ABEILLE: a novel method for ABerrant Expression Identification empLoying machine LEarning from RNA-sequencing data

Justine LABORY, PhD student

Medical Data Laboratory, IRCAN

Université Cote d'Azur, Nice, France

justine.labory@etu.univ-cotedazur.fr

Medical Context

Less than 1 person out of **2000**

30 millions

50 % **Children**

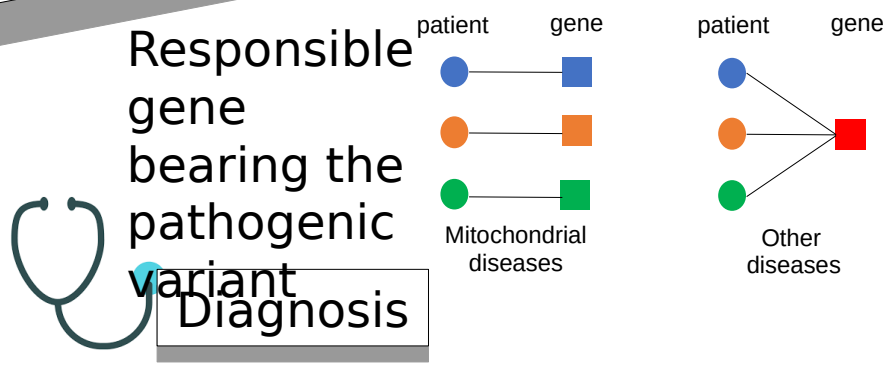
Rare diseases

responsible for a wide variety of biochemical processes

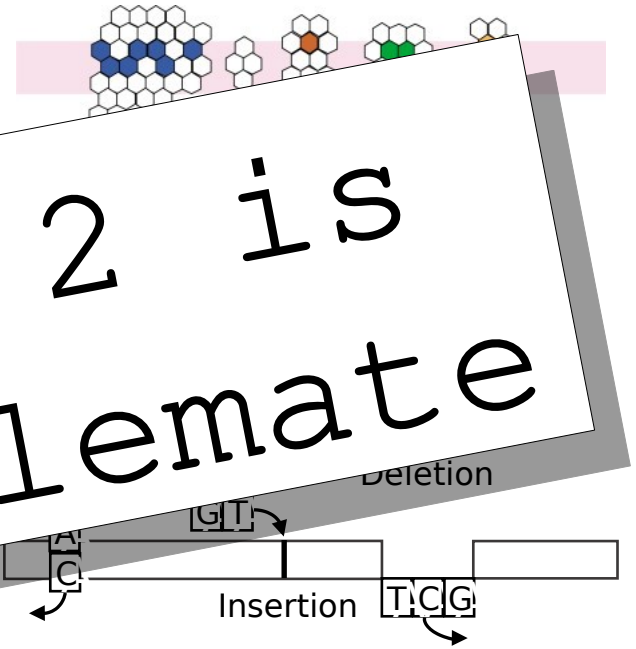
Mitochondria

1 patient out of 2 is in diagnostic stalemate

Responsible gene bearing the pathogenic variant



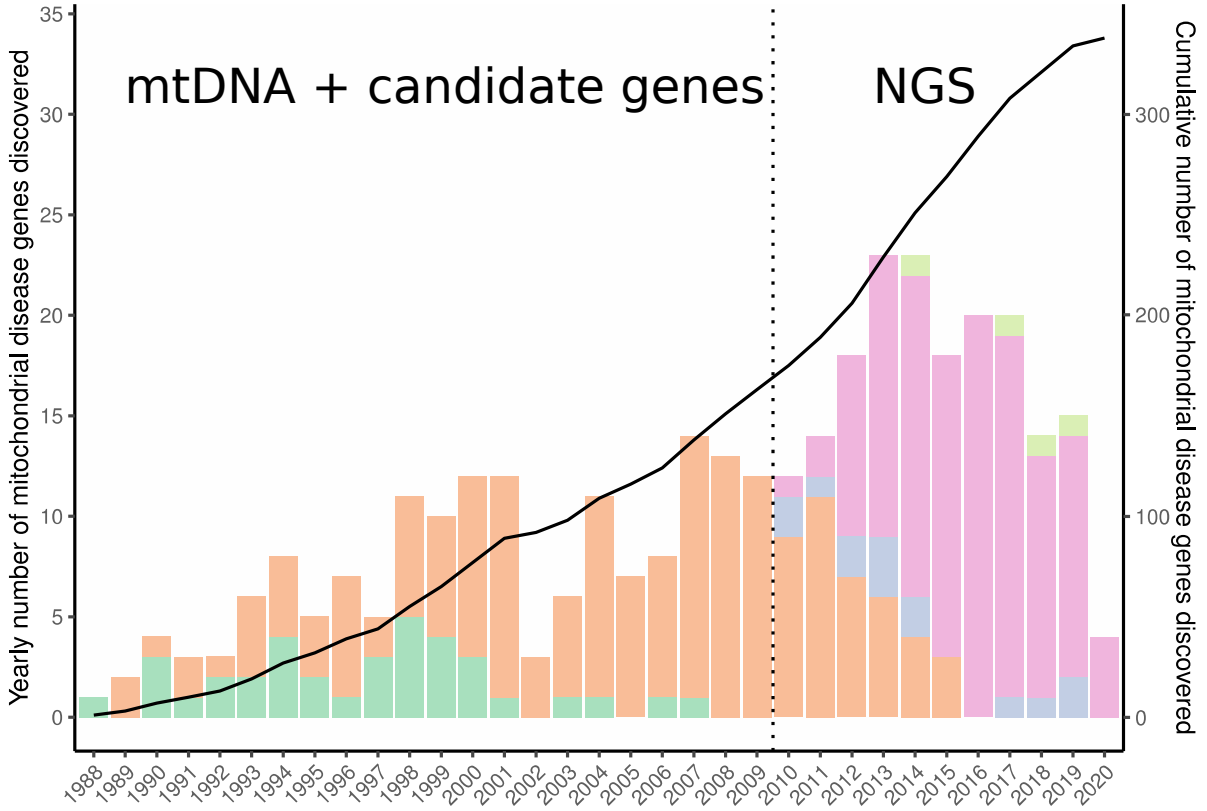
Diagnosis

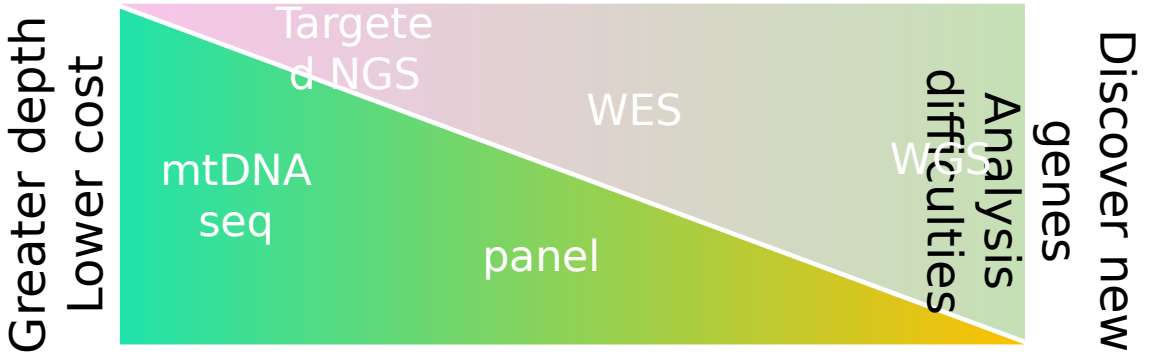


due to rare hereditary or spontaneous variants of mtDNA or nDNA

Disease

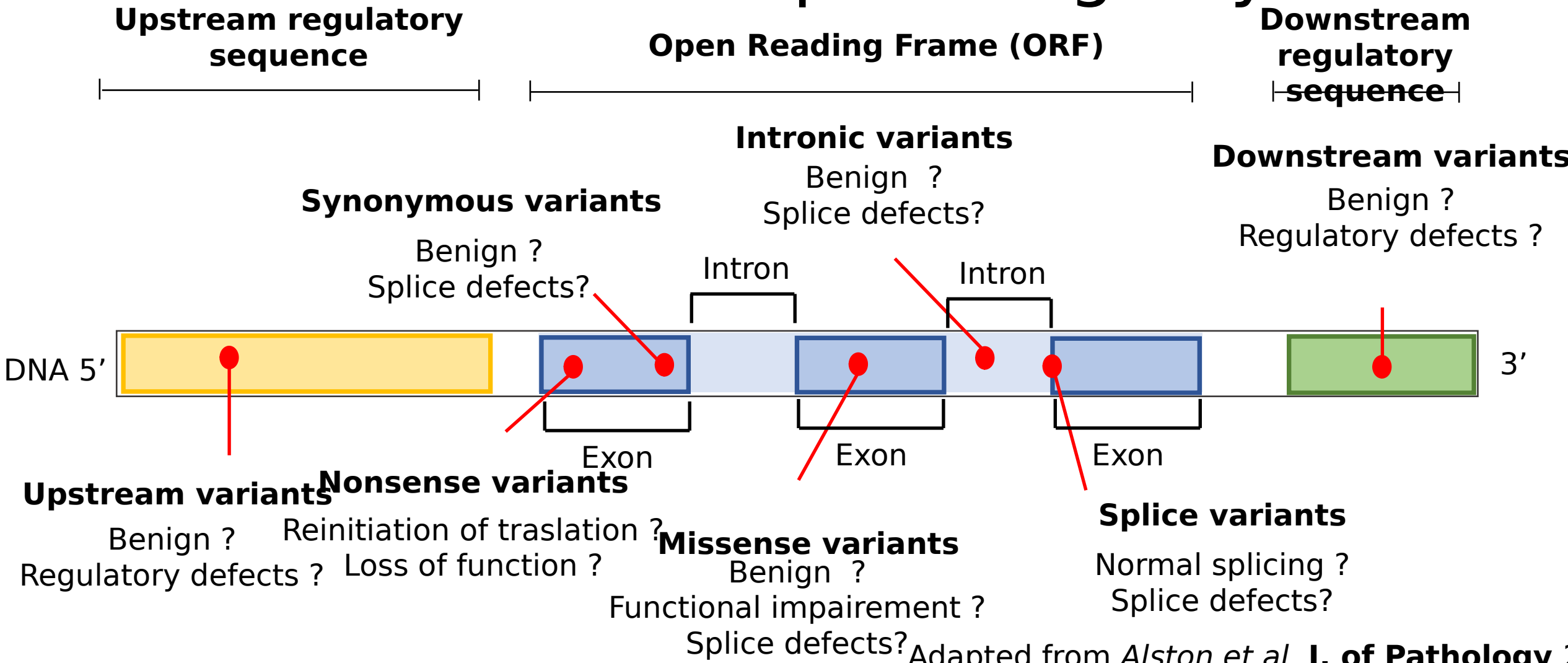
