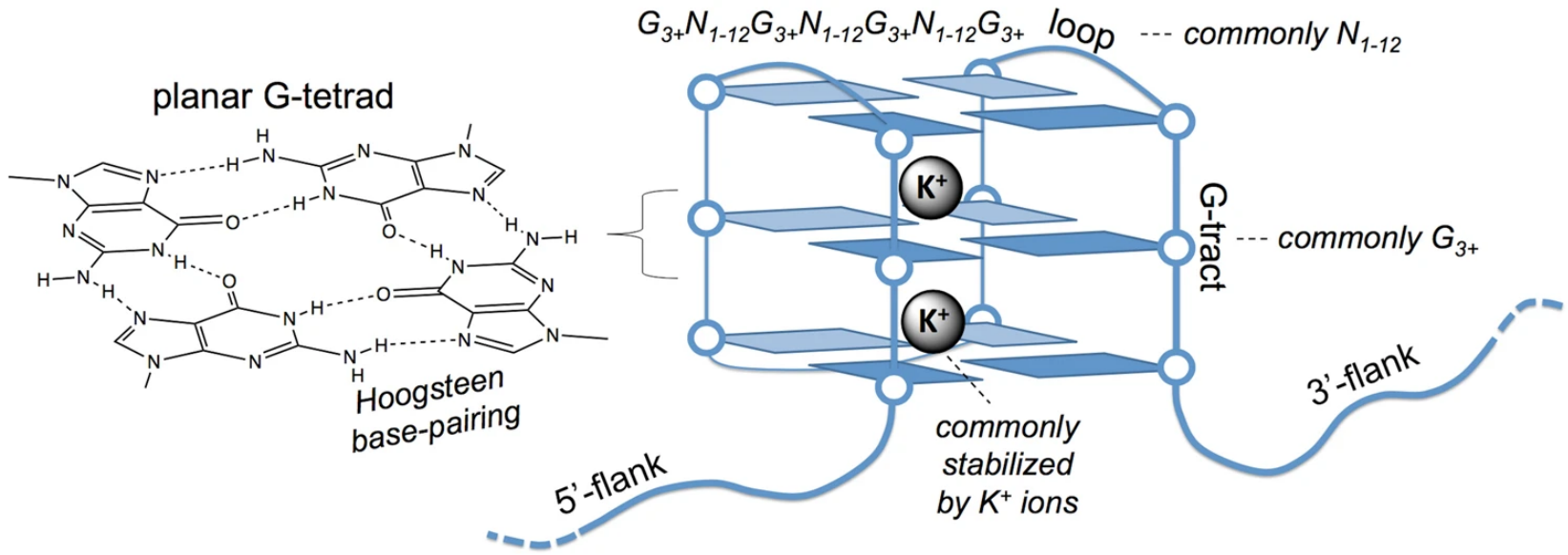
Slipped Structure



Tetraplex
(Quadruplex)



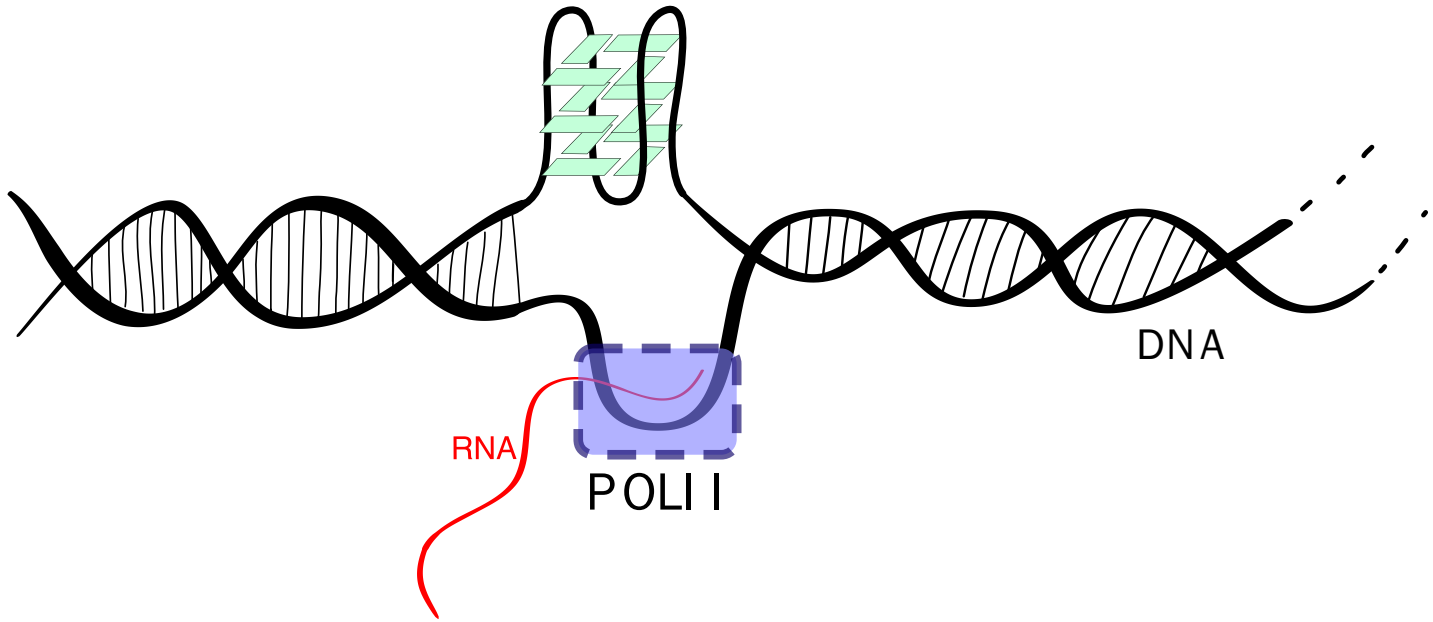
G-Quadruplex (G4): A non B-DNA Structure



Balasubramanian et al,Sci Rep, 2017

- Fold into four-stranded structures.
- Containing guanine tetrad.
- Motif $G \geq 3N_xG \geq 3N_yG \geq 3N_zG \geq 3$

Biological function of G4's



- Regulation of **gene expression** and **chromatin architecture**.
- **Telomere stability**.
- Disrupting the replication fork progression causing **Double-strand breaks (DSBs)**.

