

Méthode d'apprentissage profond pour l'analyse génomique des cancers canins comme modèles des cancers

Christophe HITTE

University of Rennes, CNRS, France



Genome Annotation

- ~**2% of DNA** is made up of protein-coding genes + **98%** is noncoding

noncoding DNA : Regulatory elements

- promoters
- enhancers
- silencers
- insulators

noncoding RNA :

- short lengths of RNAs (microARN, tRNA, rRNA)
- long noncoding RNAs (lncRNAs)

>**60%** of the genome is **transcribed into RNAs**

Genome Annotation

- **Incomplete genome annotation impact genomic analyses :**

- Gene structure identification
- Regulatory sequences identification
- Differential Gene Expression
- Variants annotation
- etc...

- Genome Wide Association Analysis

80% of the variants associated with diseases localized outside of protein-coding genes

(Manolio et al., Hindorff et al)

Genome Annotation

The computation phase of genome annotation

***Ab initio* Gene Prediction**

- Mathematical models rather than external evidence to identify the genes [GenomeScan, GeneWise, GeneID, Augustus, Splign, etc]

Evidence Alignment Stage

- ESTs, RNA-seq, and protein sequences are aligned to the assembled genome [STAR, Histar, Kallisto, etc.]
- Assemble transcripts [cufflinks, stringtie2, etc.]

Quality Control and Representation of annotated data

GTF/GFF format :

- uses controlled vocabularies
- guarantees interoperability between different analysis tools

DEEP LEARNING



- ▶ Complex patterns detection in large datasets

Outperform traditional approaches of machine learning

- Several models have been developed to predict gene expression or chromatin profiles from the DNA sequence :
Xpresso, DeepSEA, Basset, Basenji, Enformer, etc.

DEEP LEARNING



- ▶ Complex patterns detection in large datasets

Outperform traditional approaches of machine learning

- ▶ Many applications in genomics fields

Unsupervised methods *DNA sequences taxonomic classification*

Supervised methods *Genomic sequences annotation, gene expression prediction*

- Several models have been developed to predict gene expression or chromatin profiles from the DNA sequence :

Xpresso, DeepSEA, Basset, Basenji, Enformer, etc.

CROSS- / WITHIN-SPECIES

► Generalizable predictive models

Cross-species predictions

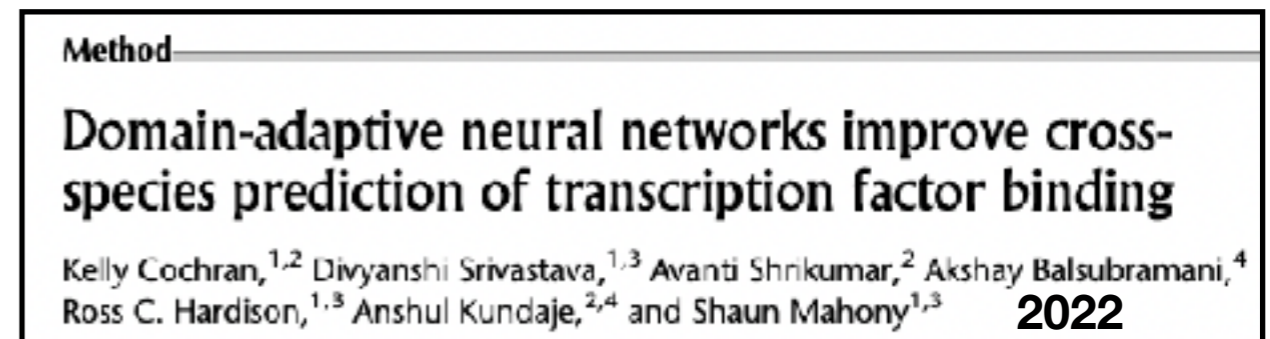
- * Take advantage of available data in one species to transfer knowledge to other species
- * Enhancers detection in mouse genomic sequences with human prediction model



► Species-specific prediction models are powerful tools

Within-species predictions

- * High performance
- * Predictions from whole DNA sequence (not only conserved regions)



• PhD Camille KERGAL : OBJECTIVES



- ▶ Creation of a deep learning model to predict canine gene expression
 - Data collection
 - Optimization strategy
- ▶ Assessment of the dog prediction model
 - Within-species and cross-species predictions
- ▶ Prediction of the impact of regulatory mutations on gene expression

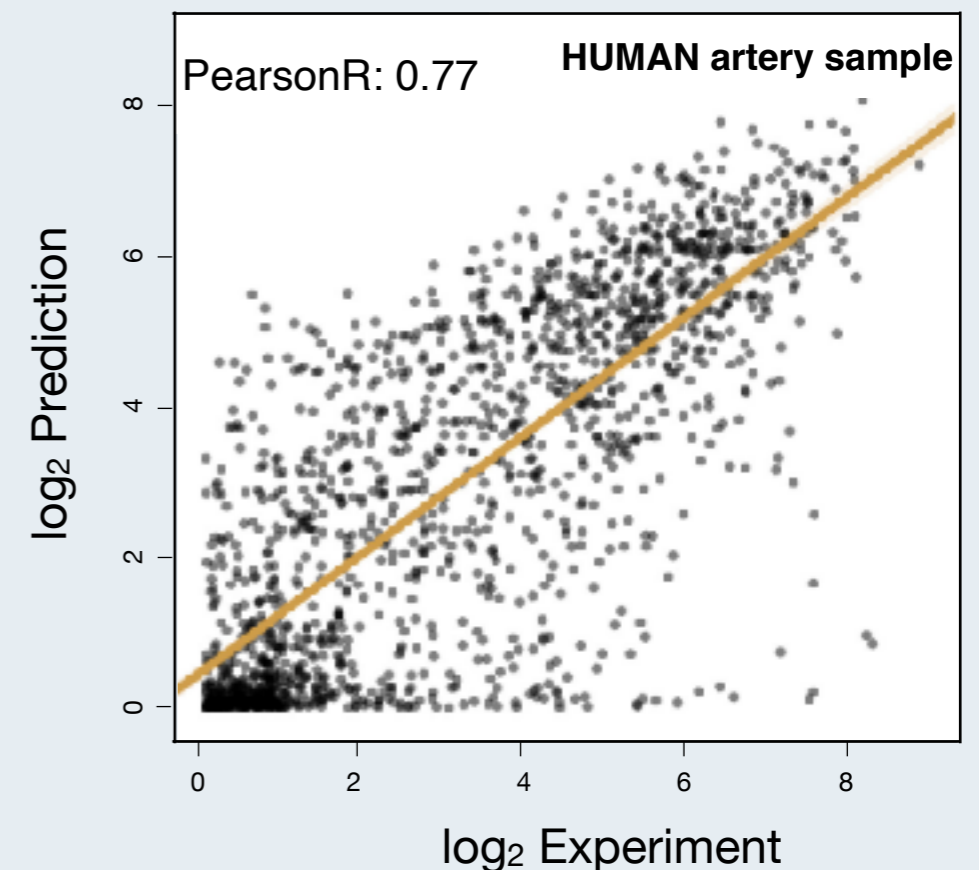
Method

Sequential regulatory activity prediction across chromosomes with convolutional neural networks



David R. Kelley,¹ Yakir A. Reshef,² Maxwell Bileschi,³ David Belanger,³ Cory Y. McLean,³ and Jasper Snoek³ 2018

- ▶ Tissue specific predictions of human gene expression and impact of mutations from DNA sequence
- ▶ Model CAGE signals as a function of DNA
- ▶ High correlations between predicted and expected expression
- ▶ Well documented tool: codes and prediction model publicly available on GitHub



MÉTHODE · BASENJI

Réseau de neurones convolutif (CNN)


CTAGCGTGGCTATCGT

0 1 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0..

1ere ÉTAPE

séquence ADN

encodage One-hot

2eme ÉTAPE

MÉTHODE · BASENJI

Réseau de neurones convolutif (CNN)


CTAGCGTGGCTATCGT

1ere ÉTAPE

séquence ADN

0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0

encodage One-hot

-2.8 -0.1 0.6 0 0 0 -2.8 -2.9 0.5 0.6 0 0 0 -2.8 -0.1 0.6 -2.8 -0.1

2eme ÉTAPE

convolution

MÉTHODE · BASENJI

Réseau de neurones convolutif (CNN)

1ere ÉTAPE

●●●●●●●●●●●●●●●●
CTAGCGTGGCTATCGT

séquence ADN

0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0

encodage One-hot

-2.8 -0.1 0.6 0 0 0 -2.8 -2.9 0.5 0.6 0 0 0 -2.8 -0.1 0.6 -2.8 -0.1

2eme ÉTAPE

convolution

0 0 0.6 0 0 0 0 0 0.5 0.6 0 0 0 0 0 0.6 0 0

activation

MÉTHODE · BASENJI

Réseau de neurones convolutif (CNN)

1ere ÉTAPE

●●●●●●●●●●●●●●
CTAGCGTGGCTATCGT

séquence ADN

0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0

encodage One-hot

-2.8 -0.1 0.6 0 0 0 -2.8 -2.9 0.5 0.6 0 0 0 -2.8 -0.1 0.6 -2.8 -0.1

2eme ÉTAPE

convolution

0 0 0.6 0 0 0 0 0 0.5 0.6 0 0 0 0 0 0.6 0 0
0 0.6 0 0 0.6 0 0 0.6 0

activation

agrégation

MÉTHODE · BASENJI

Réseau de neurones convolutif (CNN)

1ere ÉTAPE

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●
CTAGCGTGGCTATCGT

séquence ADN

0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 encodage One-hot

-2.8 -0.1 0.6 0 0 0 -2.8 -2.9 0.5 0.6 0 0 0 -2.8 -0.1 0.6 -2.8 -0.1

2eme ÉTAPE

convolution

0 0 0.6 0 0 0 0 0 0.5 0.6 0 0 0 0 0 0.6 0 0

activation

0 0.6 0 0 0.6 0 0 0.6 0

agrégation

MÉTHODE · BASENJI

Réseau de neurones convolutif (CNN)