Methods of gene disease discovery in MD





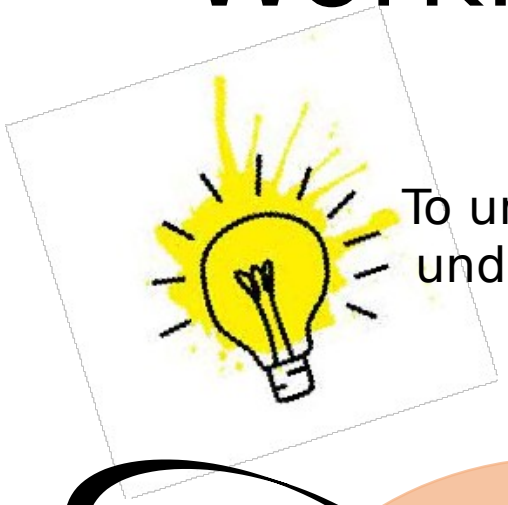
From *Stenton and Prokisch* **EBioMedicine** 2

Variant of uncertain significance: “Innocent until proven guilty”



Adapted from *Alston et al*. **J. of Pathology** 2

Working hypothesis



To understand variants meaning by understanding what they cause to genes bearing them

To identify variants

WES/WGS

Sever al VUS

Candidate genes

Identify gene with aberrant expression

RNA-seq

To improve MD diagnosis

RNA-seq to improve MD diagnosis

Bioinformatics, 2022, 1–8
https://doi.org/10.1093/bioinformatics/btac603
Advance Access Publication Date: 5 September 2022
Original Paper

OXFORD

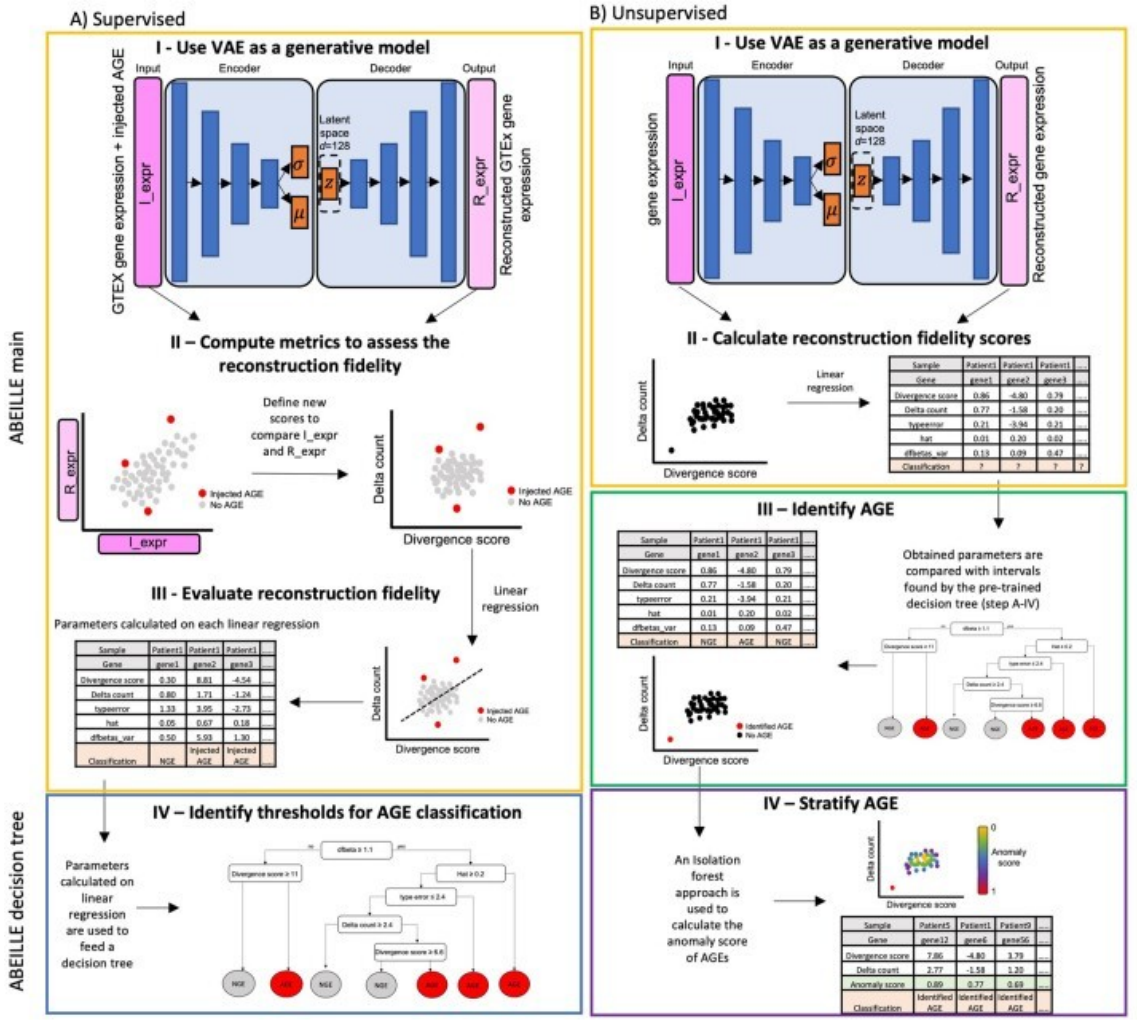
Gene expression

ABEILLE: a novel method for ABerrant Expression Identification empLoying machine LEarning from RNA-sequencing data