Algorithms for G-quadruplexes (G4) predictions

Expert system methods

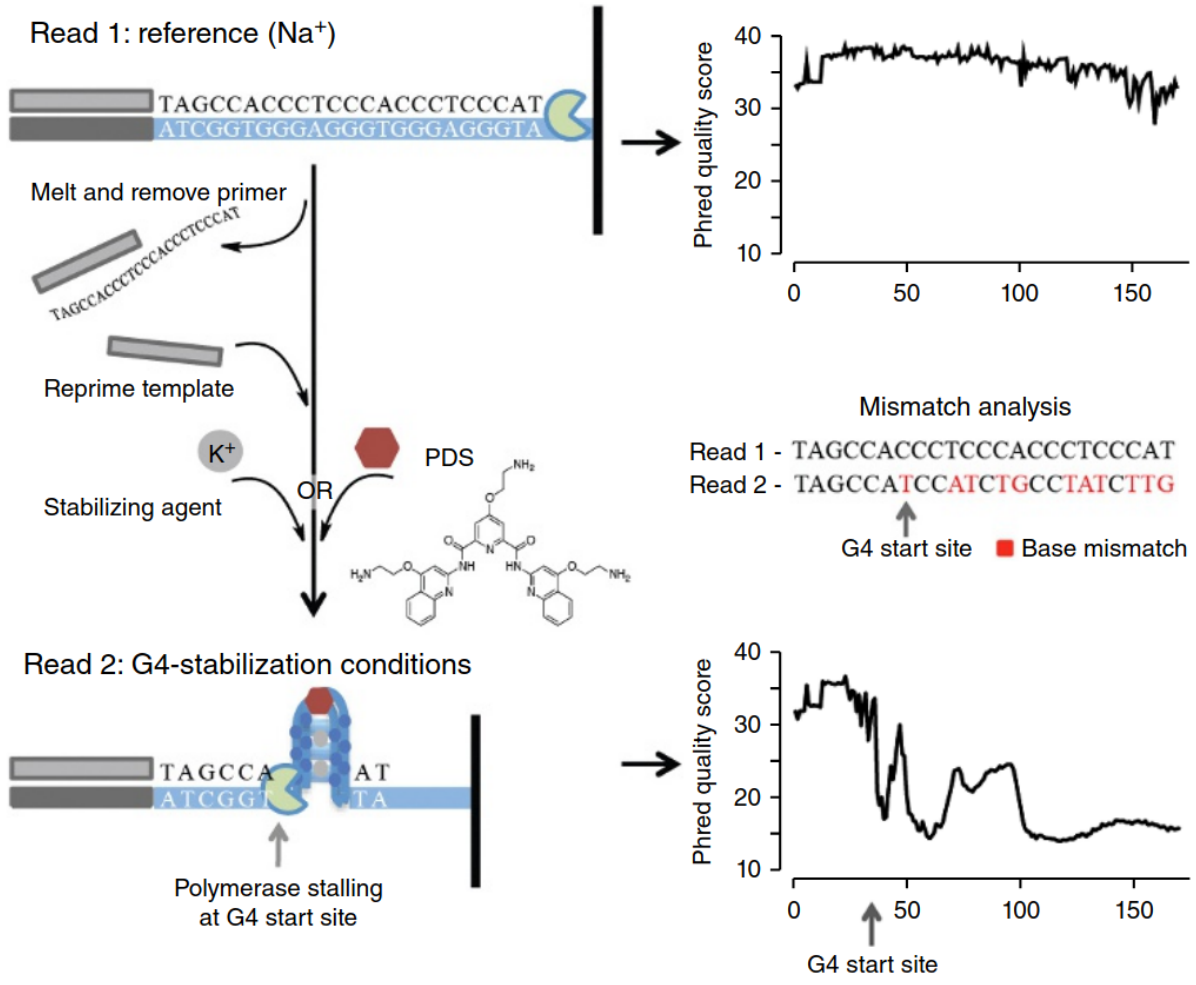
Name	Method	Implementation	Year	Link
quadparser	Regex	**Python**	2005	https://github.com/dariober/
gqrs_mapper	Score based	**Python**	2006	http://bioinformatics.ramapo.edu/QGRS
G4hunter	Score based	**Python**	2016	https://github.com/AnimaTardeb/G4Hunter
pqsfinder	Score based	**R**	2017	https://bioconductor.org/packages/release/bioc/html/pqsfinder.html
qparse	Score based	**Python**	2019	https://github.com/B3rse/qparse
G4CatchAll	Regex	**Python**	2019	https://github.com/odoluca/G4Catchall

- **Regex:** `([Gg]{3,}) (\w{1,8}) ([Gg]{3,}) (\w{1,8}) ([Gg]{3,}) (\w{1,8}) ([Gg]{3,})`
- **Score based:** Compute a score using a sliding windows over the whole genome by using **G richness** and **G skewness** (G4Hunter).

First G4 genome-wide mapping *in vitro* (G4-seq) 2014

High-throughput sequencing of DNA G-quadruplex structures in the human genome

- High-resolution sequencing–based method to detect G4s in the human genome in vitro.
- The developed method called G4-seq combining features of the polymerase stop assay with Illumina next-generation sequencing.



Algorithms for G-quadruplexes (G4) predictions

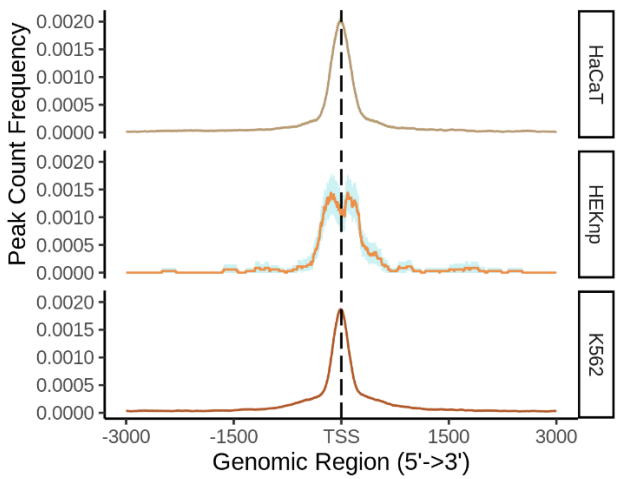
Machine learning based algorithms

Name	Method	Implementation	Year	Link
quadron	Machine Learning	**R xgboost**	2017	https://github.com/aleksahak/Quadronr
G4detector	Deep Learning	**Python / Tensorflow**	2019	https://github.com/OrensteinLab/G4detector
penguinn	Deep Learning	**Python / Tensorflow**	2020	https://github.com/ML-Bioinfo-CEITEC/penguinn

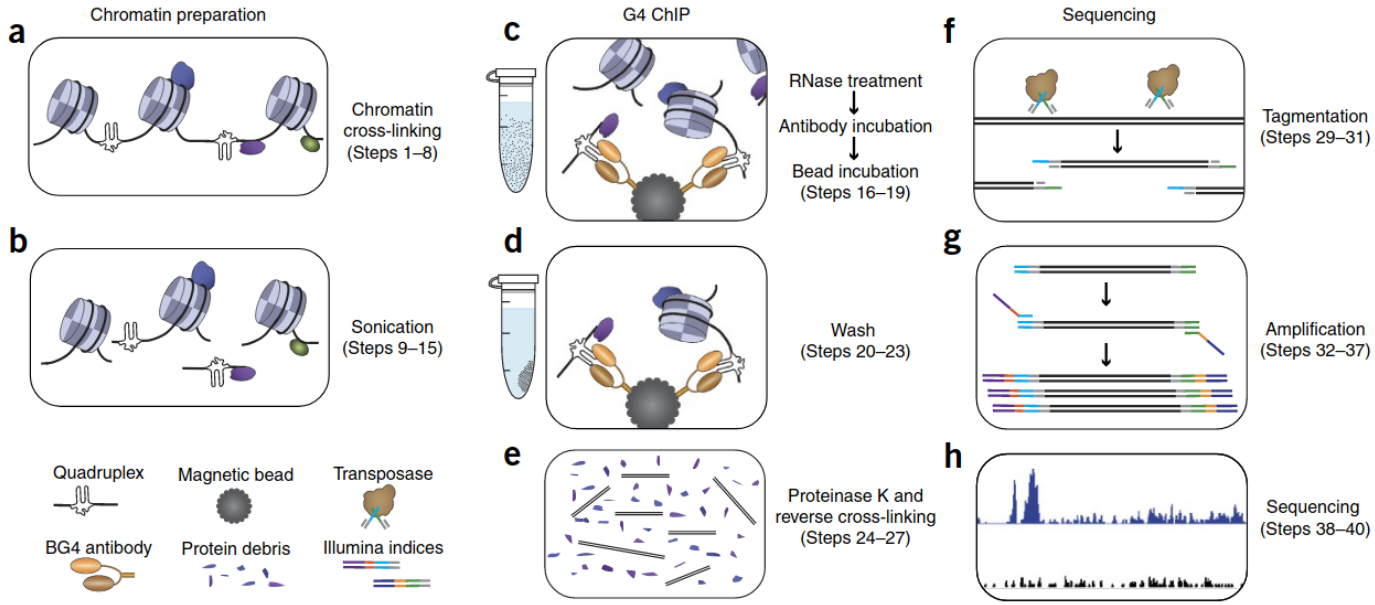
- **Quadron:** A machine learning model to predict the formation of G4s using 119 sequence-based features. I.e: the number of tetrads in the G4s , the occurrence of special kmer ...
- **Penguinn, G4detector:** Multiple layers CNN (Deep learning)

Mapping G4s in vivo with BG4-seq (2018)

- ChIP-seq for the DNA secondary structures through the use of a G4-structure-specific single-chain antibody (BG4).
- Refinements in chromatin immunoprecipitation.
- Followed by high-throughput sequencing.