1ere ÉTAPE

●●●●●●●●●●●●●●
CTAGCGTGGCTATCGT

séquence ADN

0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 encodage One-hot

2eme ÉTAPE

-2.8 -0.1 0.6 0 0 0 -2.8 -2.9 0.5 0.6 0 0 0 -2.8 -0.1 0.6 -2.8 -0.1

convolution

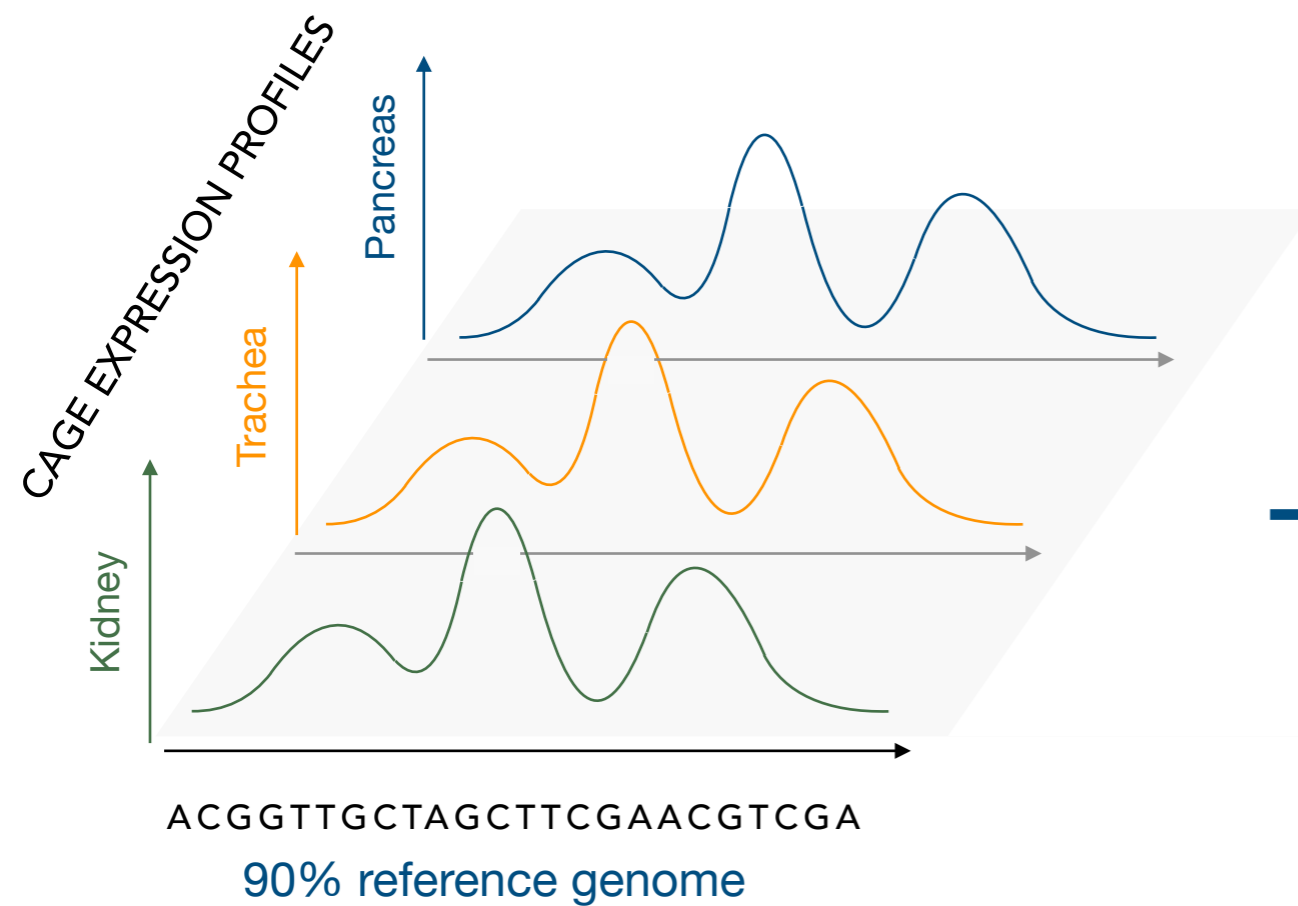
0 0 0.6 0 0 0 0 0 0.5 0.6 0 0 0 0 0 0.6 0 0

activation

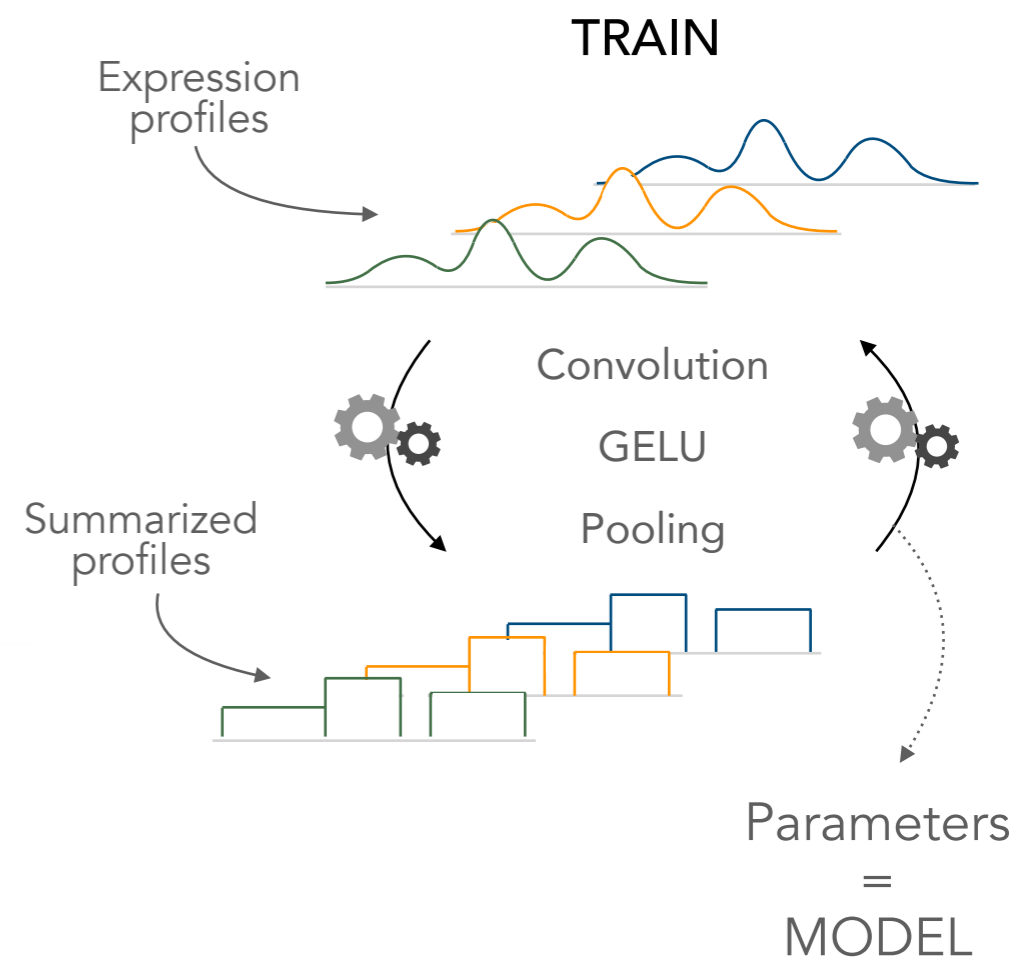
0 0.6 0 0 0.6 0 0 0.6 0

agrégation

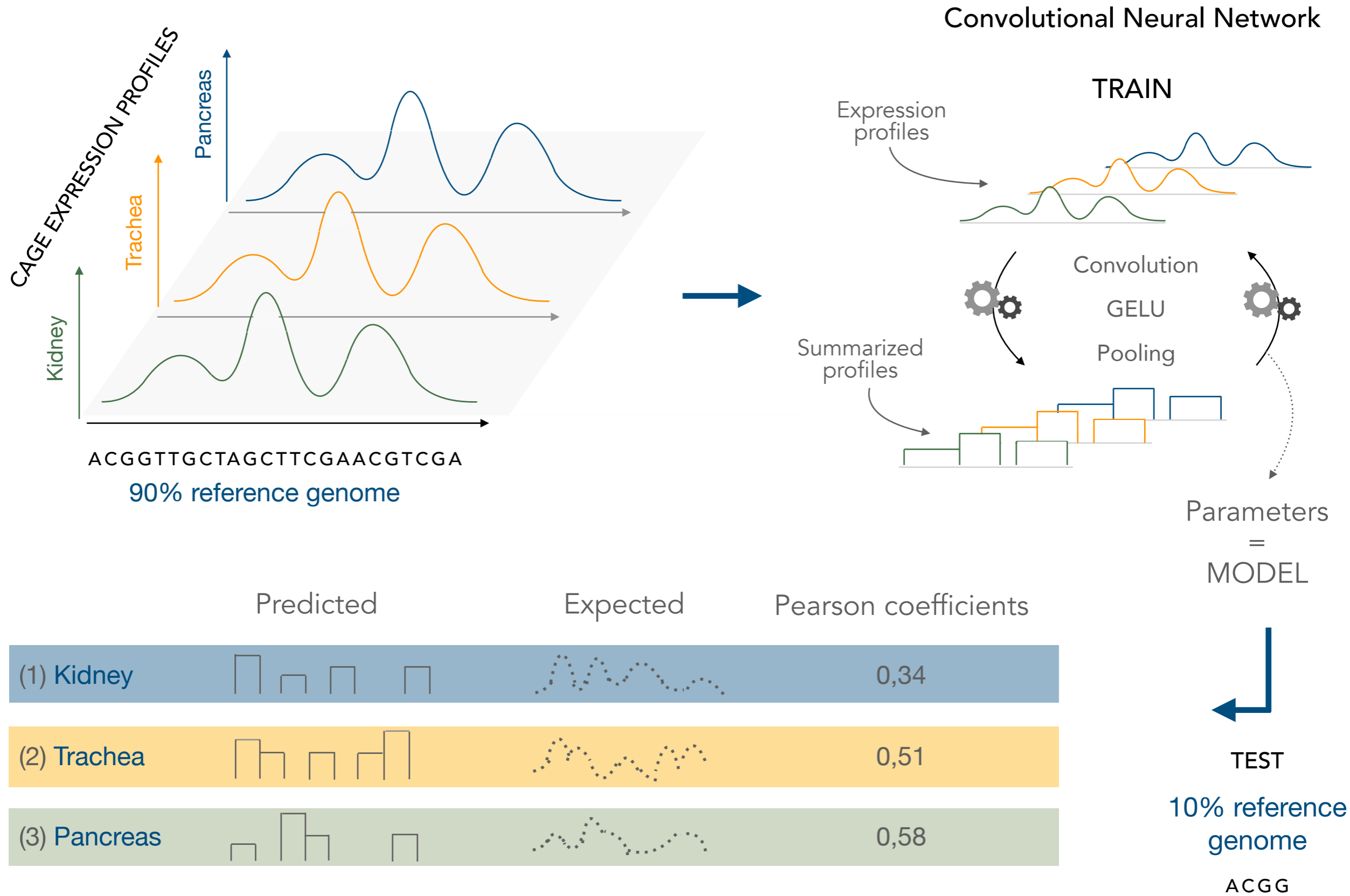
METHOD · BASENJI



Convolutional Neural Network



METHOD · BASENJI



- ▶ Development of canine gene expression prediction model

Adaptation of Basenji

Collaboration with DoGA consortium

- * Most comprehensive set of canine CAGE profiles
- * 116 canine CAGE profiles representing 37 core tissues



Dog Genome Annotation
project consortium

University of Helsinki, Karolinska Institutet

- ▶ **Basenji-Like IM**portant variant **P**rediction <https://github.com/ckergal/BLIMP>

- * Recommendations about how to process CAGE data and how to use prediction model
- * Promoting the canine prediction model