Justine Labory[1,2,†], Gwendal Le Bideau[2,†], David Pratella[1], Jean-Elisée Yao[1], Samira Ait-El-Mkadem Saadi[2], Sylvie Bannwarth[2], Loubna El-Hami[1,2], Véronique Paquis-Fluckinger[2,‡] and Silvia Bottini [1,*,‡]

ABEILLE

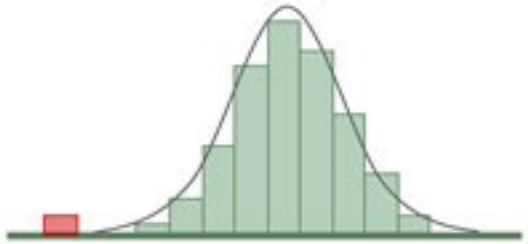
https://github.com/UCA-MSI/ABEILLE



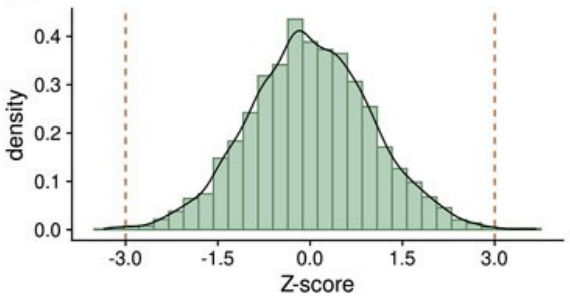
Labory et al. *Bioinformatics* 20

Methods to identify Aberrant Gene Expression (AGE): pros and cons

Statistical methods

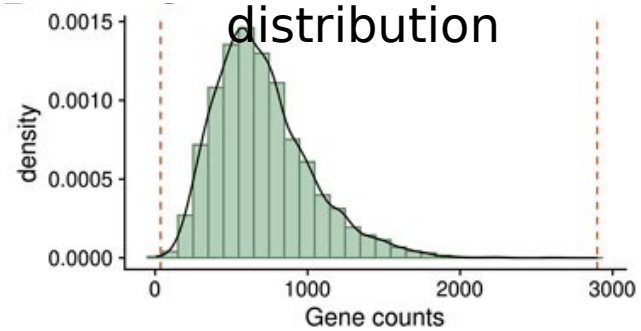


Normal distribution



SVA (Frésard et al. 2019)

Negative binomial distribution



DESeq2 (Love et al. 2014; Kremer et al. 2017)
OUTRIDER (Brechtmann et al. 2018)

No replicates

No control group

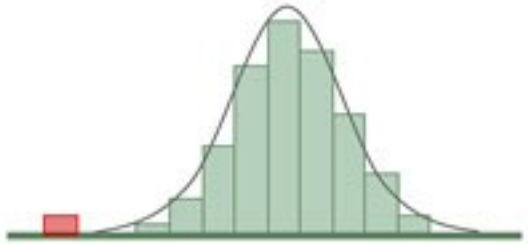
Small cohorts

No common patterns

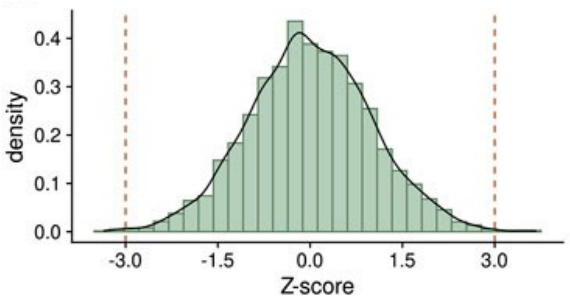
Problem

Methods to identify AGE: pros and cons

Statistical methods

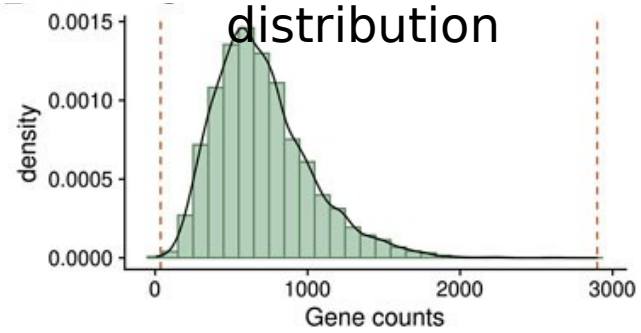


Normal distribution



SVA (Frésard et al. 2019)

Negative binomial distribution

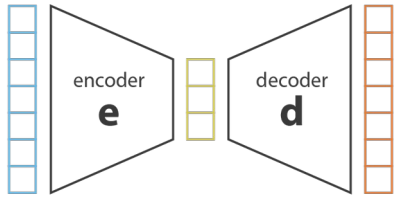


DESeq2 (Love et al. 2014; Kremer et al. 2017)
OUTRIDER (Brechtmann et al. 2018)