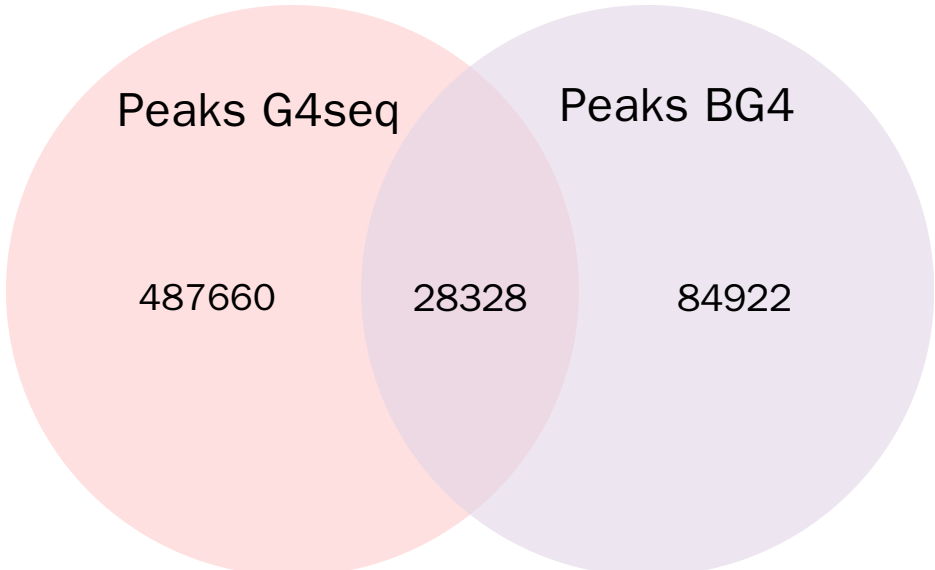


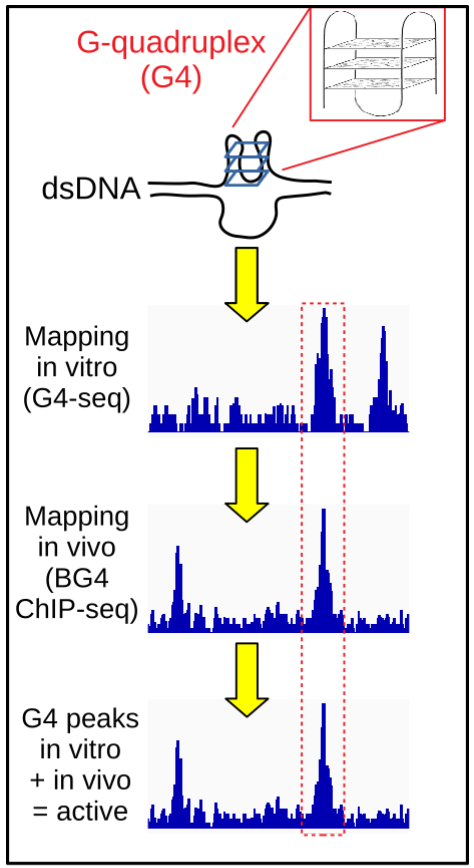
Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing



Active G4s dataset

Overlap between in vitro (G4-Seq) and in vivo (BG4-Seq) form active G4s.



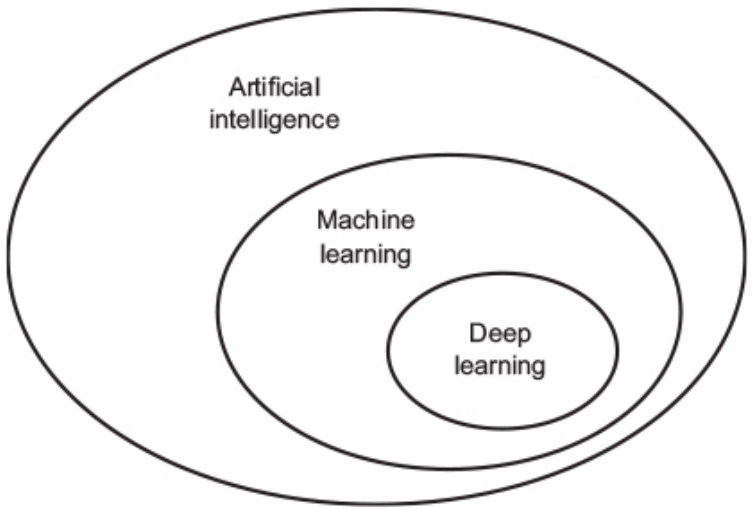


Mapping of
active G4s

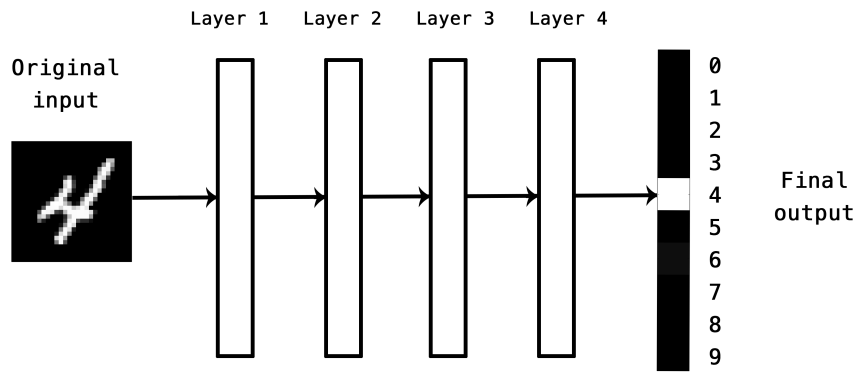
G4 predictions with DeepG4

DeepG4: a deep learning model to predict active G4s (BG4-G4-seq peaks).

What is deep learning ?

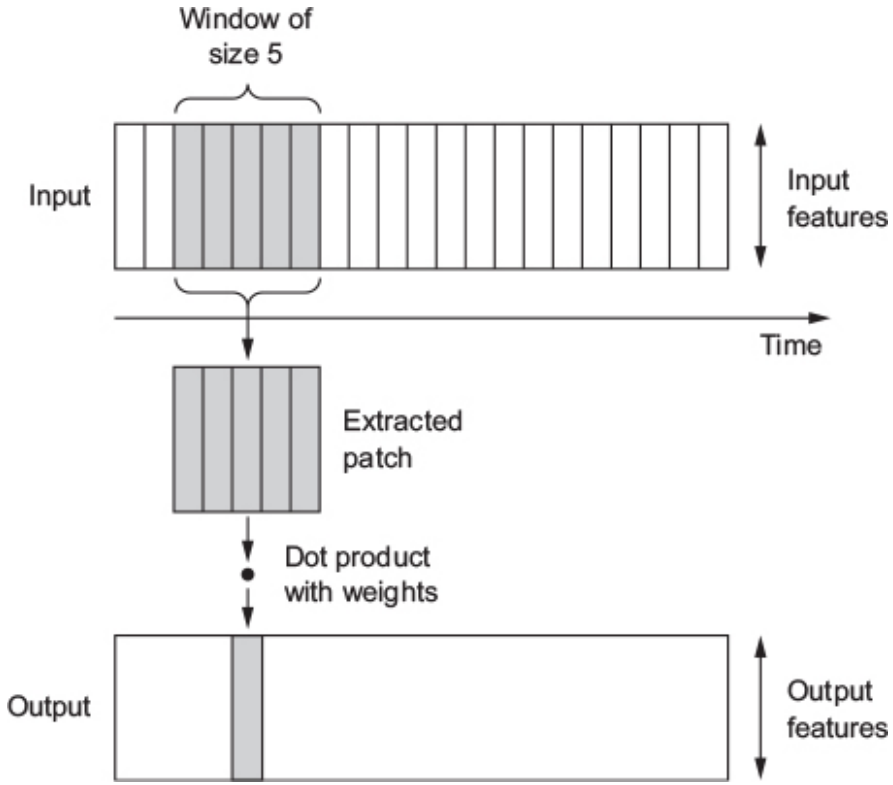


Some basic representation of a multi-layer deep learning model

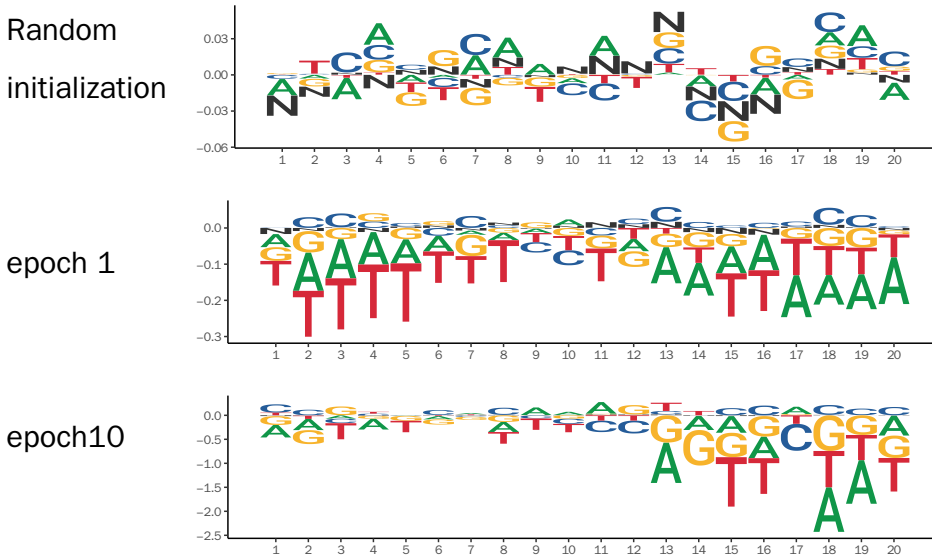


G4 predictions with DeepG4

Deep learning for DNA sequences.