RESULTS · DOG PREDICTION MODEL

Correlation between predicted and real expression

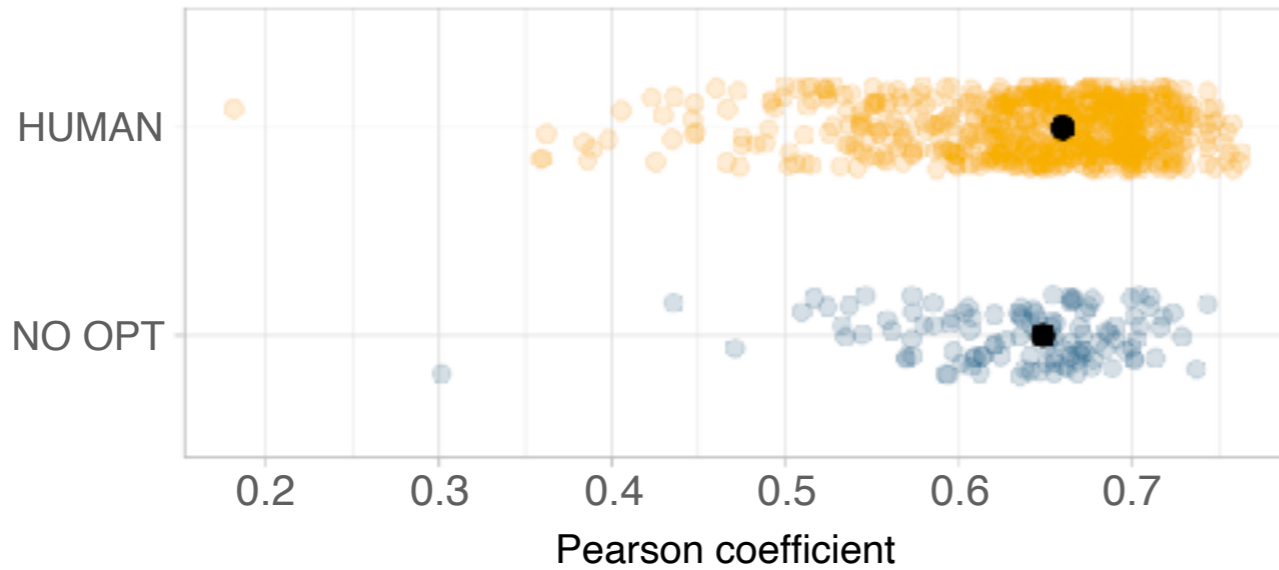


Min 0,18 Max 0,76 Median 0,64

Min 0,30 Max 0,74 Median 0,63

RESULTS · DOG PREDICTION MODEL

Correlation between predicted and real expression



Min 0,18

Max 0,76

Median 0,64

Min 0,30

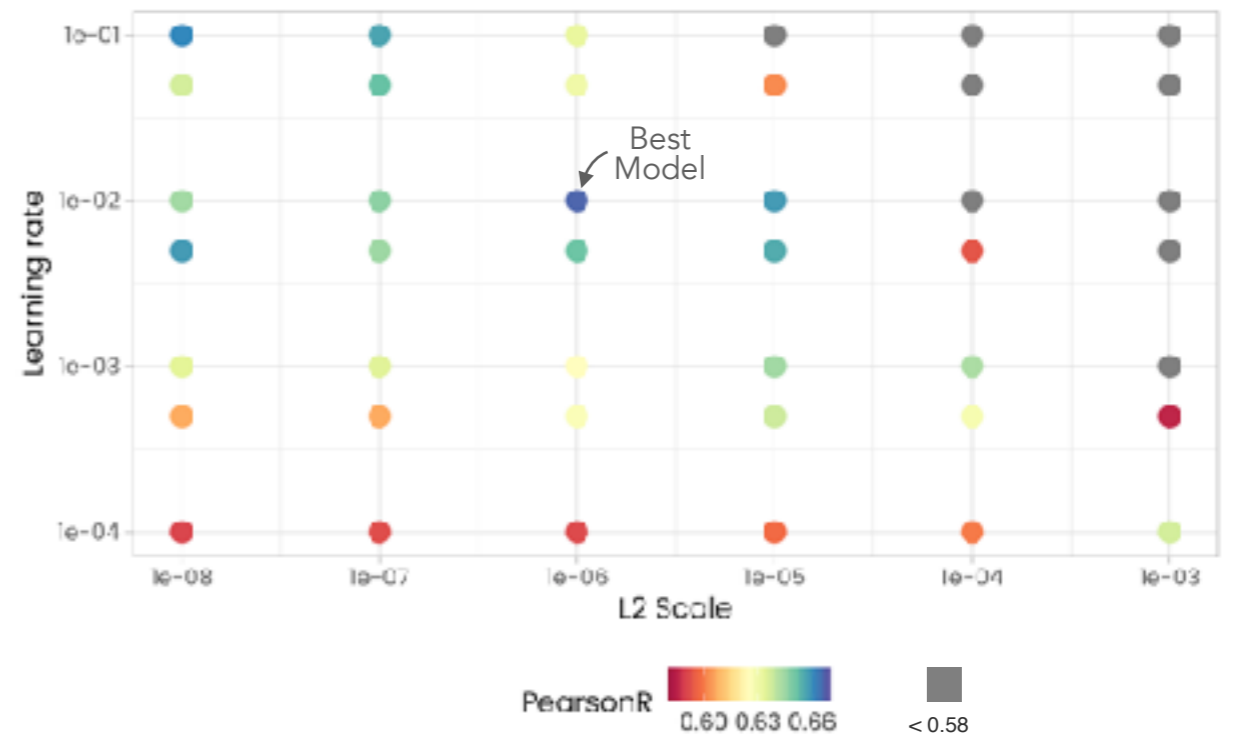
Max 0,74

Median 0,63

► Hyperparameters optimisation

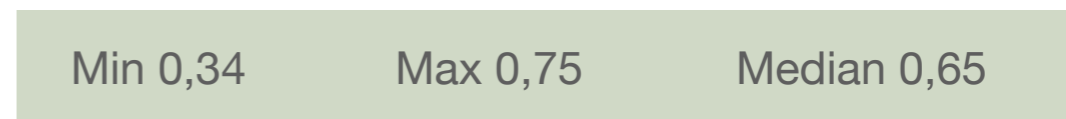
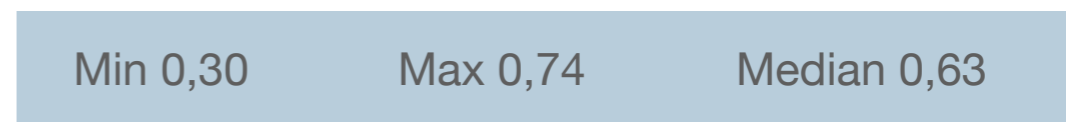
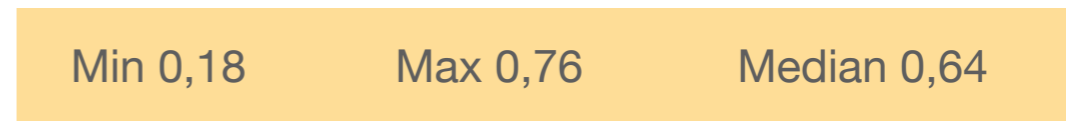
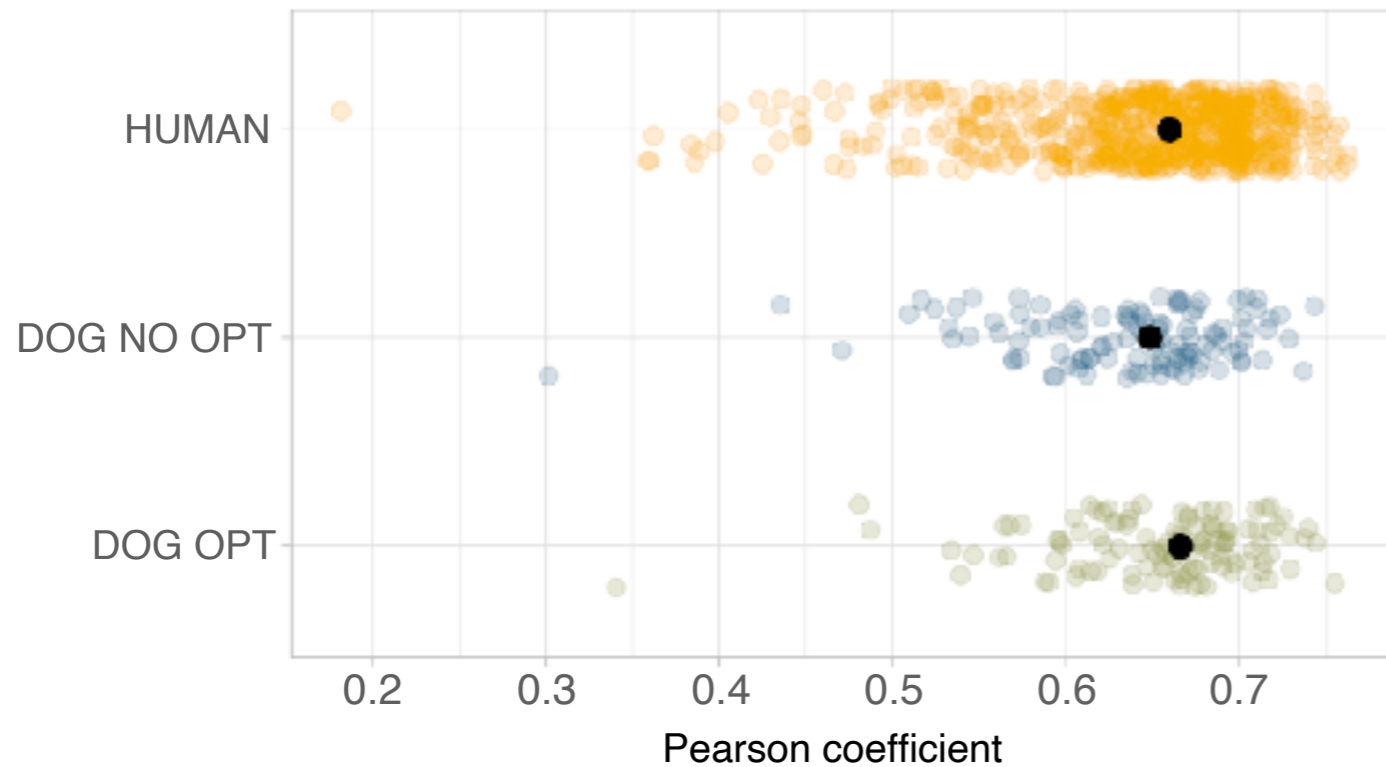
- * Focus on Learning Rate and L2 Scale
- * Selection of the model leading to the best median Pearson correlation

Median Pearson according to HP values



RESULTS · DOG PREDICTION MODEL

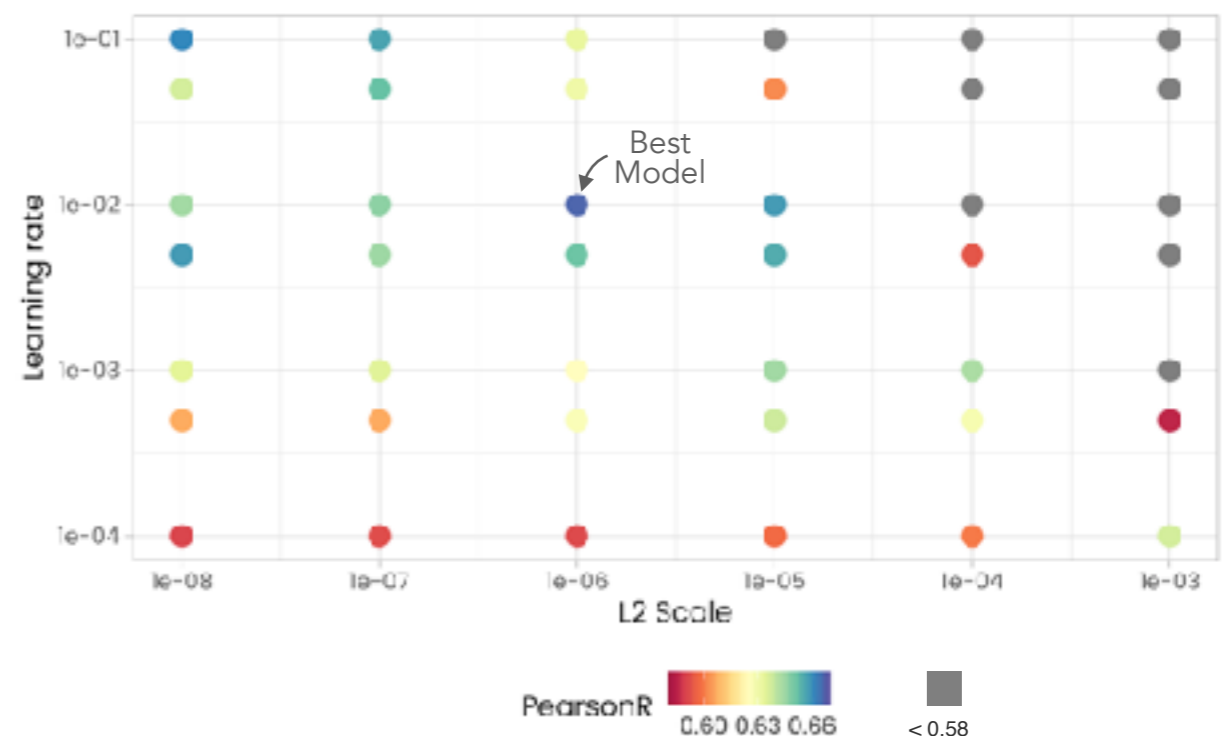
Correlation between predicted and real expression



► Hyperparameters optimisation

- * Focus on Learning Rate and L2 Scale
- * Selection of the model leading to the best median Pearson correlation