Solution

Machine learning methods

Autoencoders

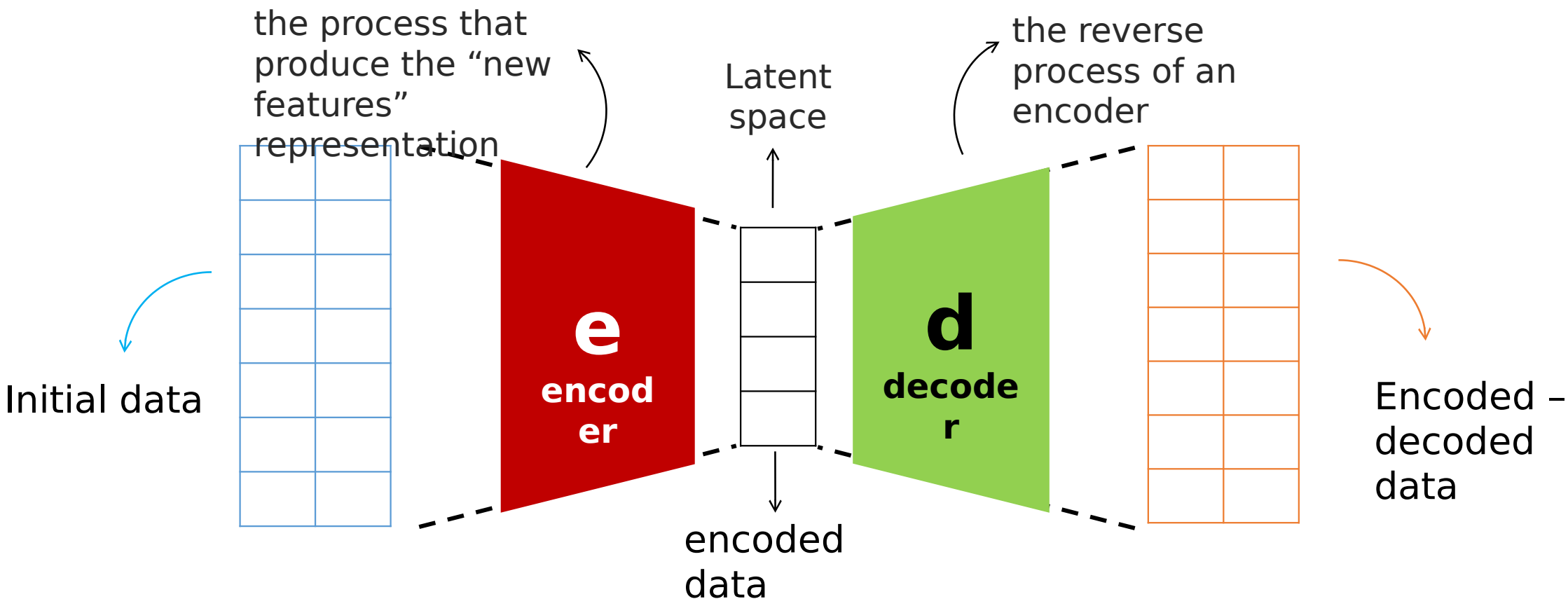


ABEILLE

ABEILLE

Labory et al. *Bioinformatics* 20

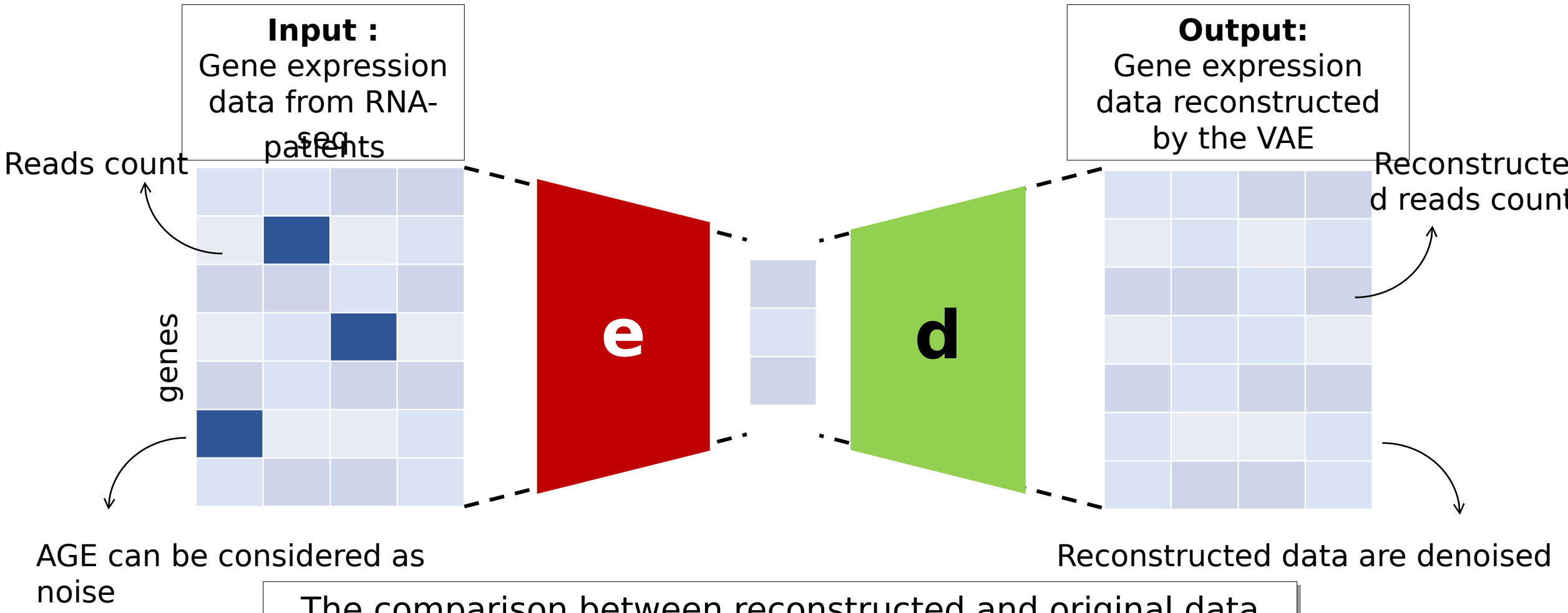
The autoencoder



Dimensionality Reduction | Denoising | Sequence to sequence prediction

Compression | Feature Extraction | Generation | Recommendation system

How to use AE to identify AGEs

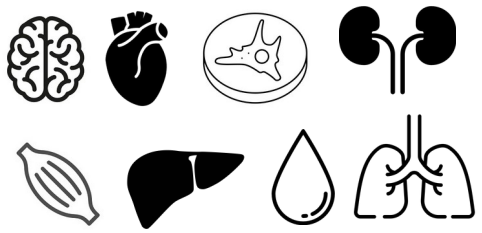


The comparison between reconstructed and original data yields the indentification of AGEs.

Supervised phase – Creation of semi-synthetics datasets

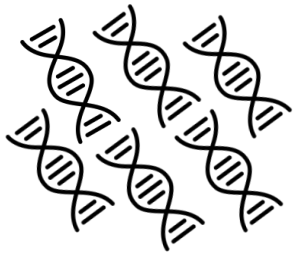
GTEx

54 tissues



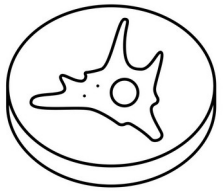
1000 individuals

56 200 transcripts



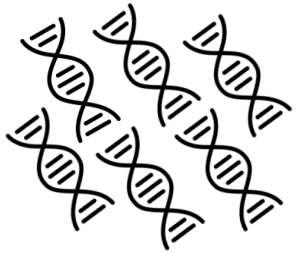
Supervised phase – Creation of semi-synthetics datasets

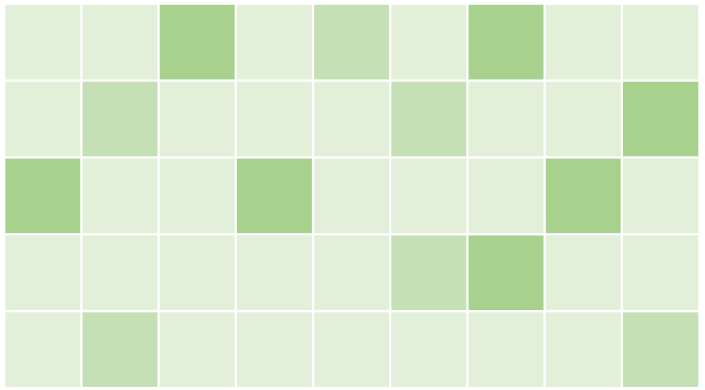
1 tissue



504 individuals

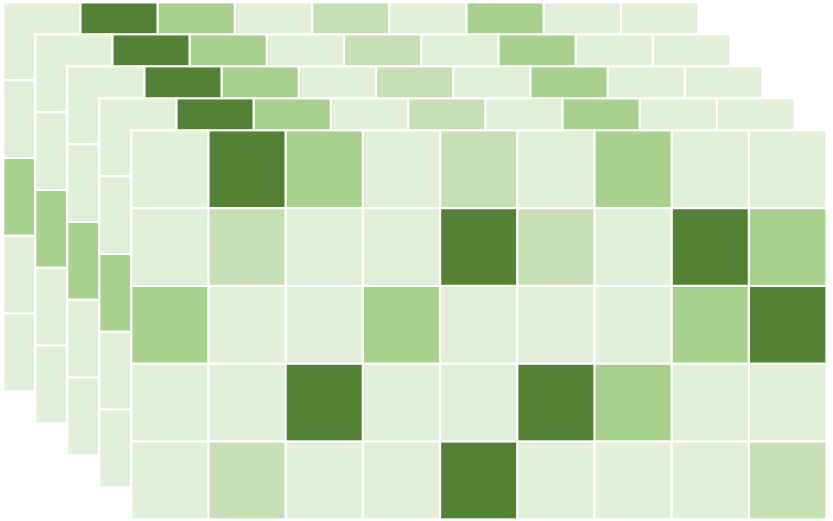
56 200 transcripts





Generate computational AGEs by replacing randomly 10 000

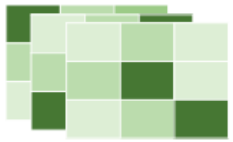
$$k_{ij}^{O} = \text{round}\left(s_i 2^{\mu_j^u \pm \exp(N)\sigma_j^u}\right)$$

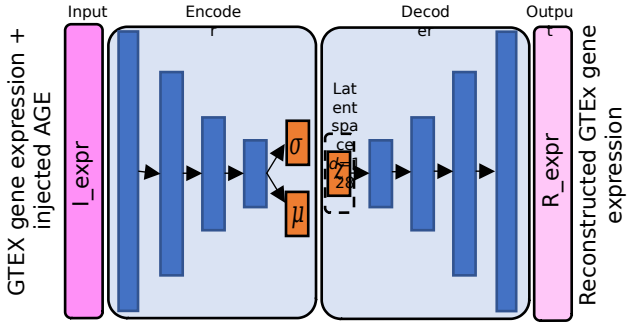


Repeat the process 20 times

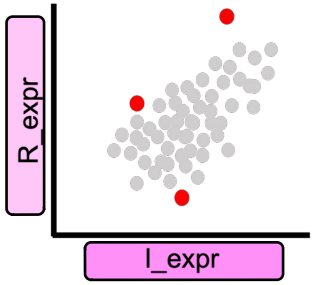
Supervised phase – Train the decision tree