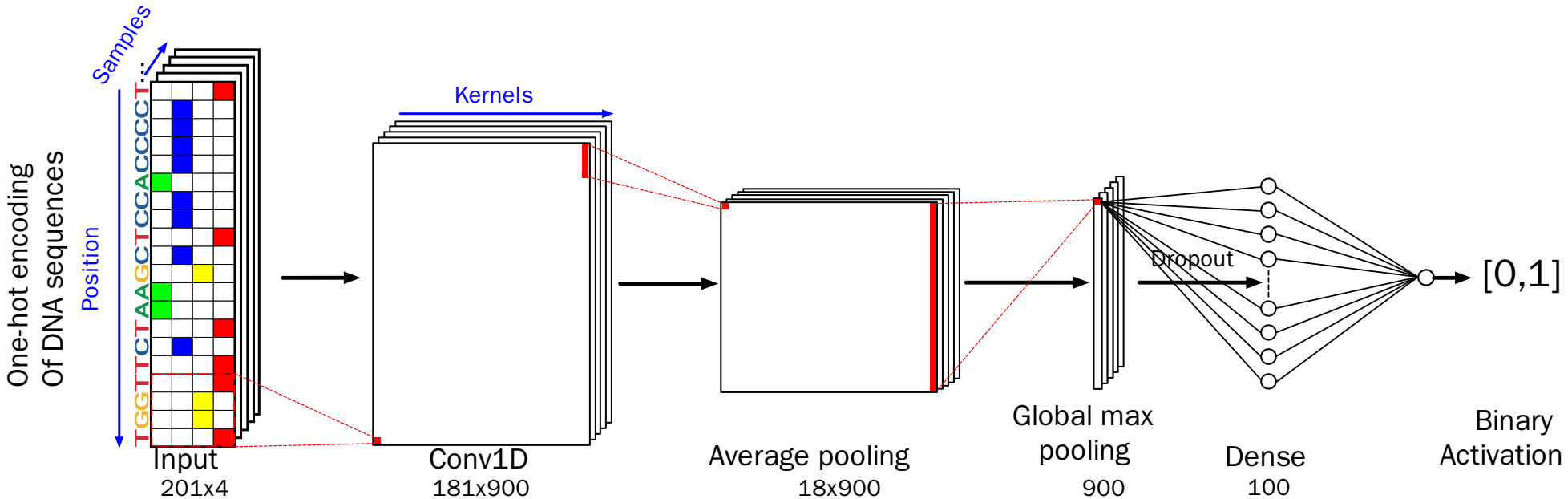


Convolutional model (CNN)



Weights can be represented as PWM and encode motifs as features for our model.

DeepG4 model architecture



1. **Conv1D**: Scan sequences using kernel (20bp).
2. **Average pooling**: Reduce dimension size and aggregate kernel signal.
3. **Global max pooling**: Output max activation signal for each kernel.
4. **Dropout**: Regularization layer.
5. **Dense layer** `(100 units,linear)`: Combination of weighted kernel signal.
6. **Dense layer** `(1 unit, sigmoid)`: Output a probability.

DeepG4: Performances

Input

Control sequences: randomly selected genomic sequences that matched sizes, GC, and repeat contents similar to actives G4s (R package gkmSVM).

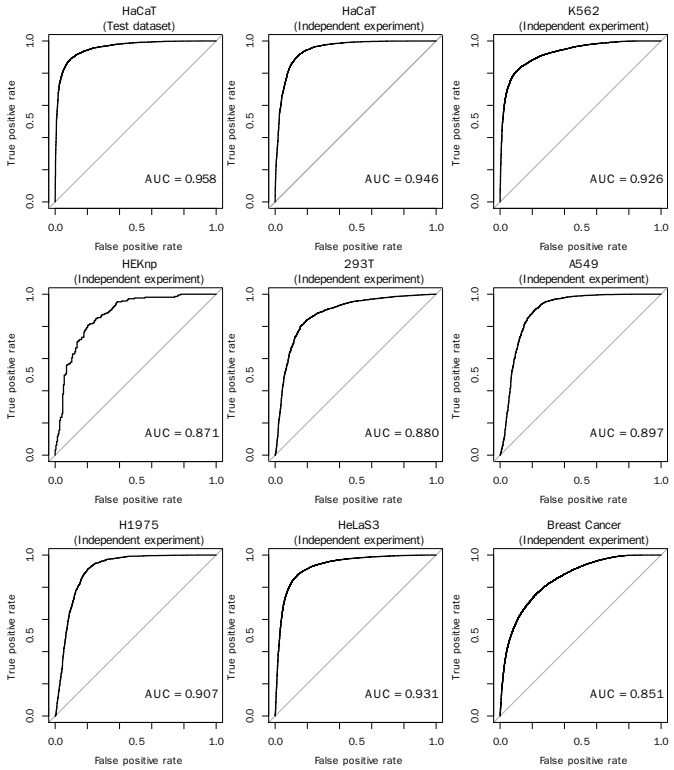
- **HaCat G4:** train/validation/test dataset.
- **Independent experiment:** HaCat, HEKnp, K562.
- **G4P experiments:** HeLaS3, 293T, A549, H1975 *(Ke-wei Zheng et al, NAR 2020)*.
- **qG4 experiment:** Breast cancer.