Median Pearson according to HP values



• (PhD Camille KERGAL) OBJECTIVES

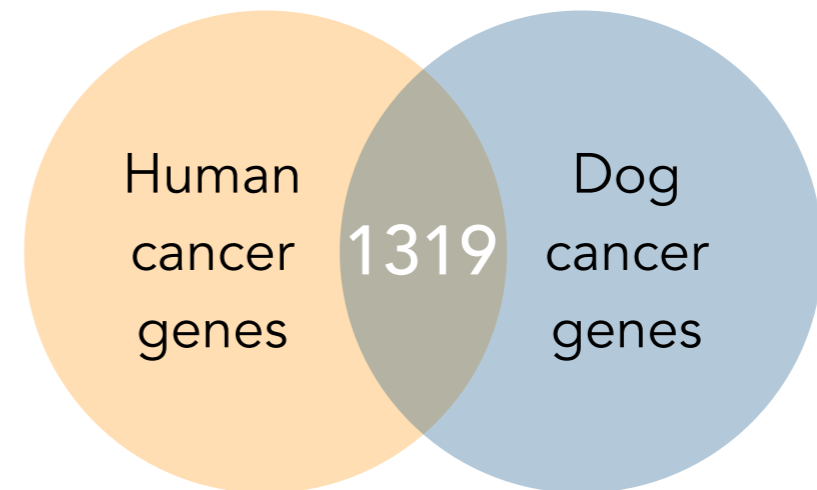
- ▶ Creation of a deep learning model to predict canine gene expression
 - Data collection
 - Optimization strategy *In order to obtain the most powerful prediction model*

- ▶ Assessment of the dog prediction model
 - Within-species and cross-species predictions

- ▶ Orthologous genes between human and dog

Focus on genes implicated in cancers

Wong et al. 2021

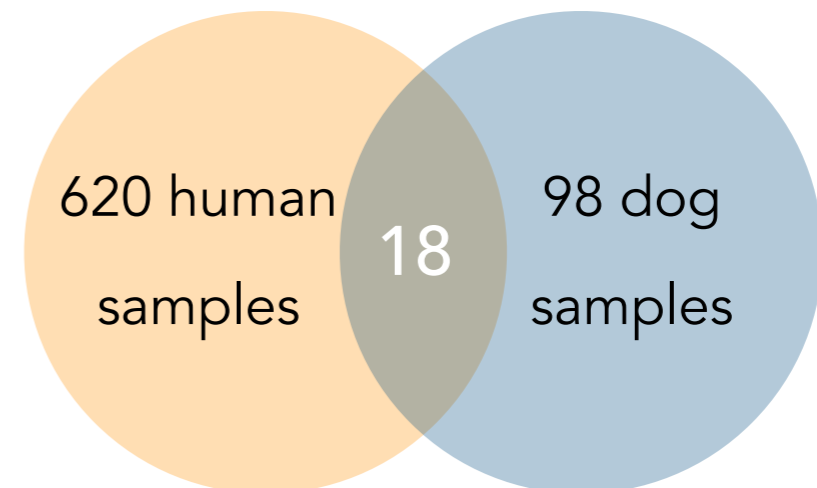


KIT
CDK4
BRAF
NRAS
MDM2
...

- ▶ Matching tissues between both species

18 common samples

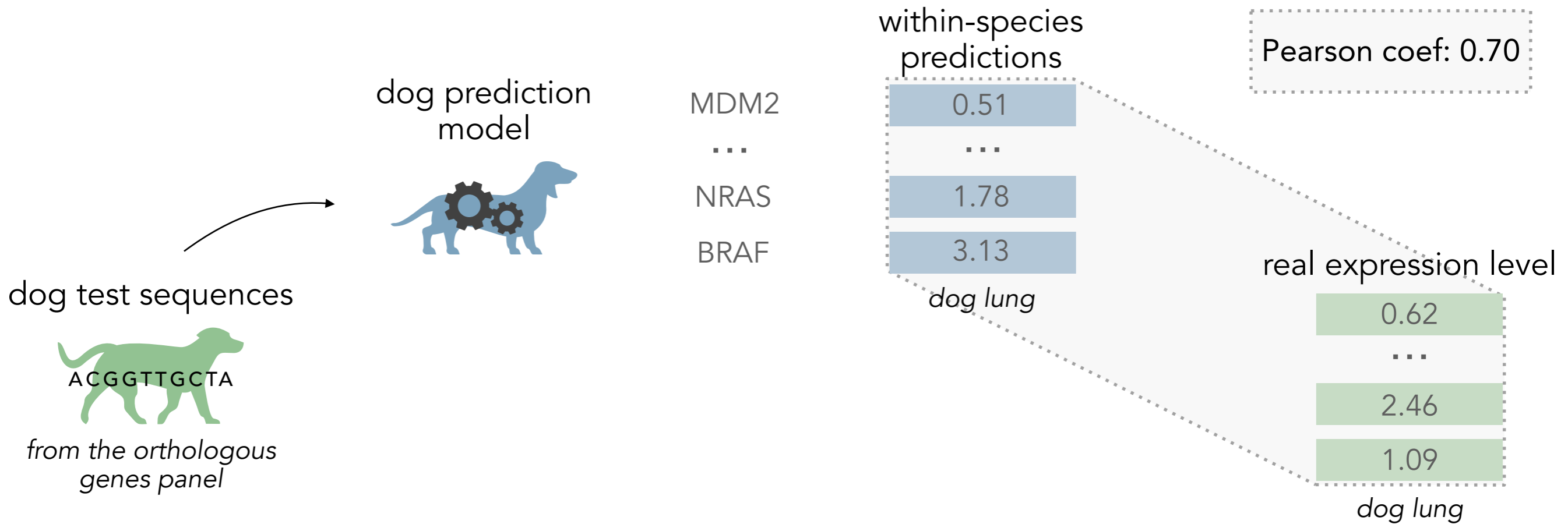
- * 638 composing the human prediction model
- * 116 composing the dog prediction model



RETINA
SKIN
LUNG
...