I - Use VAE as a generative model

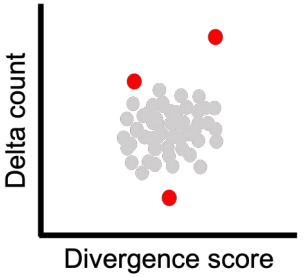




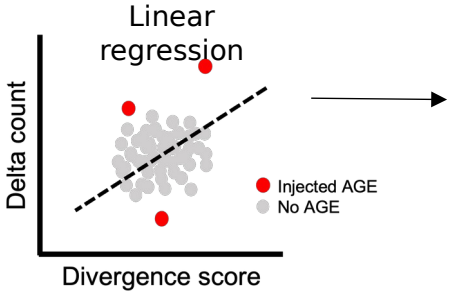
II – Compute metrics to assess the reconstruction fidelity



Define new scores to compare I_expr and R_expr



III - Evaluate reconstruction fidelity

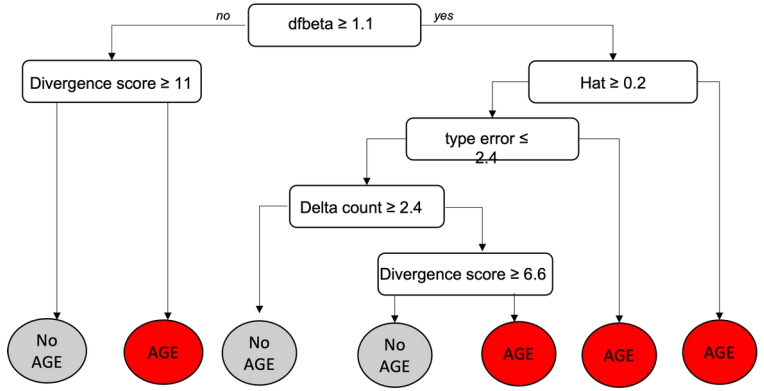


Parameters calculated on each linear regression

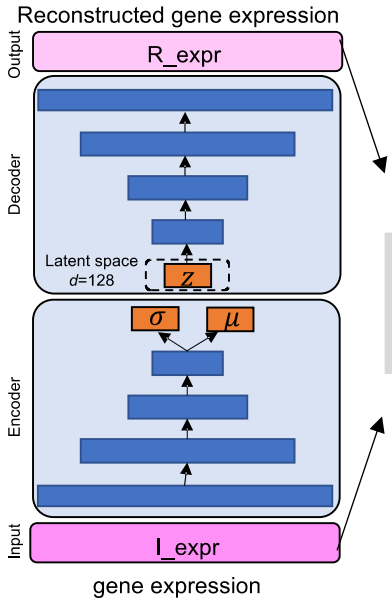
Sample	Patient1	Patient1	Patient1	.....
Transcript	gene1	gene2	gene3	.....
divergence_score	0.30	8.81	-4.54	.....
delta_count	0.80	1.71	-1.24	.....
typeerror	1.33	3.95	-2.73	.....
hat	0.05	0.67	0.18	.....
dfbetas_var	0.50	5.93	1.30	.....
Classification	No AGE	Injected AGE	Injected AGE	.....