Tools

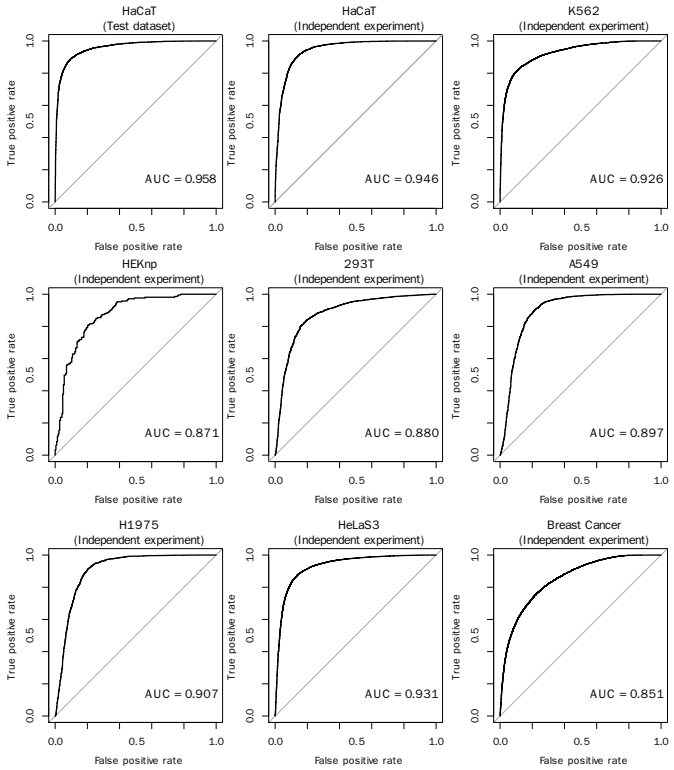
G4 detection algorithms

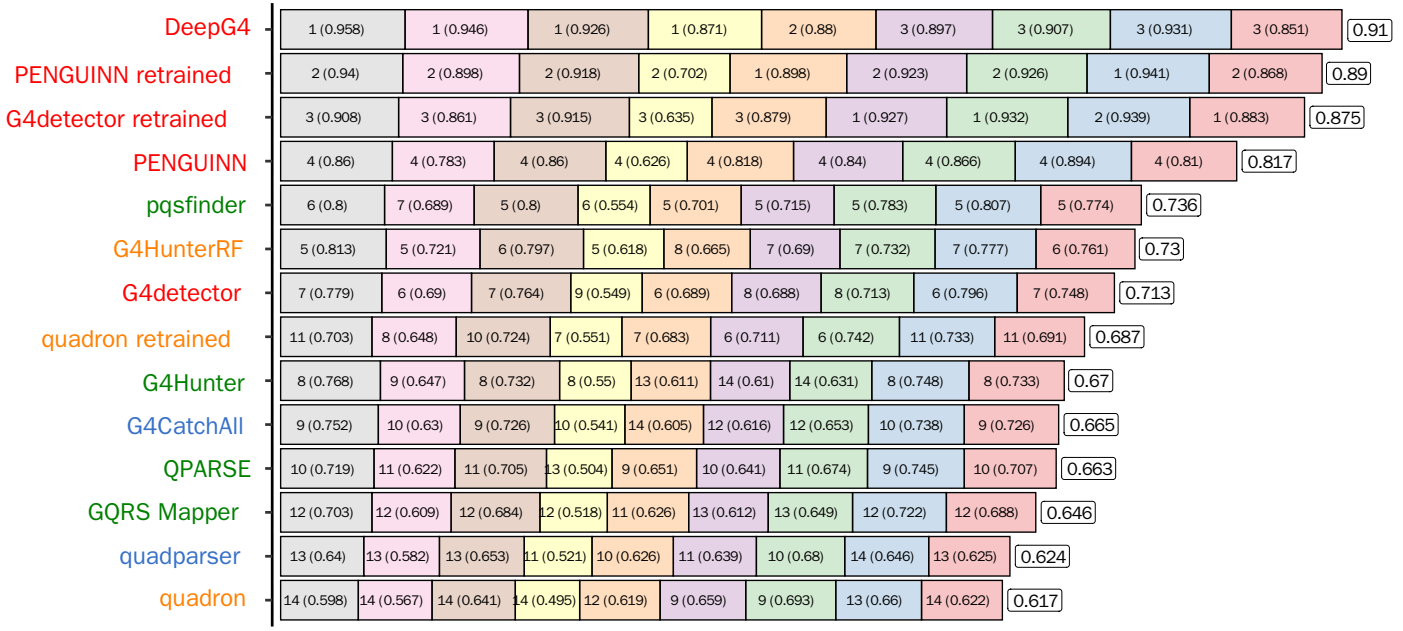
Name	Method	Implementation
DeepG4	Deep Learning	R/Tensorflow
penguinn_retrained	Deep Learning	Python / Tensorflow
penguinn	Deep Learning	Python / Tensorflow
G4detector_retrained	Deep Learning	Python / Tensorflow
G4detector	Deep Learning	Python / Tensorflow
quadron_retrained	Machine Learning	R xgboost
quadron_score	Machine Learning	R xgboost
G4hunterRF	Machine Learning	R ranger / python
G4hunter	Score based	Python
qparse	Score based	Python
pqsfinder	Score based	R
gqrs_mapper	Score based	Python
quadparser	Regex	Python
G4CatchAll	Regex	Python

DeepG4: Performances

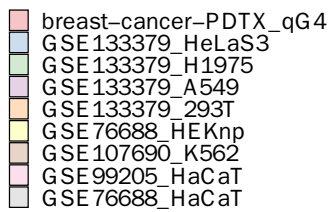


DeepG4: Performances



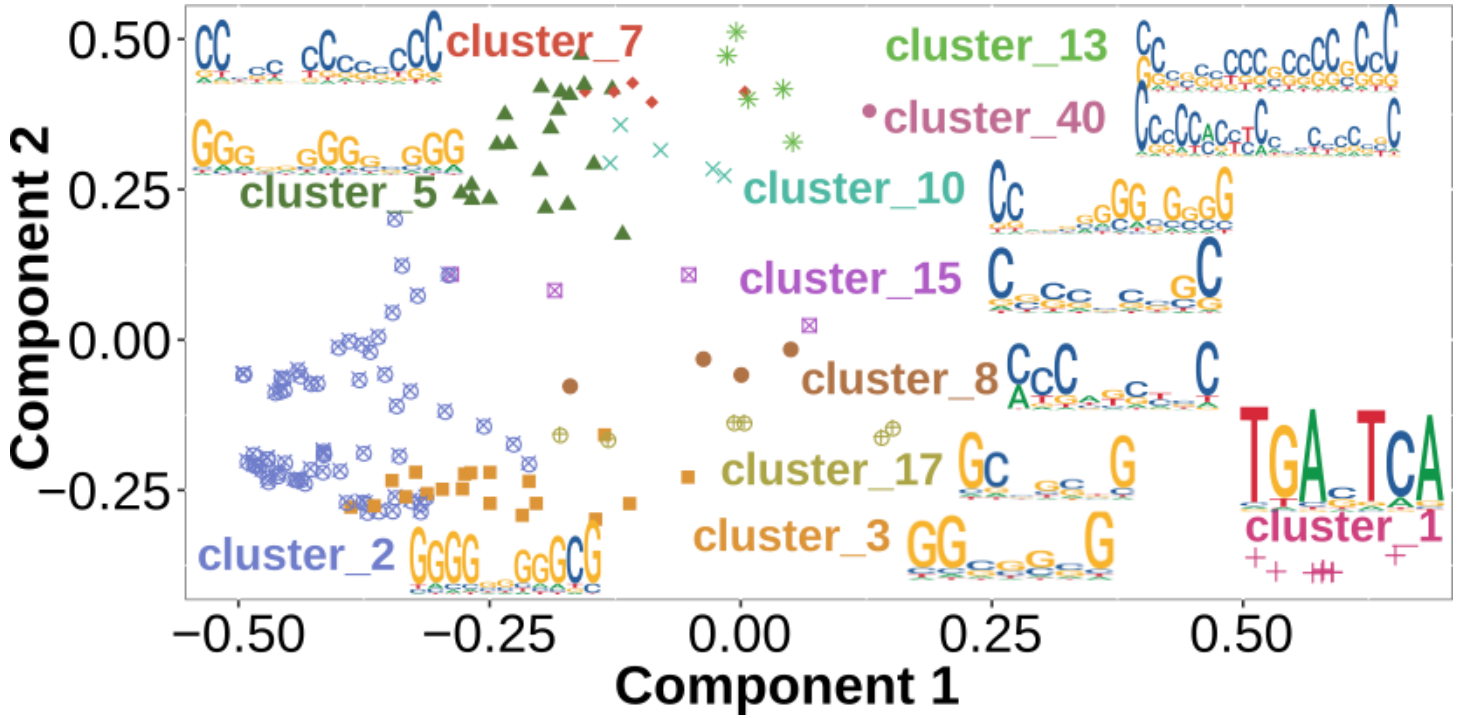


AUC (rankings) for 9 different datasets



DeepG4: Feature extraction

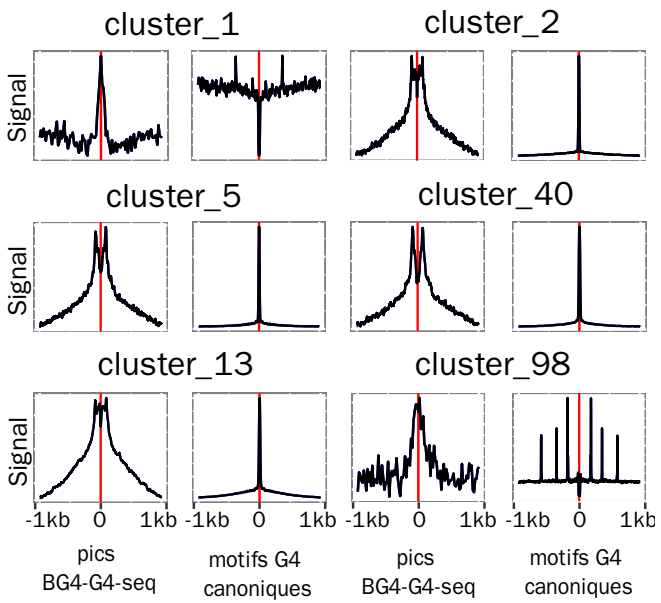
- **Motifs are extracted from kernels**.
- 900 kernels associated into **163 clusters** using matrix clustering (RSAT).
- Represented into **163 root motifs**.