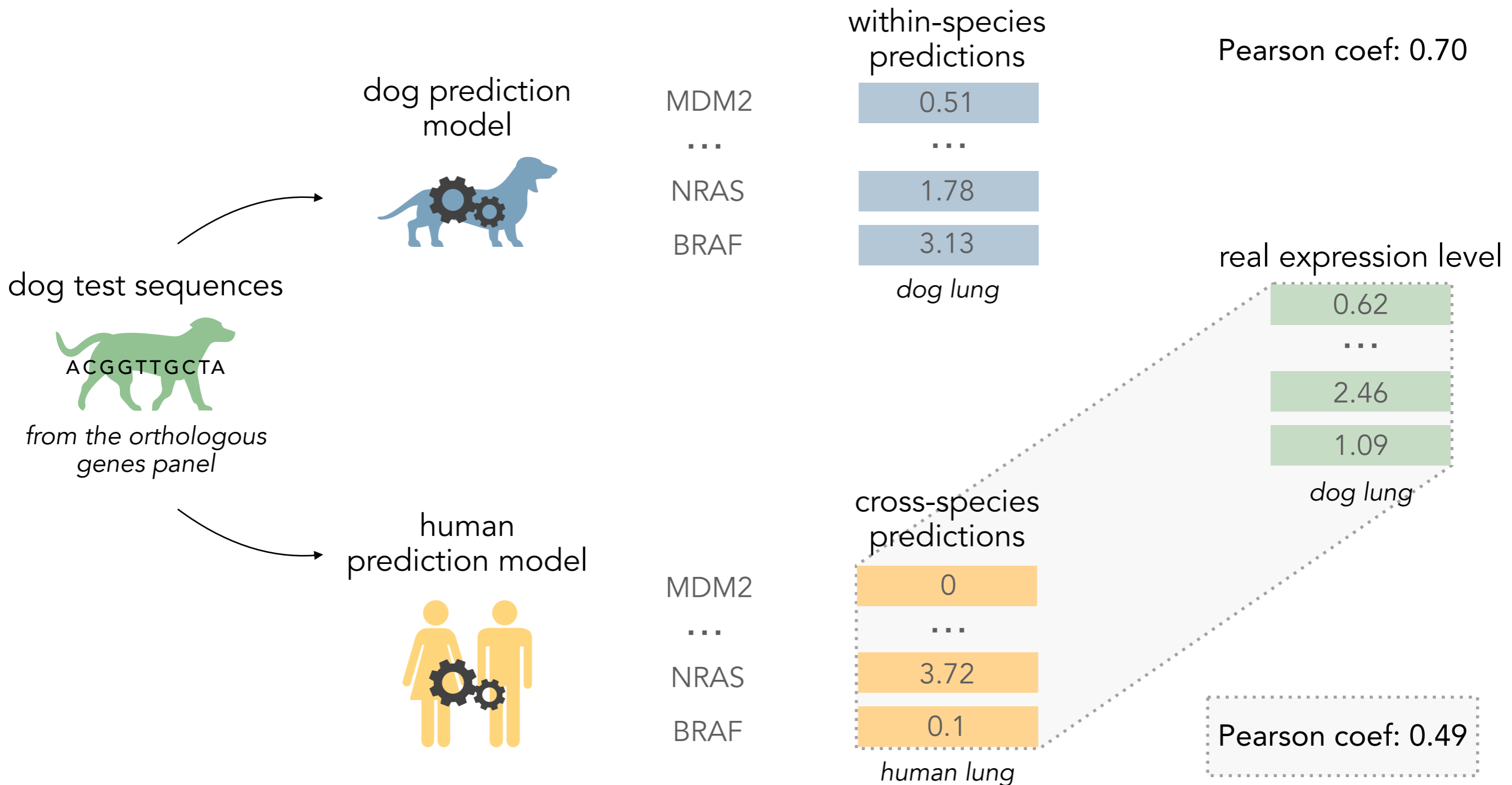
METHOD · CROSS-SPECIES / WITHIN-SPECIES

- ▶ Within-species (dog specific) compared to cross-species predictions



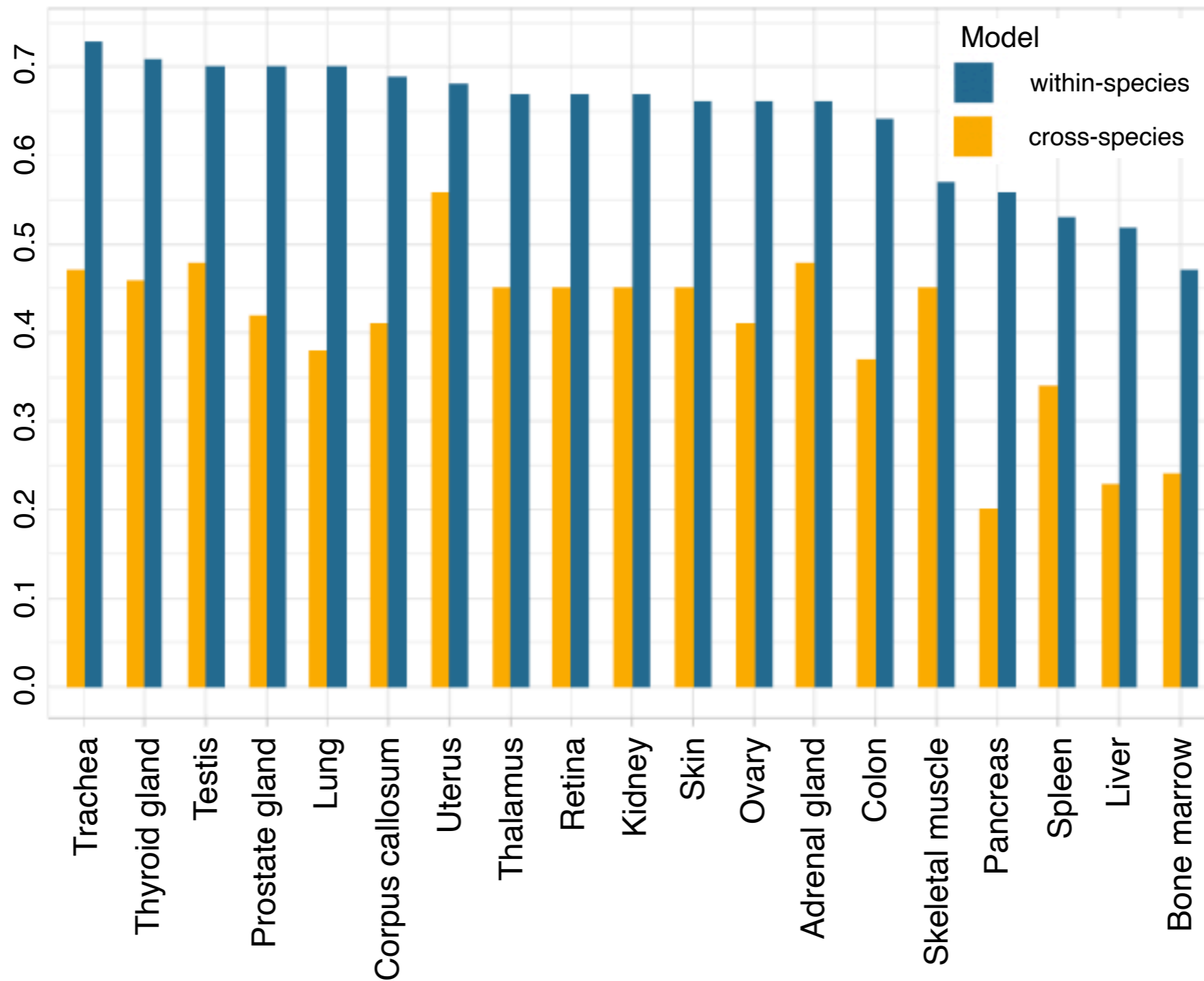
METHOD · CROSS-SPECIES / WITHIN-SPECIES

- ▶ Within-species (dog specific) compared to cross-species predictions



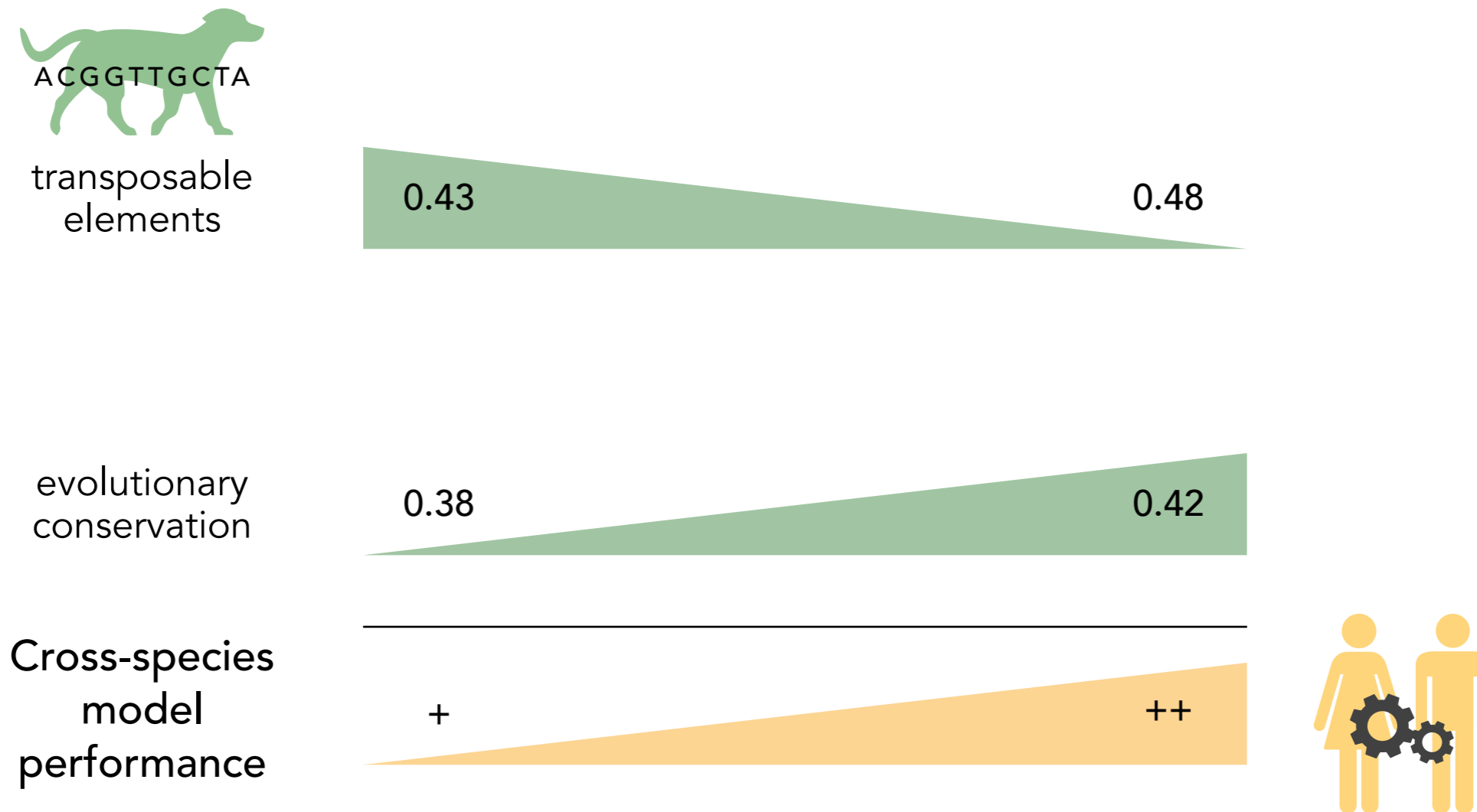
RESULTS · CROSS-SPECIES / WITHIN-SPECIES

- ▶ Canine sequence prediction assessment for each matching tissues



RESULTS · CROSS-SPECIES / WITHIN-SPECIES

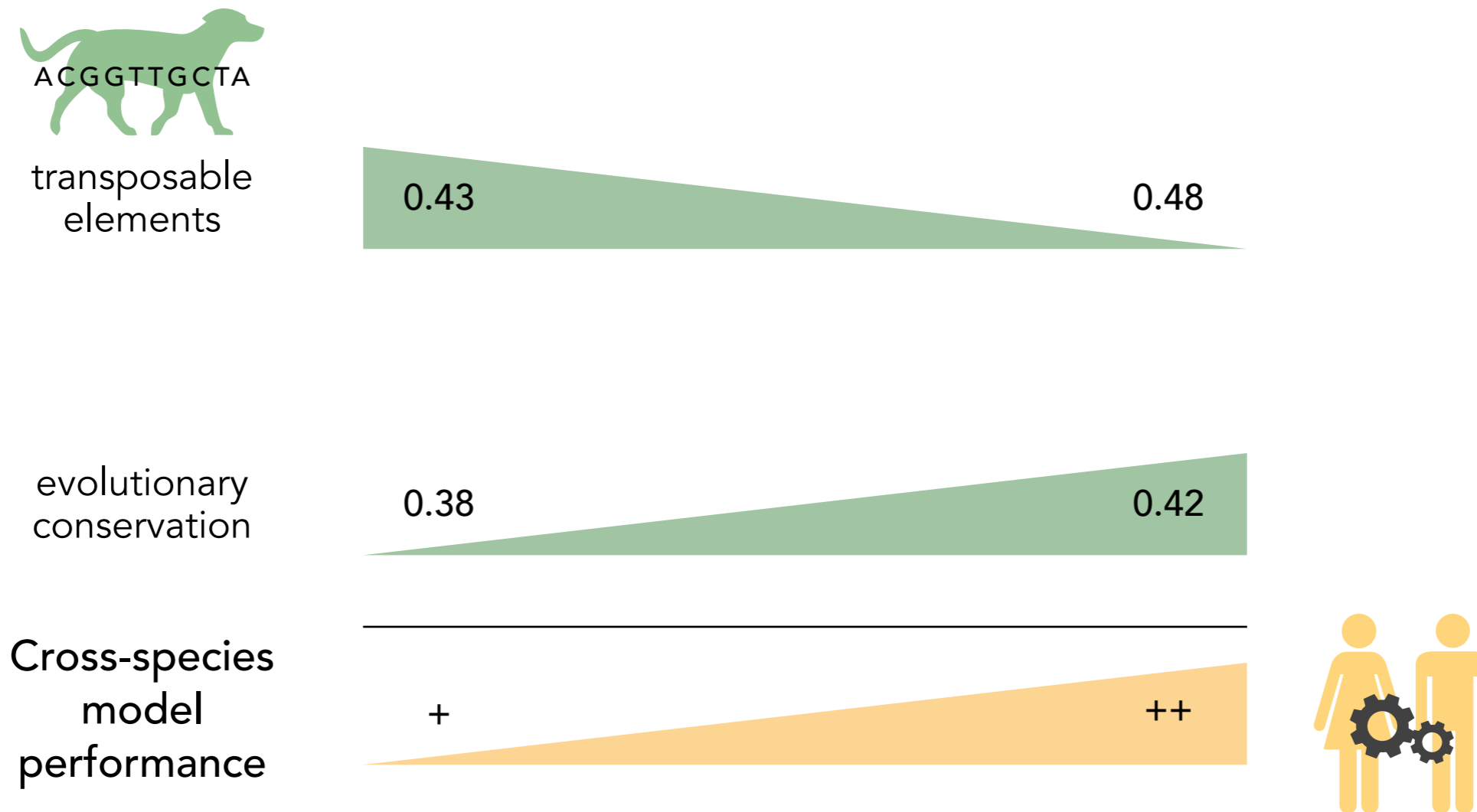
- ▶ Canine genomic features influencing cross-species predictions



- ▶ Cross-species approach leads to better results when dog sequences features approximate human genomics features

RESULTS · CROSS-SPECIES / WITHIN-SPECIES

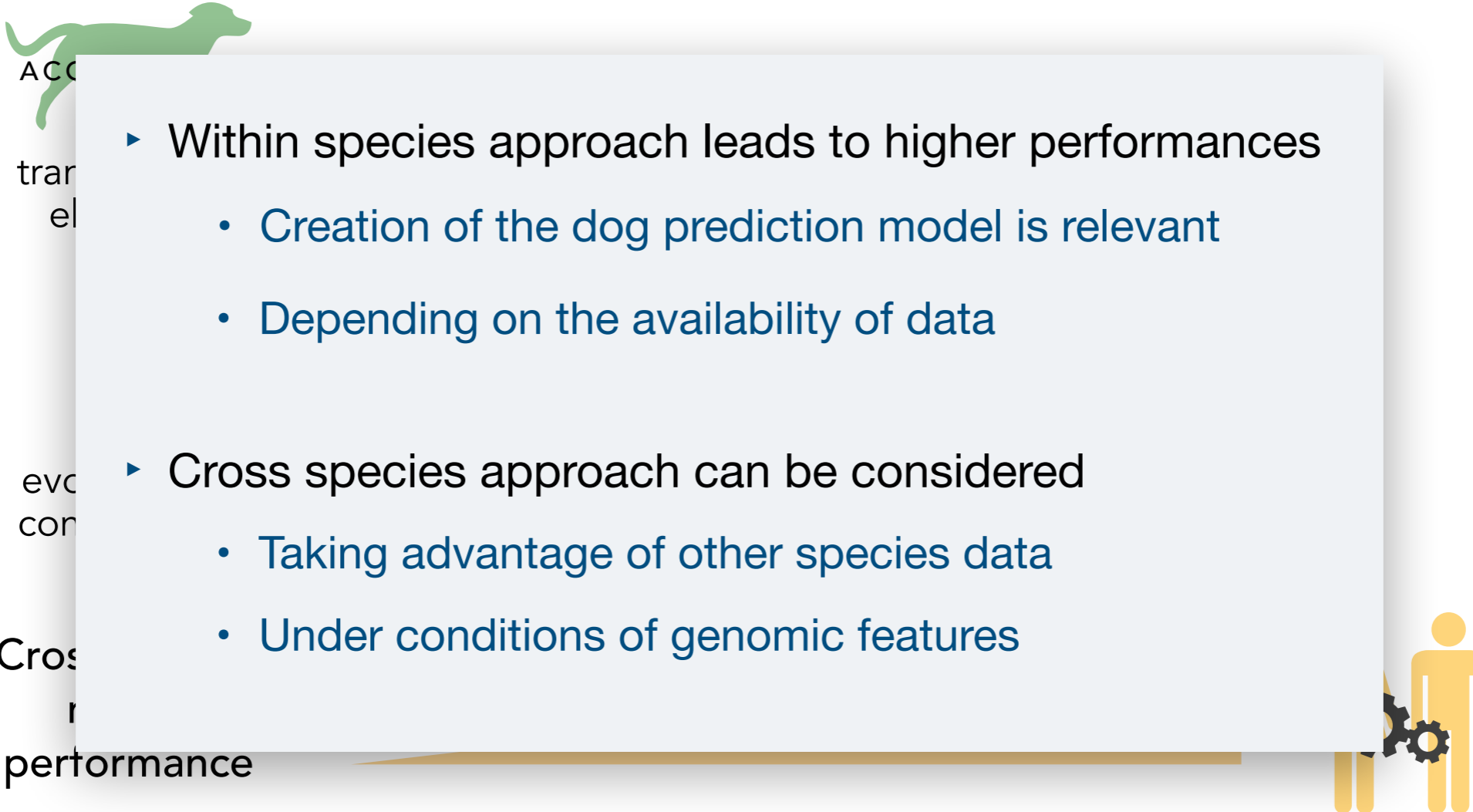
- ▶ Canine genomic features influencing cross-species predictions