IV – Identify thresholds for AGE classification

Parameters calculated on linear regression are used to feed a decision tree



Unsupervised phase - ABEILLE framework



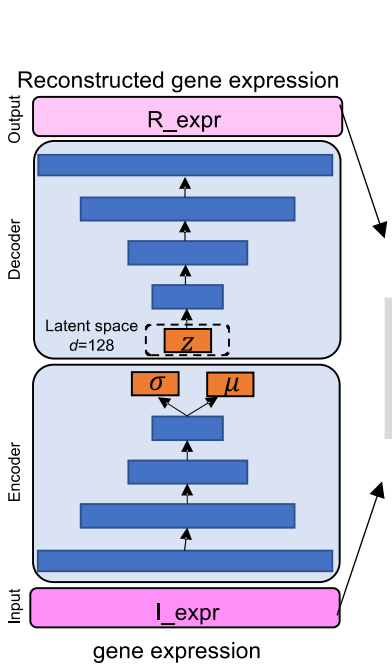
Compute the 2 novel quantities

Linear regression

AGE detection

AGE classification

Unsupervised phase - ABEILLE framework



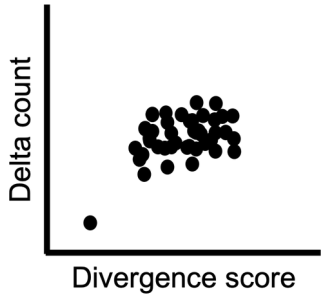
Divergence score
Delta Count

Compute the 2 novel quantities

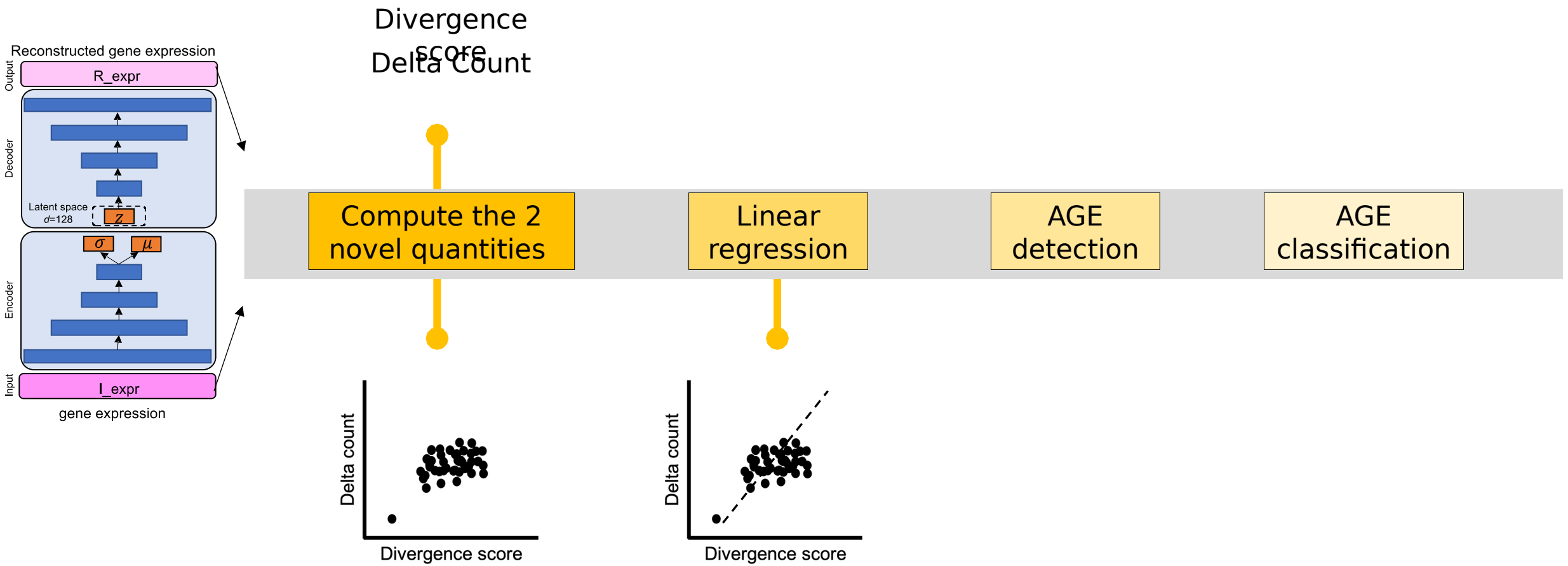
Linear regression

AGE detection

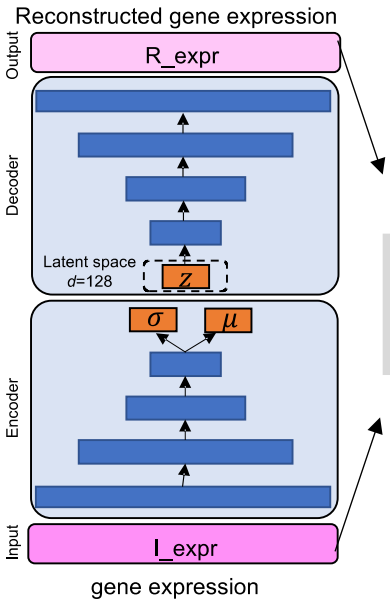
AGE classification



Unsupervised phase - ABEILLE framework

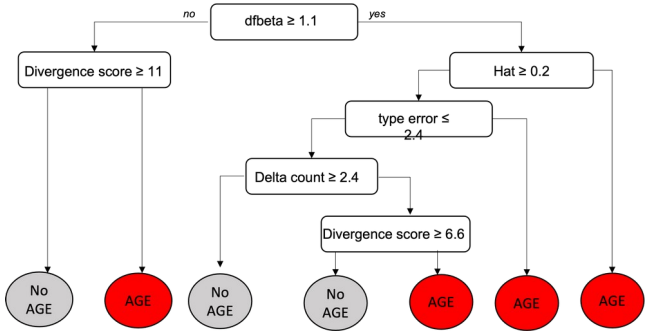


Unsupervised phase - ABEILLE framework



Divergence score

Delta Count

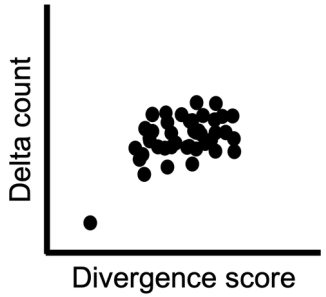


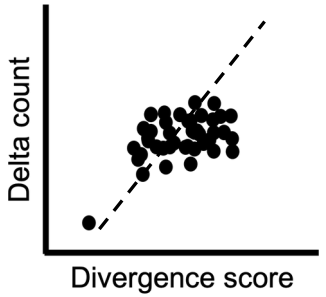
Compute the 2 novel quantities

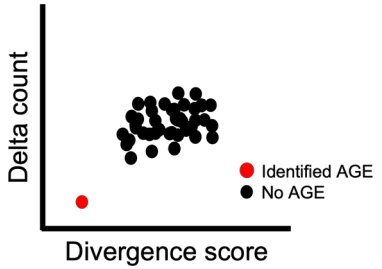
Linear regression

AGE detection

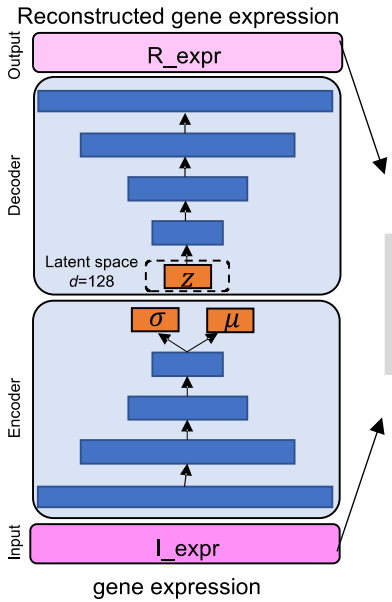
AGE classification





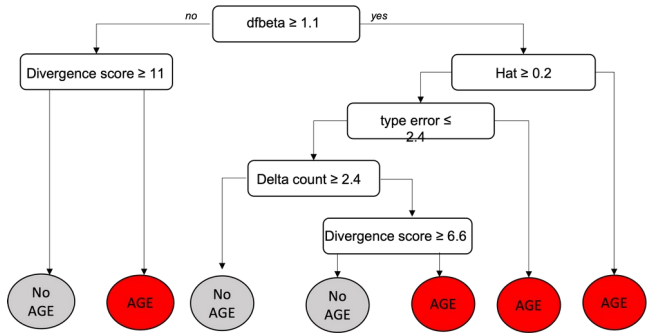


Unsupervised phase - ABEILLE framework



Divergence score

Delta Count



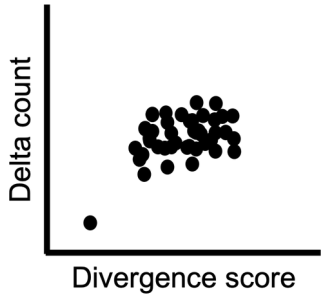
Anomaly Score

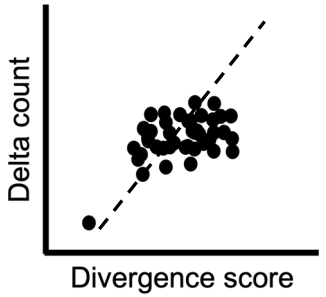
Compute the 2 novel quantities

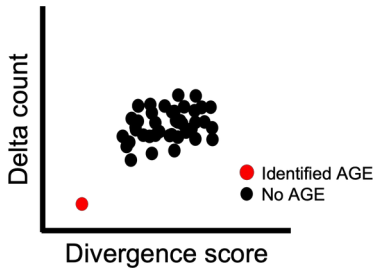
Linear regression

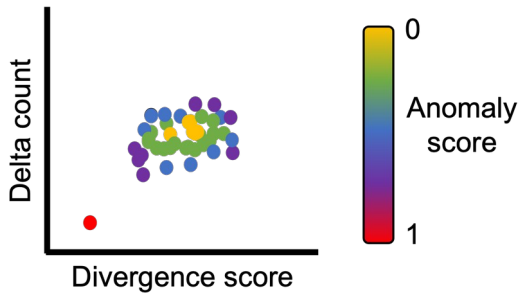
AGE detection

AGE classification



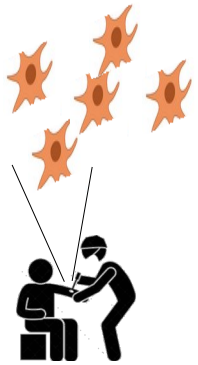




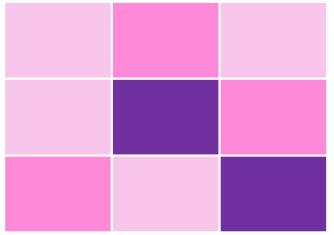


Case study

119 patients with MD suspicion
(from Kremer et al. *Nat Comm* 2017)