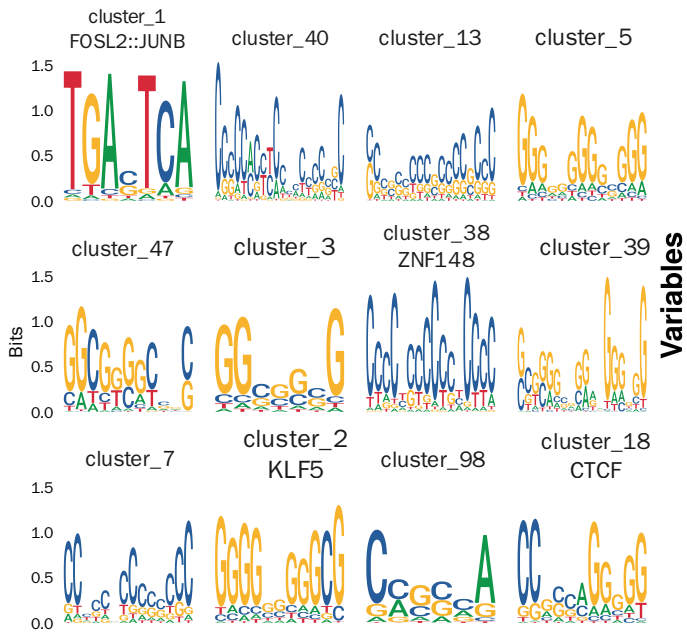


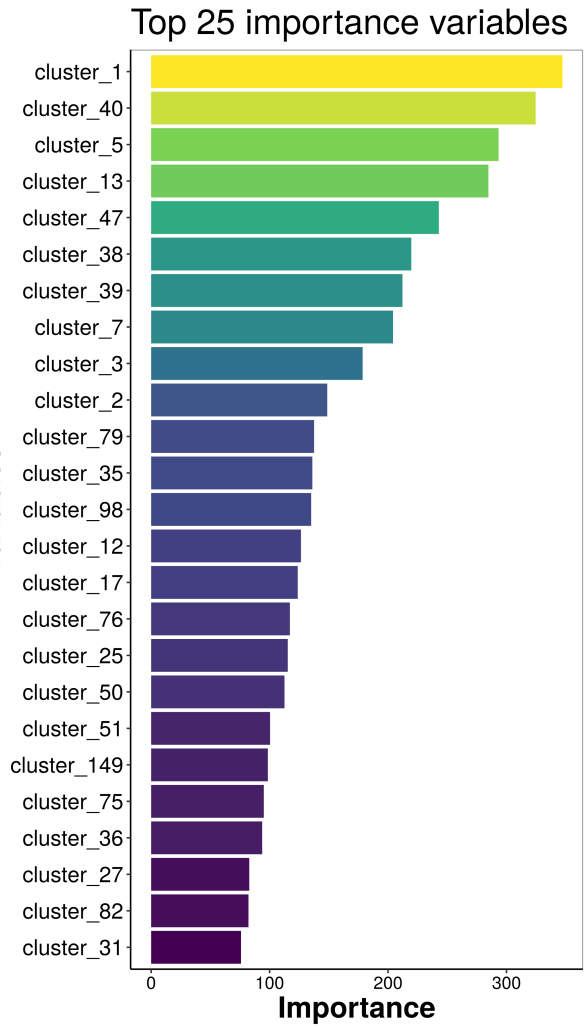
Multidimensional scaling (MDS) of DeepG4 clusters.

DeepG4: Feature importance

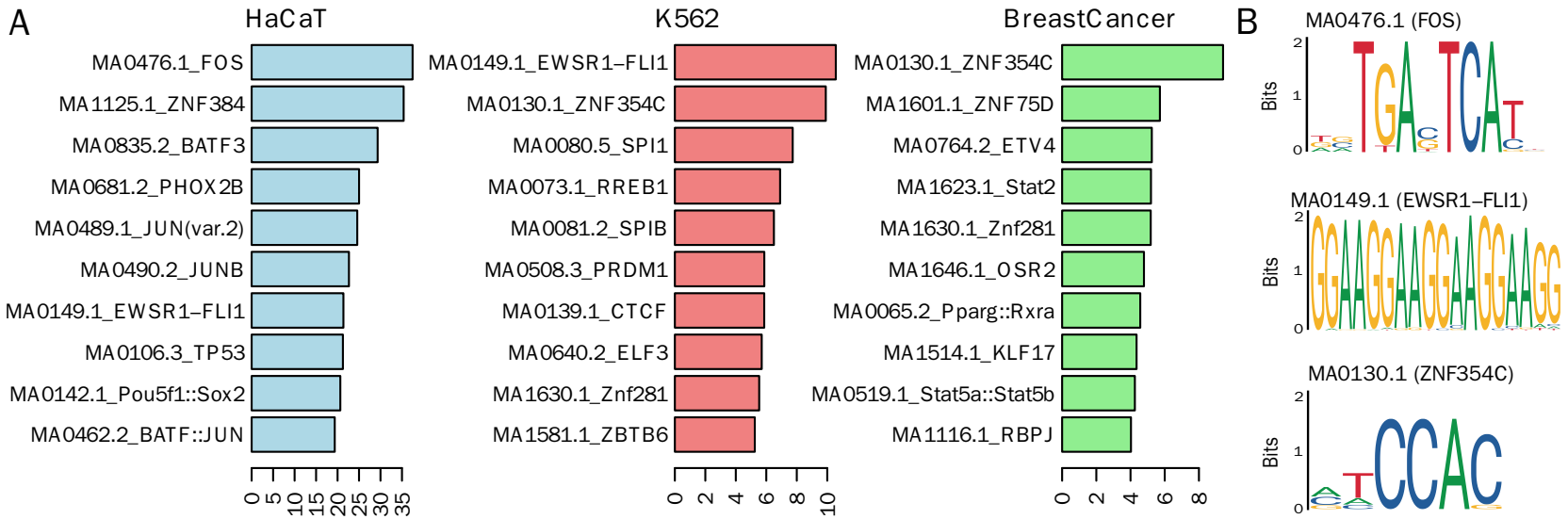
- Known TFBS motifs (identified with TomTom) are good predictors.
- De novo and G4-like motifs also found as good predictors.







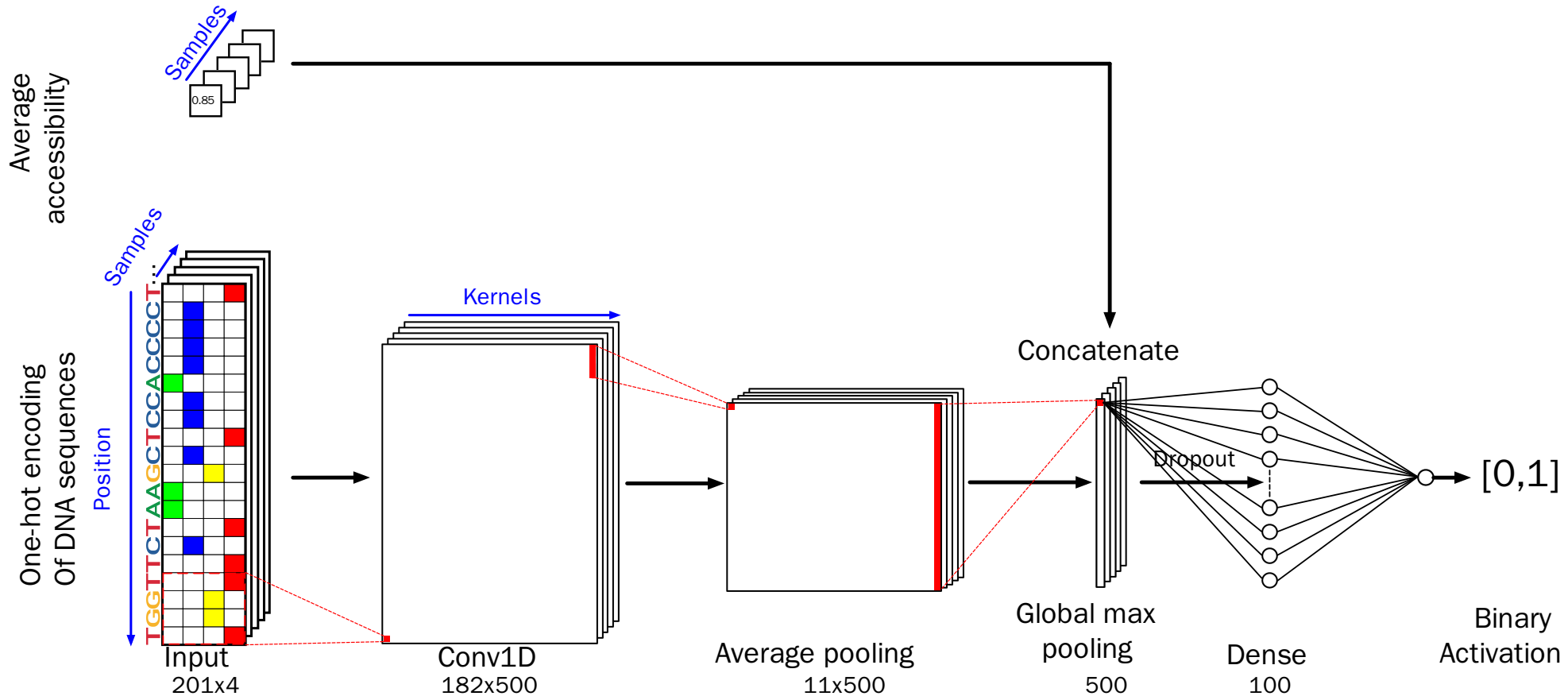
Cell type specific transcription factor motif predictors of active G4s