- ▶ Cross-species approach leads to better results when dog sequences features approximate human genomics features

RESULTS · CROSS-SPECIES / WITHIN-SPECIES

- ▶ Canine genomic features influencing cross-species predictions



ACC

tran

el

- ▶ Within species approach leads to higher performances
 - Creation of the dog prediction model is relevant
 - Depending on the availability of data
- ▶ Cross species approach can be considered
 - Taking advantage of other species data
 - Under conditions of genomic features

evc

con

Cros

r

performance

- ▶ Cross-species approach leads to better results when dog sequences features approximate human genomics features

OBJECTIVES · THIRD PART

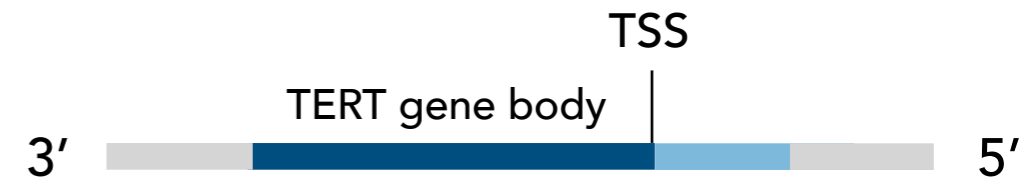
- ▶ Creation of a deep learning model to predict canine gene expression
 - Data collection
 - Optimization strategy *In order to obtain the most powerful prediction model*

- ▶ Assessment of the dog prediction model
 - Within-species and cross-species predictions *Performance comparison*

- ▶ Prediction of the impact of regulatory mutations on gene expression
 - *In silico* saturated mutagenesis in promoters of cancer genes
 - Analysis of variants predicted as impacting

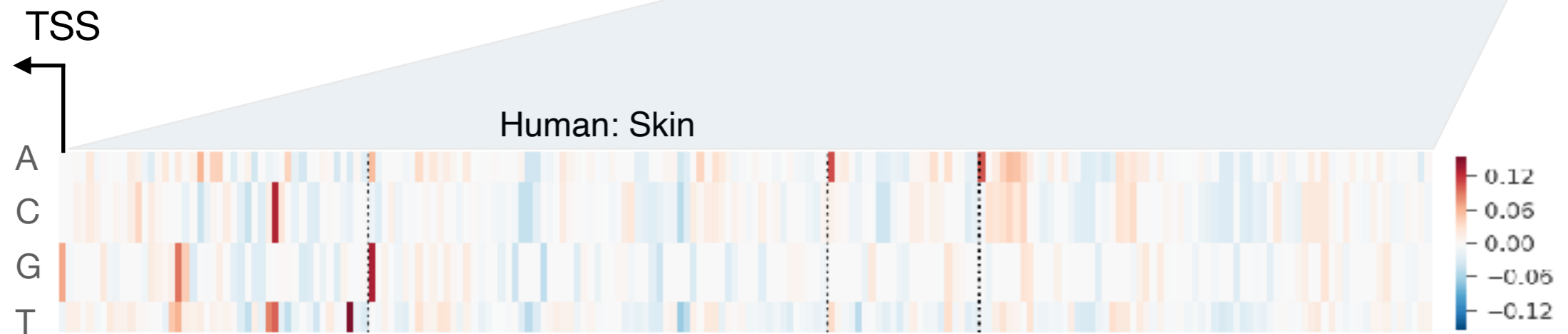
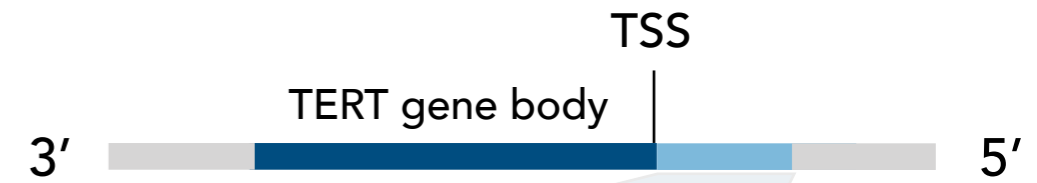
METHOD · IN SILICO SATURATED MUTAGENESIS

- ▶ TERT : gene implied in various cancers



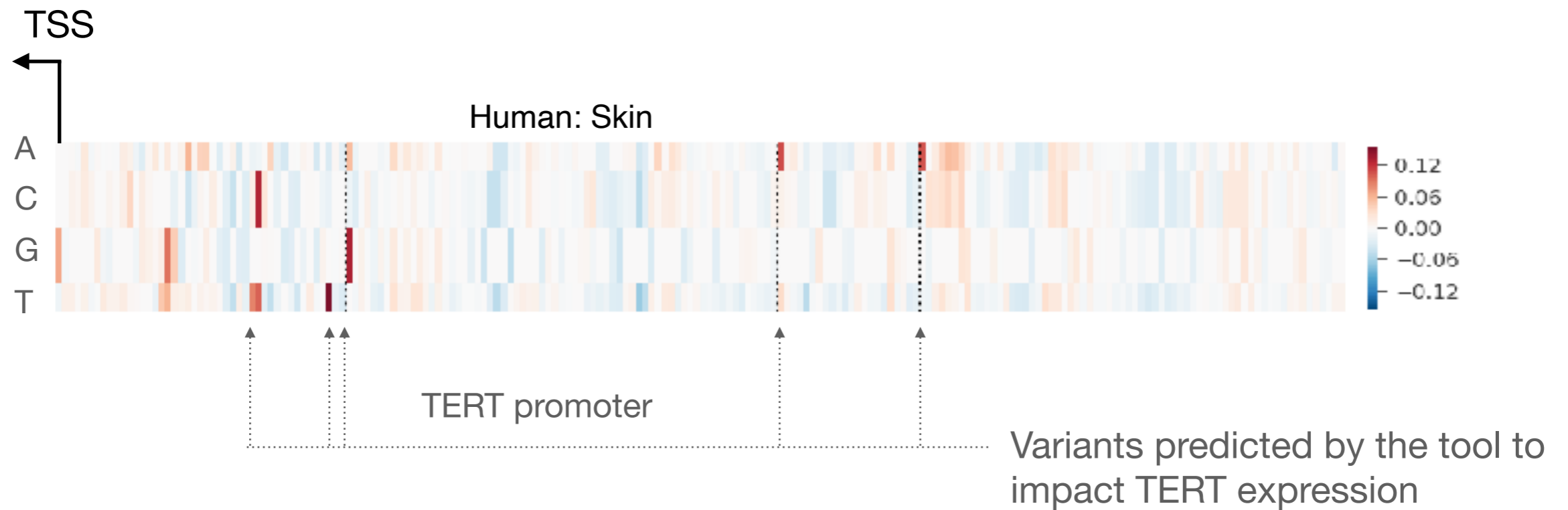
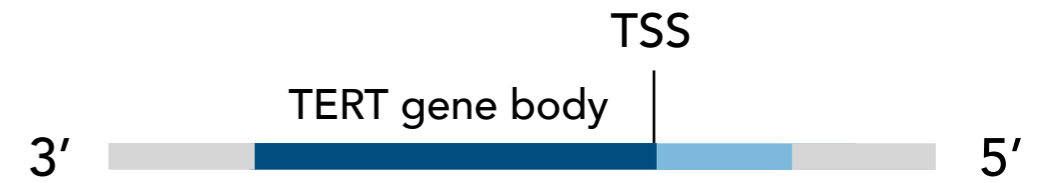
METHOD · IN SILICO SATURATED MUTAGENESIS

- ▶ TERT : gene implied in various cancers



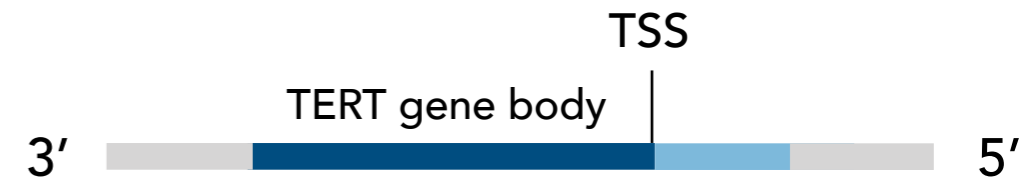
METHOD · IN SILICO SATURATED MUTAGENESIS

- ▶ TERT : gene implied in various cancers

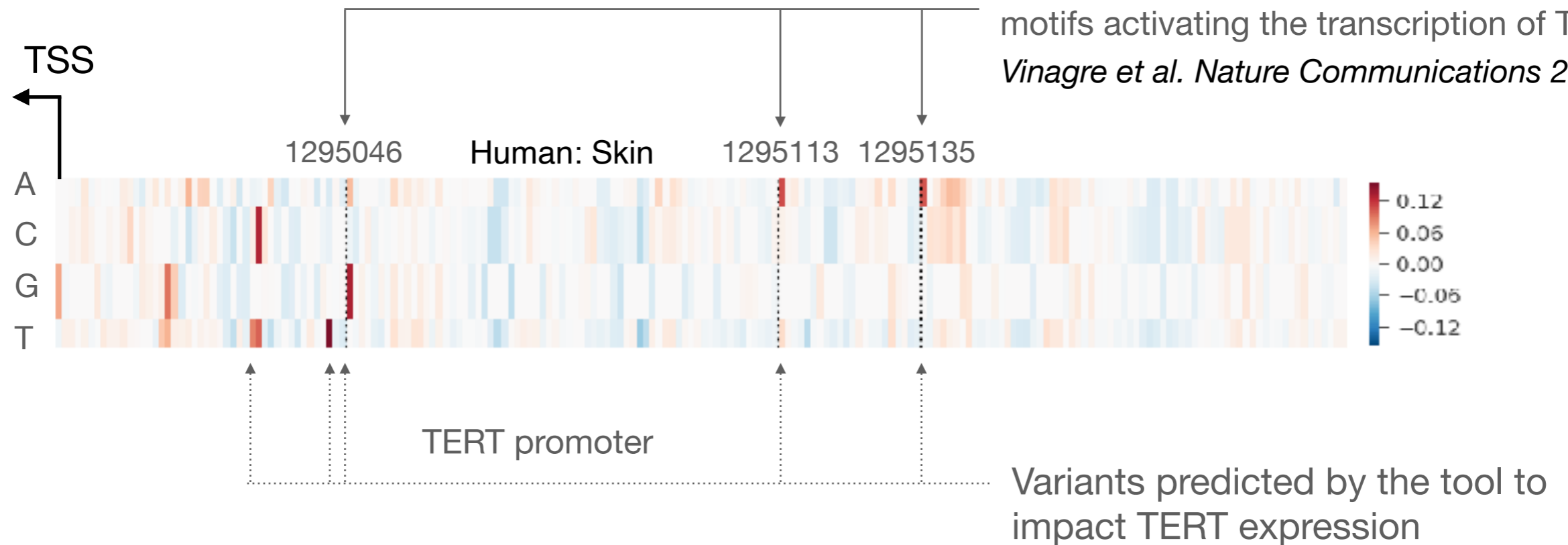


METHOD · IN SILICO SATURATED MUTAGENESIS

- ▶ TERT : gene implied in various cancers



Variants known to create new binding motifs activating the transcription of TERT
Vinagre et al. Nature Communications 2013