RNA-seq

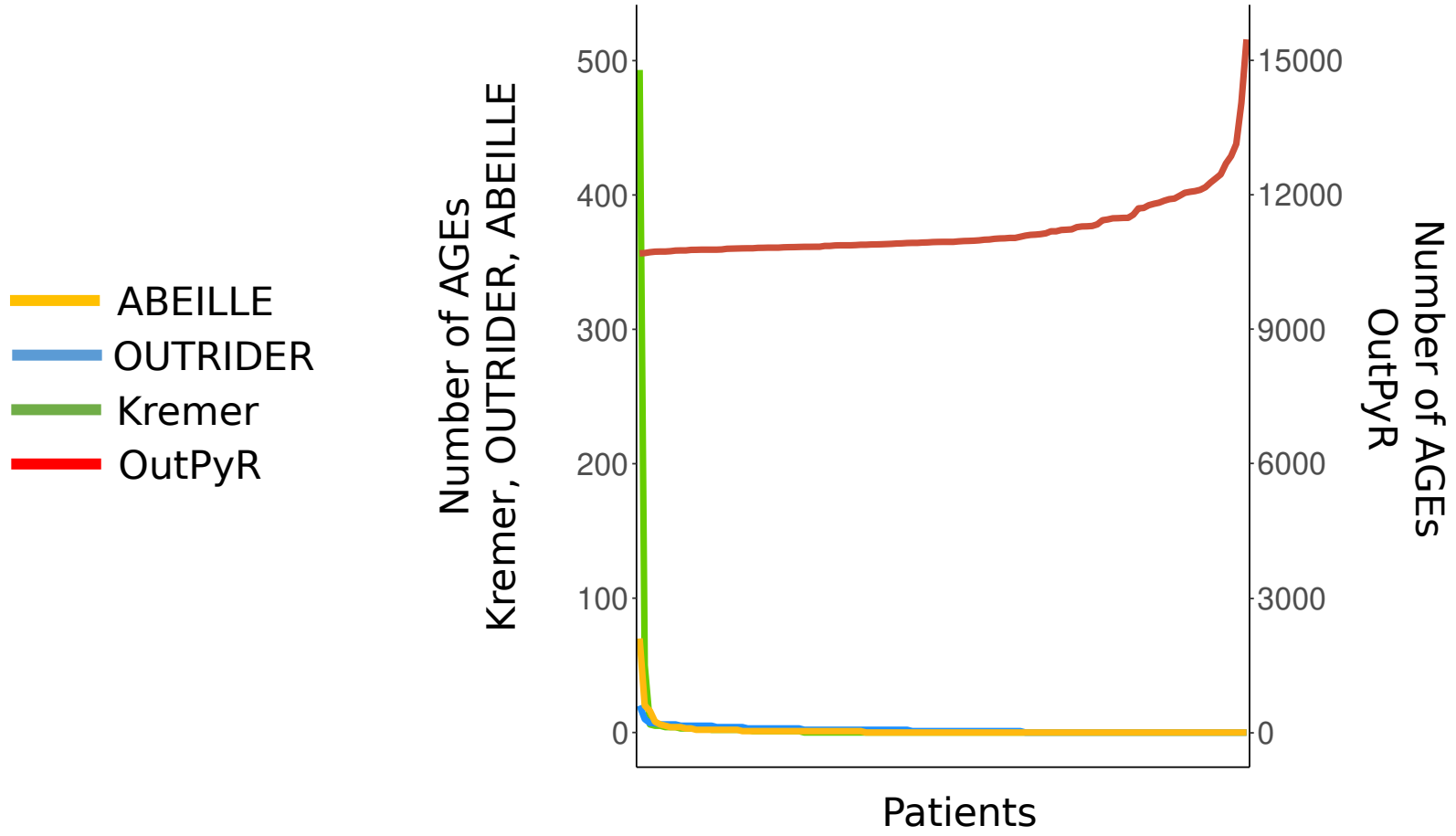


Validation of 5 candidate genes in 6 patients

Goal : Compare ABEILLE to other methods

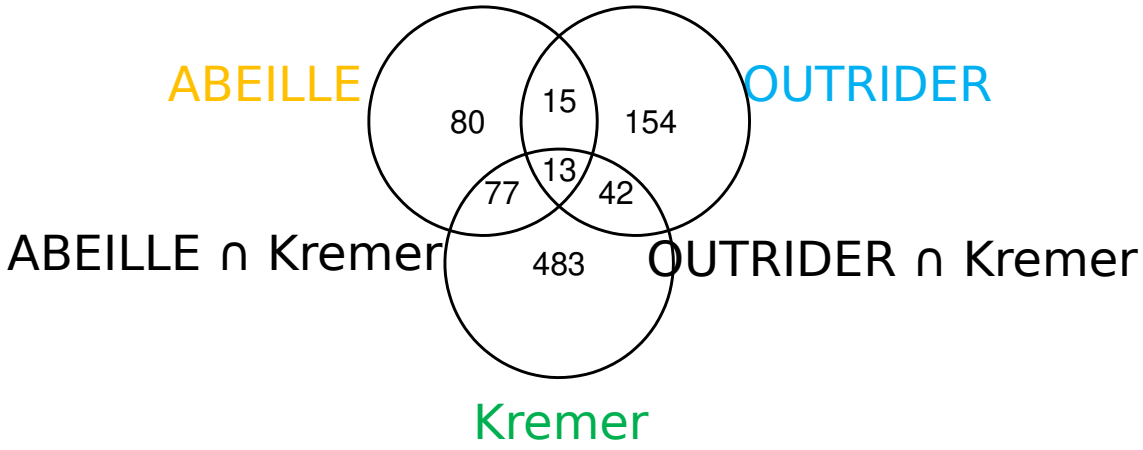
ABEILLE
OUTRID
ERC
OutPyR
Kremer

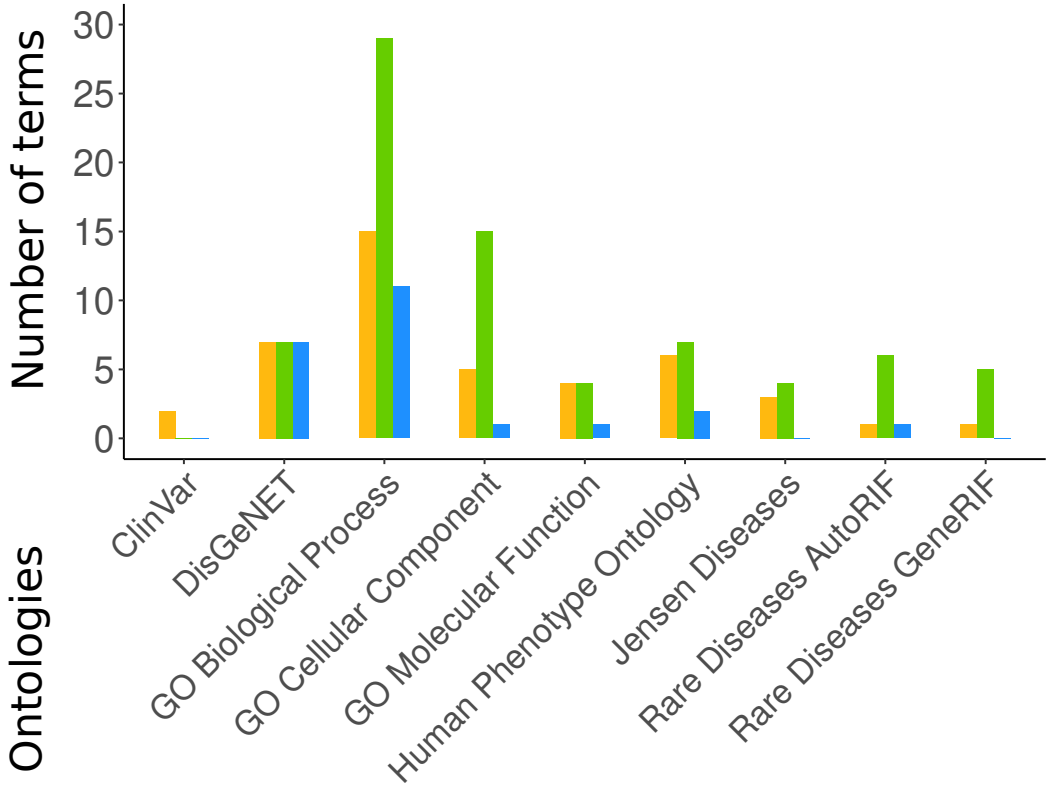
Performances of the four tools on real dataset



These observations rule out OutPyR as a tool for AGE identification in this context.