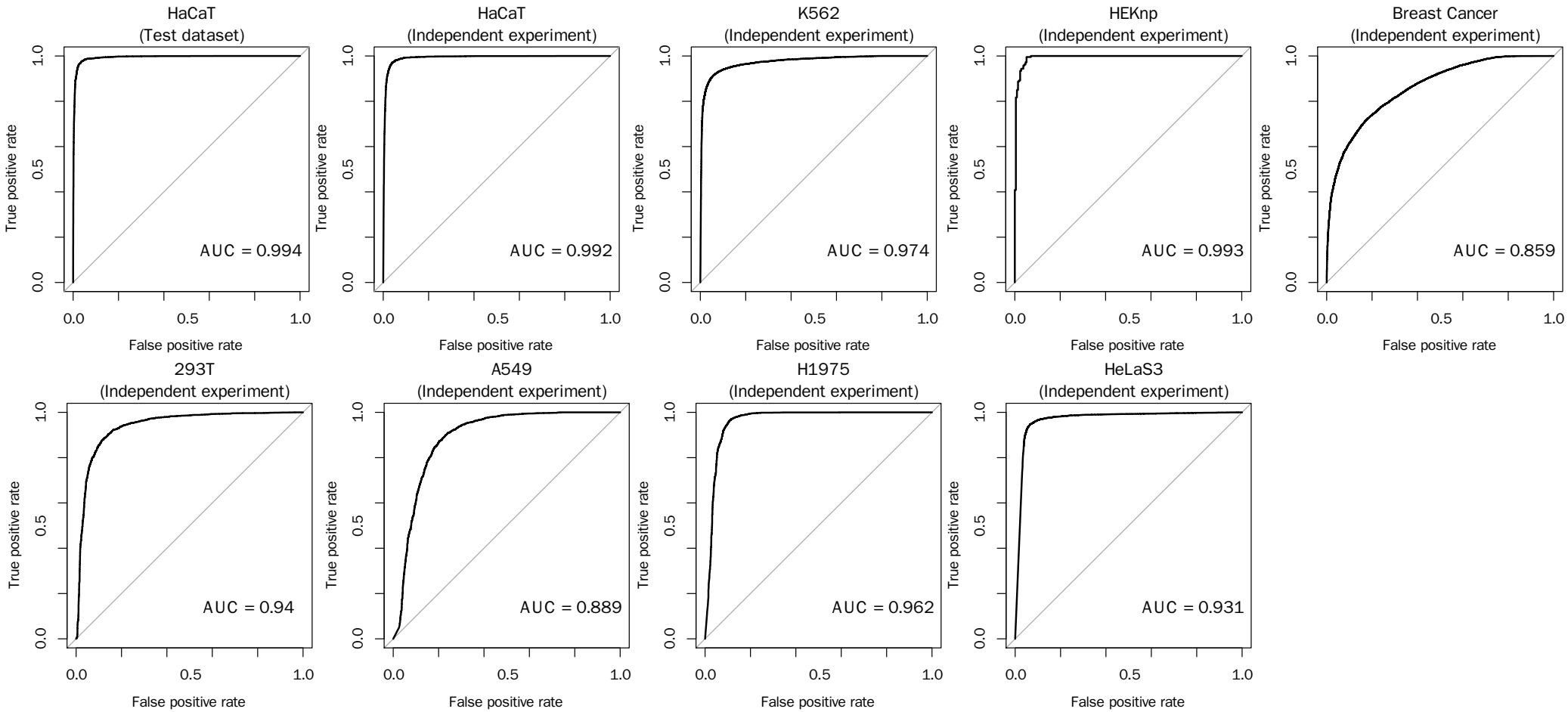
Random Forest classifier:

- One cell type vs all others cells types.
- Use TFBS motifs as features.
- Importance weighted by motif abundance in the positive set.

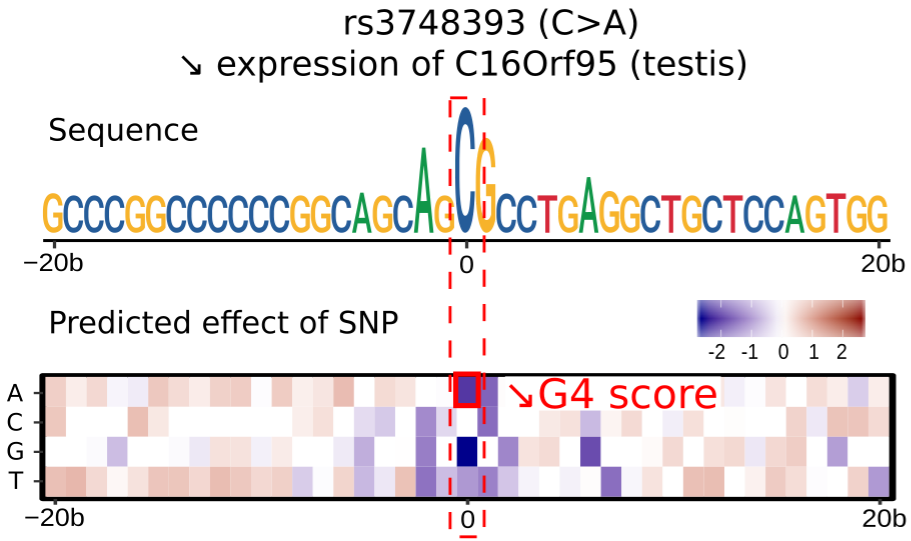
DeepG4 model architecture with accessibility