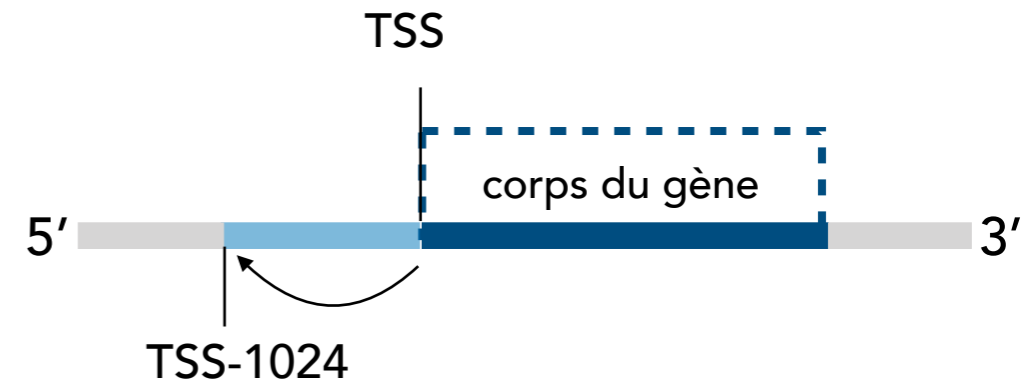
- ▶ Predictions of non-coding mutations impact in TERT promoters relevant with the human model

RÉSULTATS · PREDICTION D'IMPACT de MUTATION

► Application de mutagenèse saturée *in silico*

- * Gènes orthologues

- * Promoteurs : 1024 nucléotides en amont du TSS



► Modèles de prédiction de l'expression

- * Utilisation des deux modèles : modèle humain / modèle canin

- * Prédictions dans 19 tissus

RÉSULTATS · MUTATIONS IMPACTANTES

- ▶ L'outil prédit un score d'impact pour chaque mutation possible

1024 positions génomiques X 1317 gènes X 4 mutations X 19 tissus

+ 100 millions de prédictions

Nombreuses mutations sans impact sur le niveau d'expression

(Allèle muté identique à l'allèle de référence)

- ▶ Caractérisation des variants selon la prédiction d'impact

sans impact

écart-type entre 0 et 4

impact **faible**
à **modéré**

entre 4 et 8

impact **fort**

> 8

RÉSULTATS · MUTATIONS IMPACTANTES

- ▶ L'outil prédit un score d'impact pour chaque mutation possible

1024 positions génomiques X 1317 gènes X 4 mutations X 19 tissus

+ 100 millions de prédictions

Nombreuses mutations sans impact sur le niveau d'expression

(Allèle muté identique à l'allèle de référence)

- ▶ Caractérisation des variants selon la prédiction d'impact

sans impact

écart-type entre 0 et 4

impact faible
à modéré

entre 4 et 8

impact fort

> 8

- ▶ Prédiction de mutations à fort impact

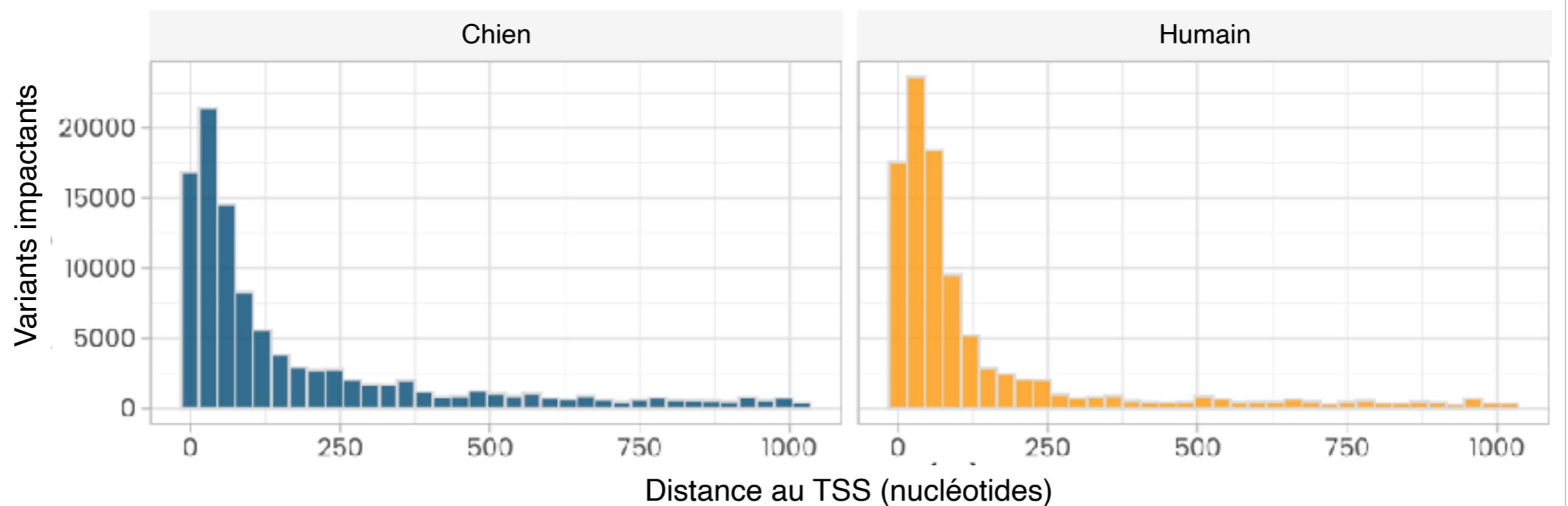
103 000 (7,6%) chez le chien / 98 000 (7,2%) chez l'humain

RÉSULTATS · MUTATIONS IMPACTANTES

- ▶ Étude comparative entre l'humain et le chien

- * 7,6% des mutations possibles sont prédites avec un fort impact chez le chien / 7,2% chez l'humain

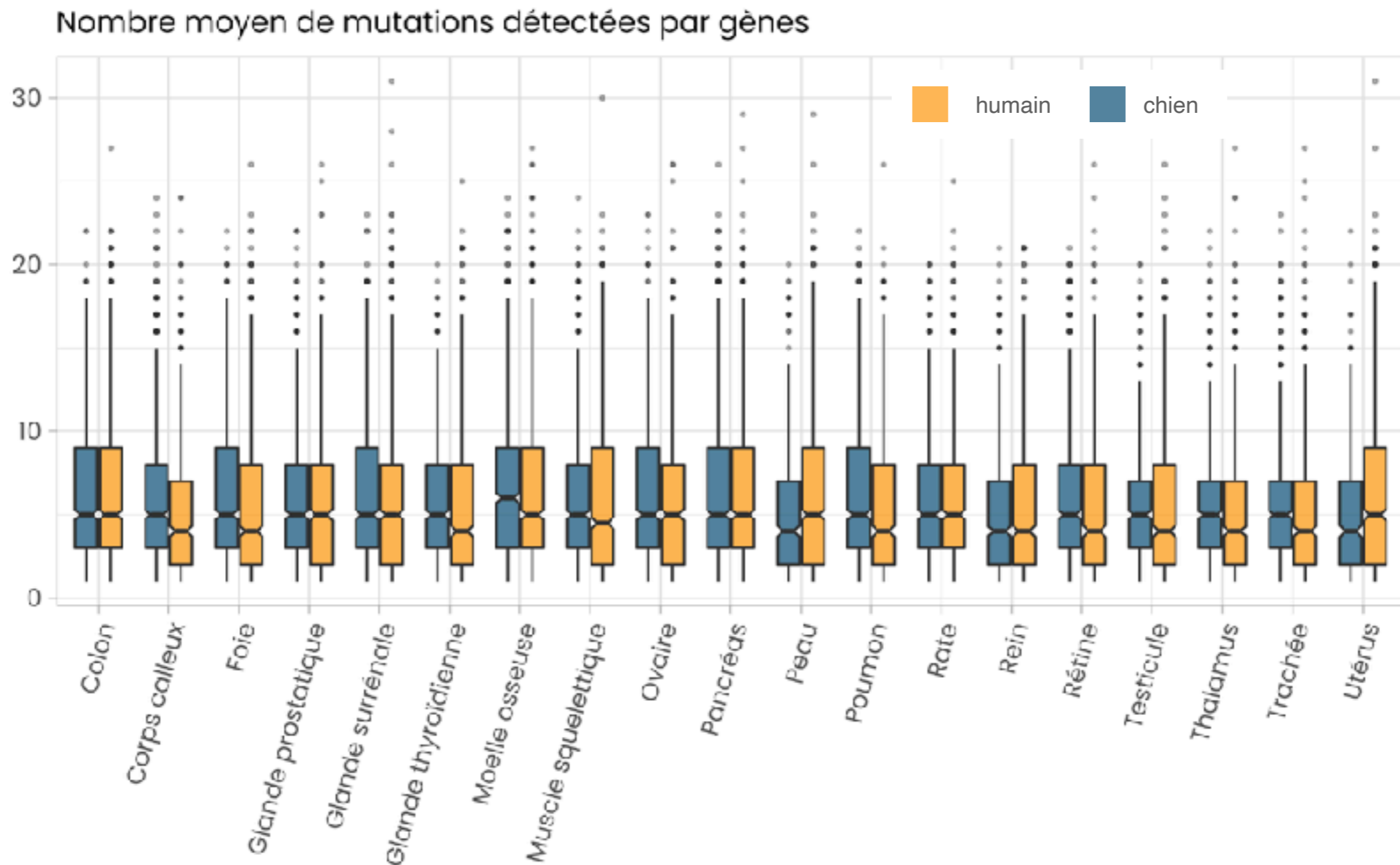
Distance au TSS des variants fortement impactants



RÉSULTATS · MUTATIONS IMPACTANTES

► Étude comparative entre l'humain et le chien

- * 7,6% des mutations possibles sont prédites avec un fort impact chez le chien / 7,2% chez l'humain



RÉSULTATS · MUTATIONS IMPACTANTES

- ▶ Étude comparative entre l'humain et le chien

- * 7,6% des mutations possibles sont prédites avec un fort impact chez le chien / 7,2% chez l'humain

Variants impactants

- ▶ Répartition similaire des variants impactants entre l'humain et le chien
- ▶ Enrichissement des variants impactants très proches du TSS :
TATA box : -10 TSS [TATAAT]
TTGACA motif :-35 TSS
- ▶ Motifs régulateurs conservés entre les deux espèces :
313 gènes

RÉSULTATS · ANALYSE VARIANTS DOG10K

- ▶ Données Dog10K : chien/loup/chien-village :
n=2000 WGS -> 33 millions de variants



- ▶ 15 500 variants identifiés parmi les promoteurs du panel de gènes (1317)