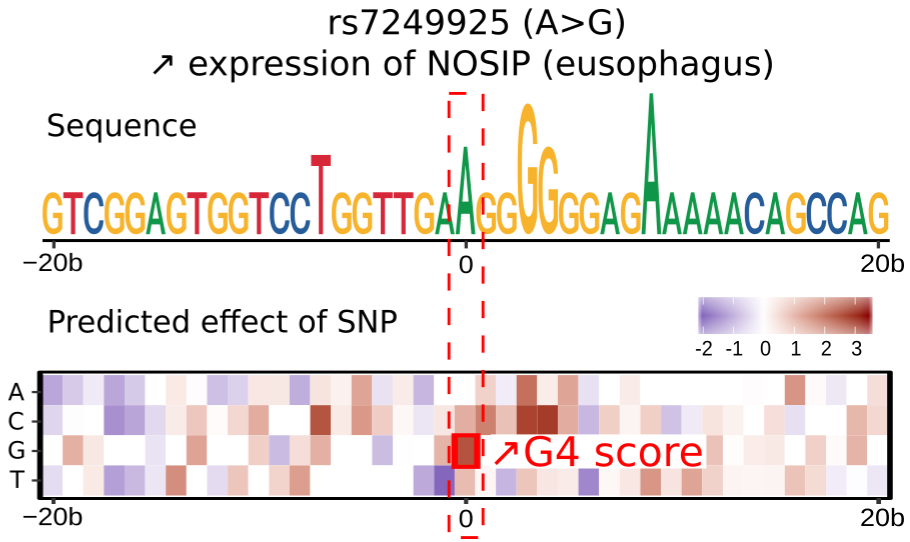


DeepG4: Performances with accessibility



DeepG4: SNP effect on active G4s

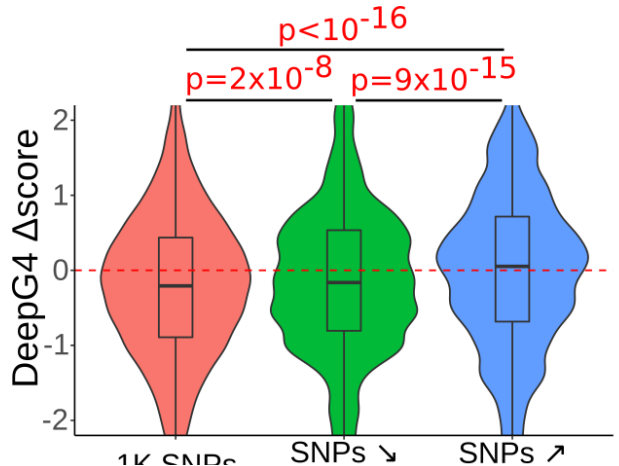




- C>A lead to a decrease in G4 activity.
- A>G lead to an increase in G4 activity.

SNPs could alter the G4 structure stability.

SNPs eQTL (GTEx) increasing gene expression presented high G4 activity.



Thanks !

Vincent ROCHER, Matthieu Genais, Elissar Nassereddine and Raphaël Mourad

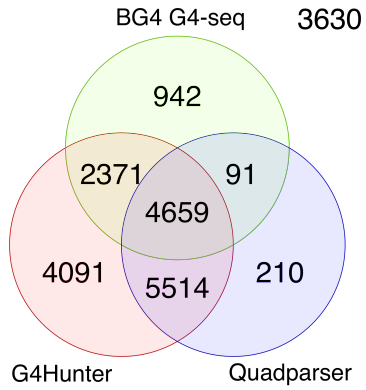
CBI-Toulouse | Chromatin and DNA Repair | 19/03/2021

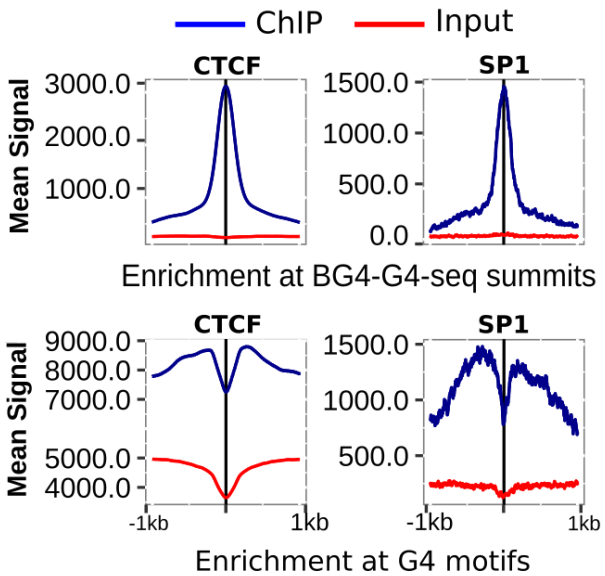
Possible upgrades

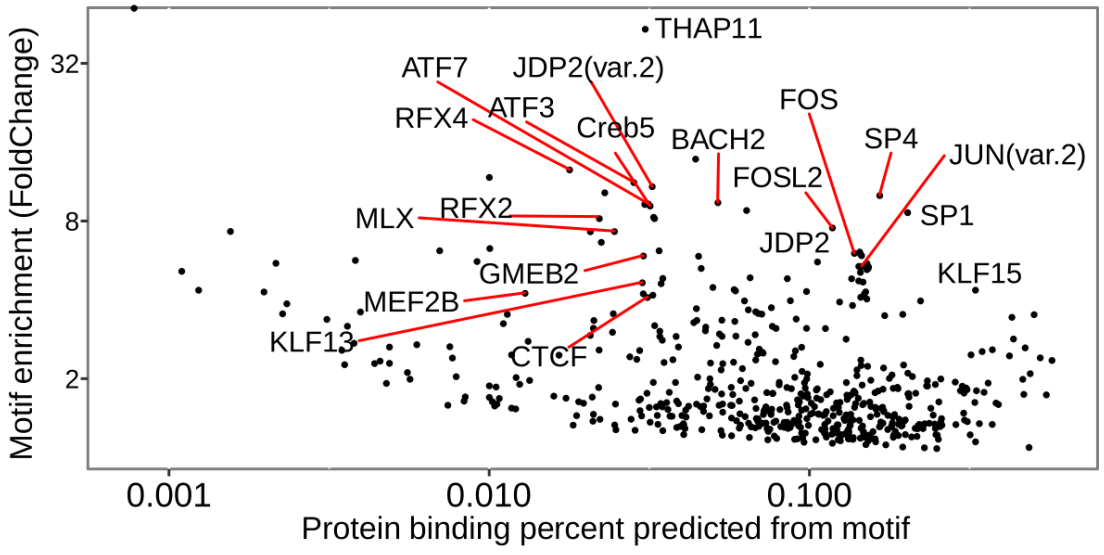
- Quasi-SVM as last layer (replacing Dense).
- Filled weights with JASPAR PWMs to help training.
- Parallel convolution layer with differents kernels sizes.
- Add DNA accessibility as input with ATAC-seq.

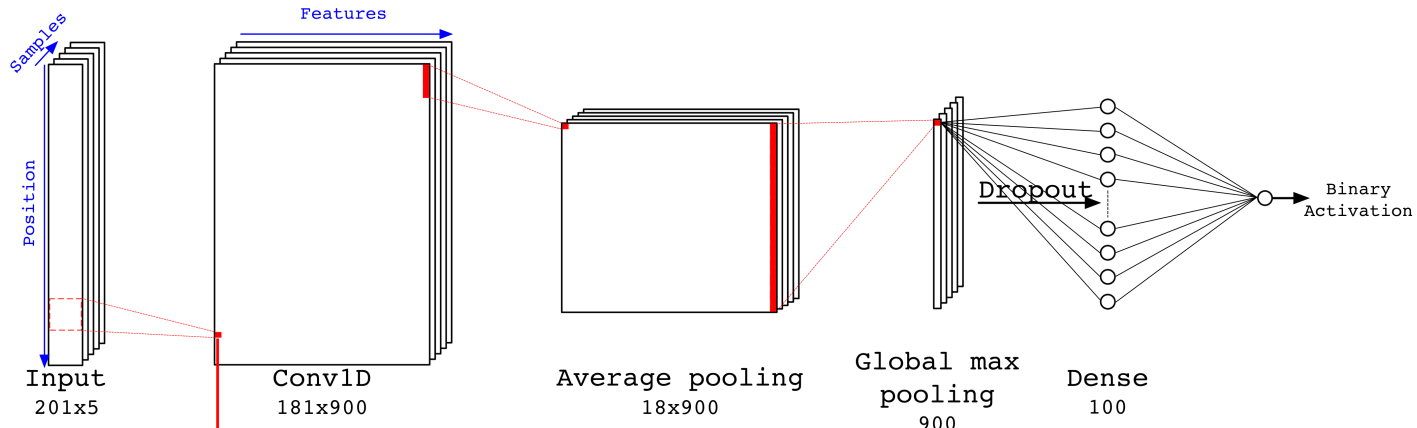
Sequence features enriched at active G4s

- Active G4s are enriched in **promoters**.
- Current algorithms failed to predict non-negligible fraction of **active G4s (11%)**.
- And more than 50% of their results are **false positives**.



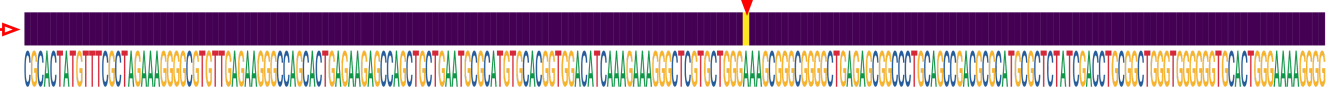






Extract
Feature map

Max activation
position



Motif extraction