* 3 catégories :

sans impact

écart-type entre 0 et 4

impact
faible à modéré

entre 4 et 8

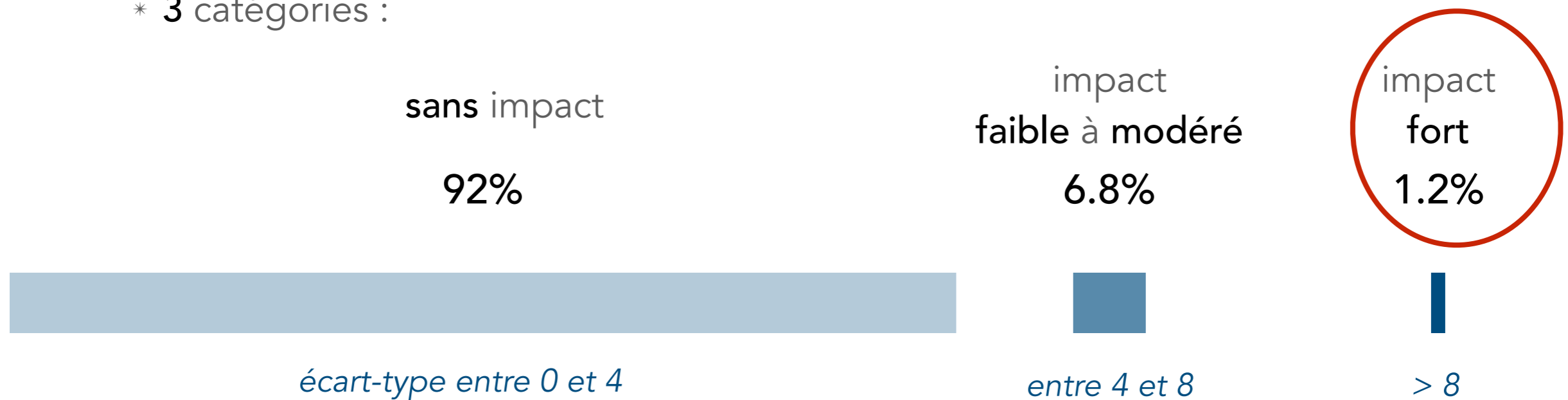
impact
fort

> 8

RÉSULTATS · ANALYSE VARIANTS DOG10K

- ▶ Données Dog10K : chien/loup/chien-village : n=2000
- ▶ 15 500 variants identifiés parmi les promoteurs du panel de gènes (1317)

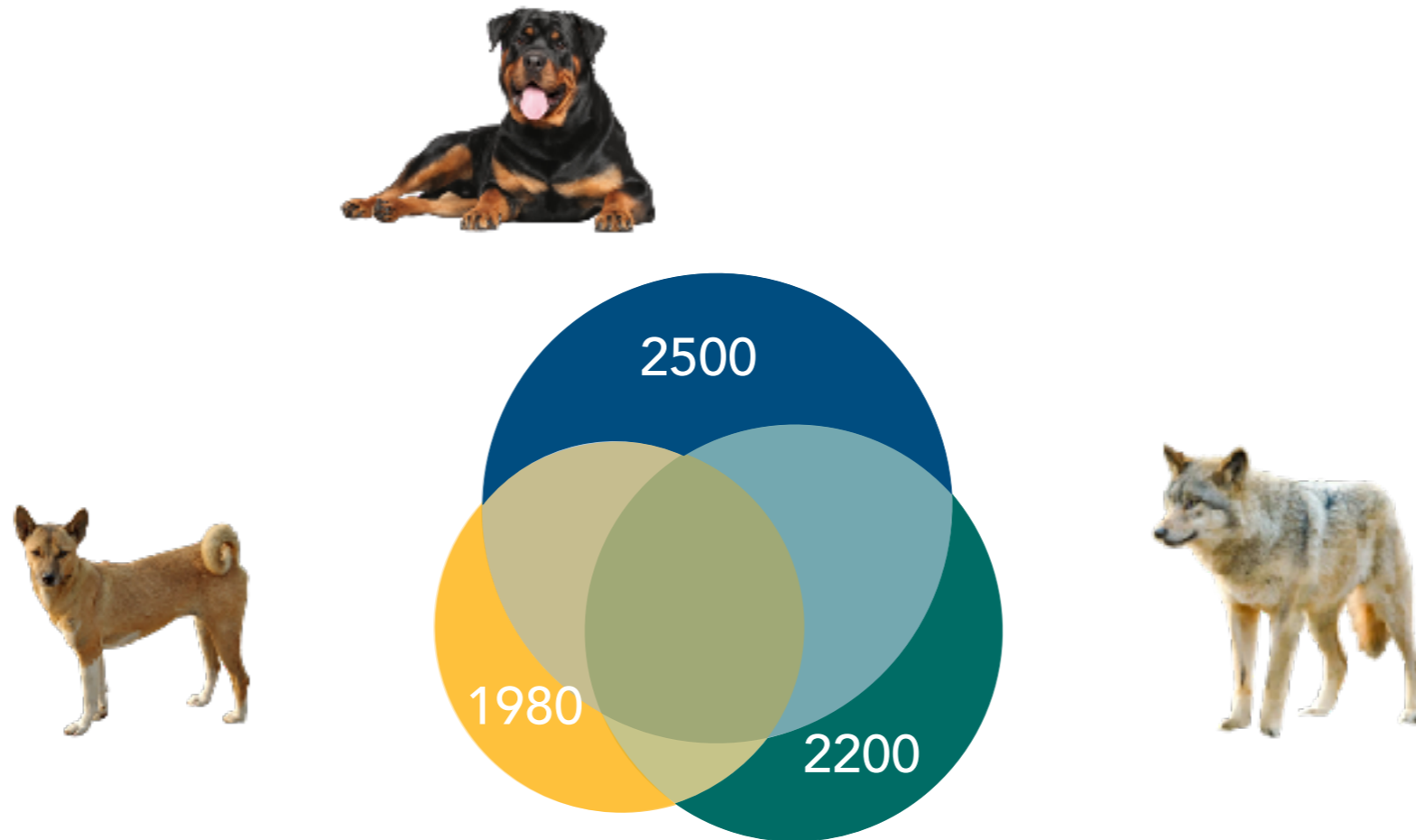
* 3 catégories :



* Identification de **186** variants représentés dans la base de données Dog10K et prédits avec un **impact fort** sur l'expression des gènes cancers

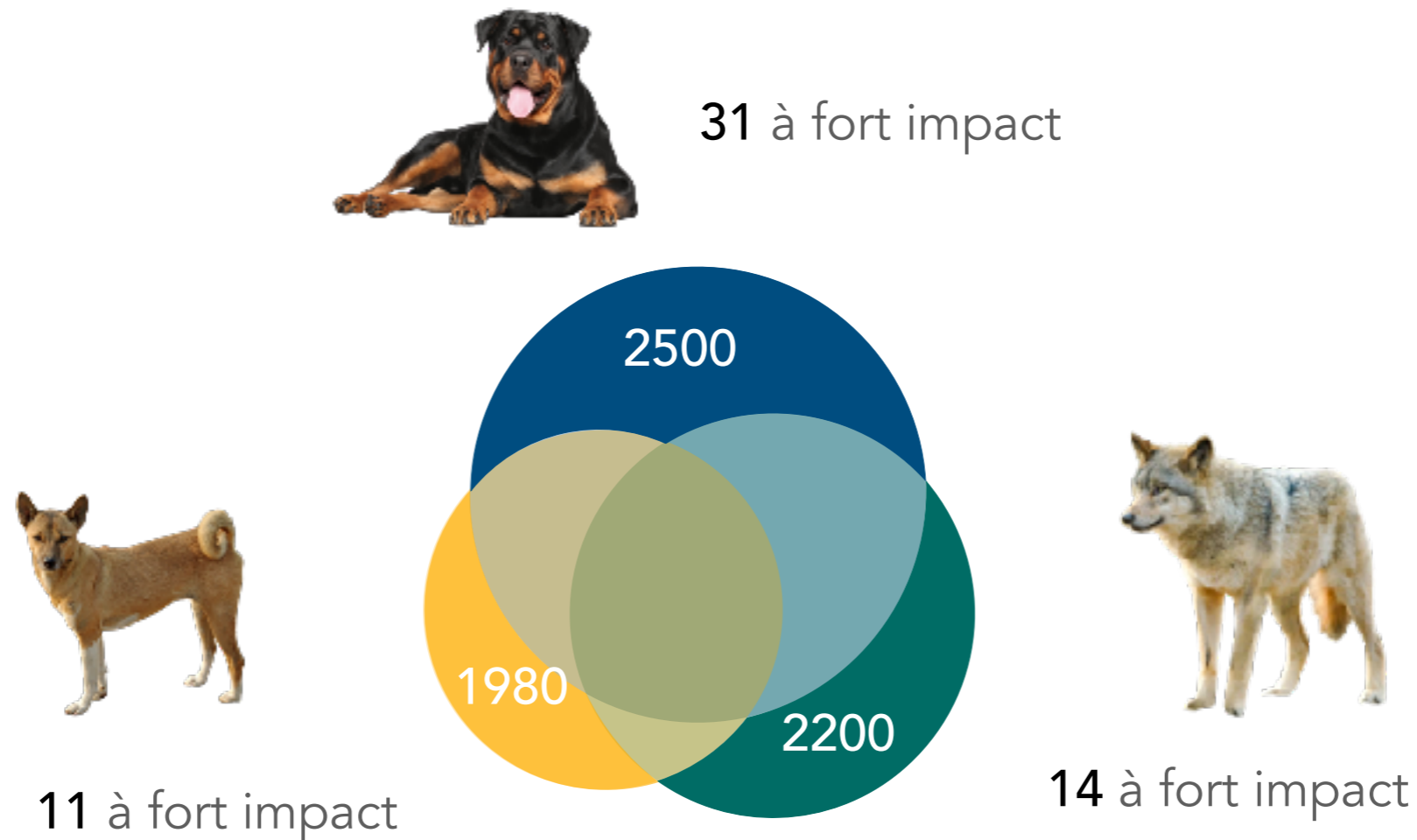
RÉSULTATS · ANALYSE VARIANTS DOG10K

- ▶ Données Dog10K : 15 500 variants identifiés dans la population des canidés



RÉSULTATS · ANALYSE VARIANTS DOG10K

- ▶ Données Dog10K : 15 500 variants identifiés dans la population des canidés



- * **Chiens** : nombre de variants à fort impact statistiquement plus **élevé** que chez les **loups** ou les **chiens de village**

RÉSULTATS · ANALYSE VARIANTS DOG10K

- ▶ Données Dog10K : 15 500 variants identifiés dans la population des canidés



31 à fort impact

- ▶ Hiérarchisation des variants du Dog10K
Catégorisation selon l'impact
- ▶ Maintien de variants délétères
Conséquence probable de sélection artificielle

* **Chiens** : nombre de variants à fort impact statistiquement plus **élevé** que chez les **loups** ou les **chiens de village**

CONCLUSION

canFam3 / canFam4

- ▶ Creation de modèles optimisés de l'expression des gènes canins
 - * Performances **équivalentes** avec le modèle de l'expression des gènes humains
 - * Intérêt de l'approche **intra-espèce**

- ▶ Creation de modèles optimisés de l'expression des gènes canins
 - * Performances **équivalentes** avec le modèle de l'expression des gènes humains
 - * Intérêt de l'approche **intra-espèce**

- ▶ Analyse des régions régulatrices des gènes de cancer
 - * Évaluation de la **mutagenèse saturée *in silico*** de l'outil Basenji
 - * Utilisation avec le modèle humain et le modèle canin

- ▶ Creation de modèles optimisés de l'expression des gènes canins
 - * Performances **équivalentes** avec le modèle de l'expression des gènes humains
 - * Intérêt de l'approche **intra-espèce**

- ▶ Analyse des régions régulatrices des gènes de cancer
 - * Évaluation de la **mutagenèse saturée *in silico*** de l'outil Basenji
 - * Utilisation avec le modèle humain et le modèle canin

- ▶ Identification de mutations impactantes
 - * Analyse comparative entre l'humain et le chien
 - * Croisement avec des variants présents dans la population canine

CONCLUSION

- ▶ Creation of optimized prediction models of canine gene expression

- * **Equivalent** performance with the human gene expression model

*canFam3 and
UU_Cfam_GSD*

- * Relevance of the **within species** approach

versions



CONCLUSION

- ▶ Creation of optimized prediction models of canine gene expression

- * **Equivalent** performance with the human gene expression model

*canFam3 and
UU_Cfam_GSD
versions*

- * Relevance of the **within species** approach

- ▶ Analysis of cancer genes regulatory regions

- * Assessment of the *in silico* saturated mutagenesis of the Basenji tool

- * Identification of non-coding mutations potentially involved in cancers

- * Promoting the canine prediction model with **BLIMP**



*Basenji-Like **IM**portant variant **P**redictor*



<https://github.com/ckergal/BLIMP>

Acknowledgements



- Camille Kergal
- Thomas Derrien
- Marie Dominique Galibert
- Dog Genetics Team



- Hannes Lohi
- Matthias Hortenhuber

► Hosting structures and funding