Performances of ABEILLE and OUTRIDER

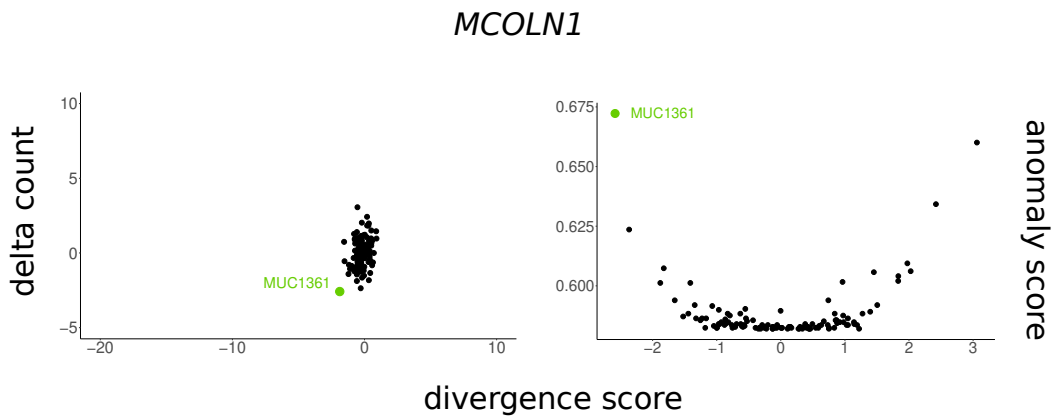


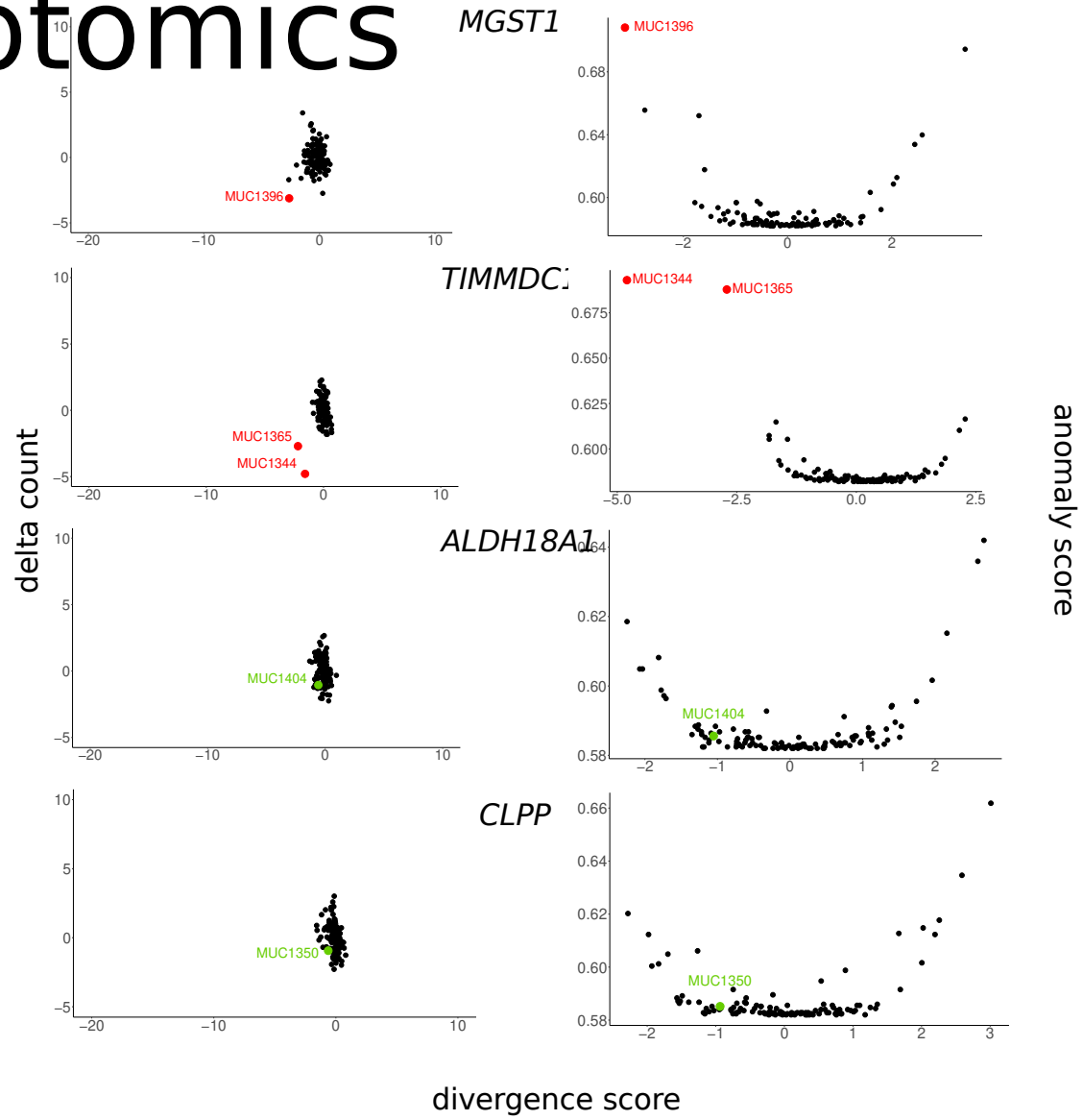


AGEs found by ABEILLE are more enriched in terms related to mitochondrial biology than the AGEs found by OUTRIDER.

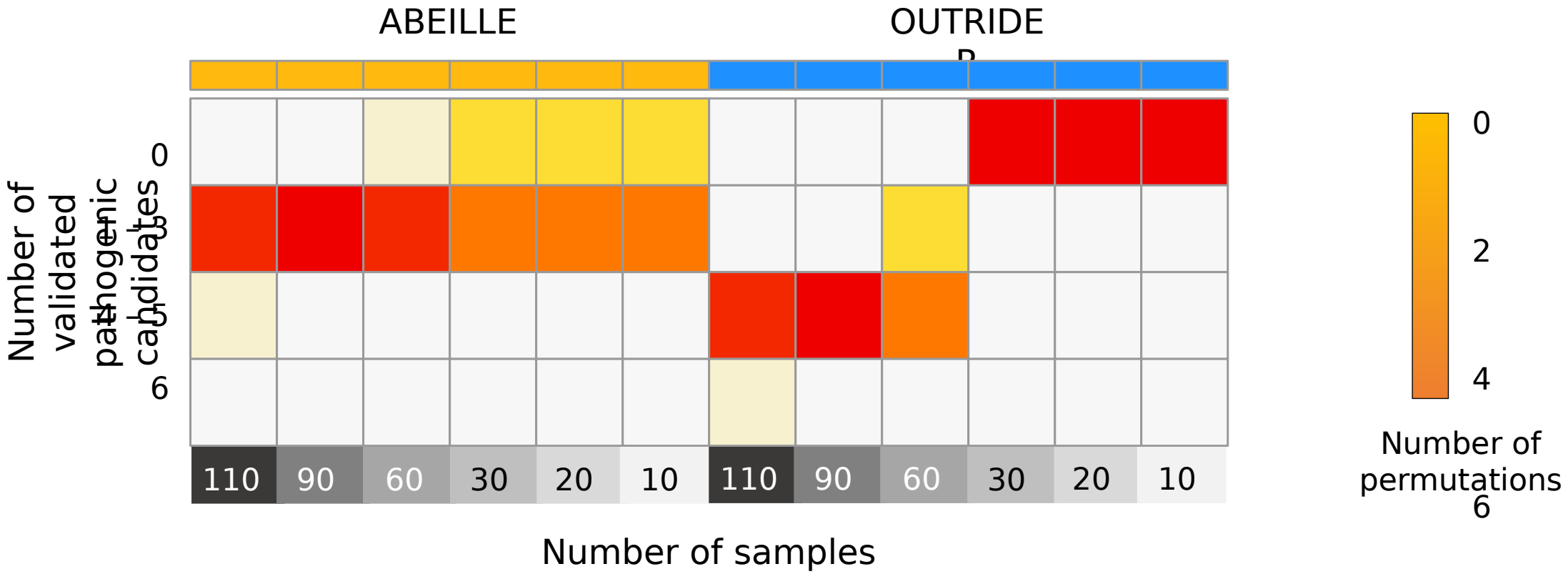
Integration of genomics and transcriptomics

Validated pathogenic genes	Detected by	ABEILLE	OUTRIDER
MGST1	AGE	✓	✓
TIMMDC1	AGE	✓	✓
ALDH18A1	MAE	✗	✓
CLPP	AS	✗	✓
MCOLN1	AGE	✓	✓



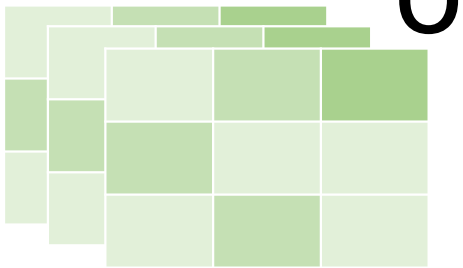


AGE detection on small dataset size



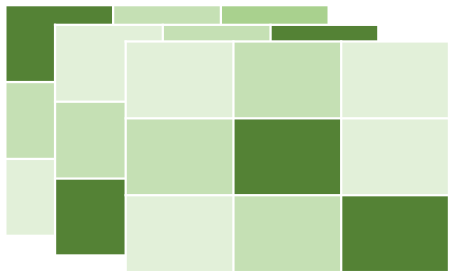
The performances of ABEILLE do not depend on the number of samples

Performances of ABEILLE and OUTRIDER on semi-synthetic datasets

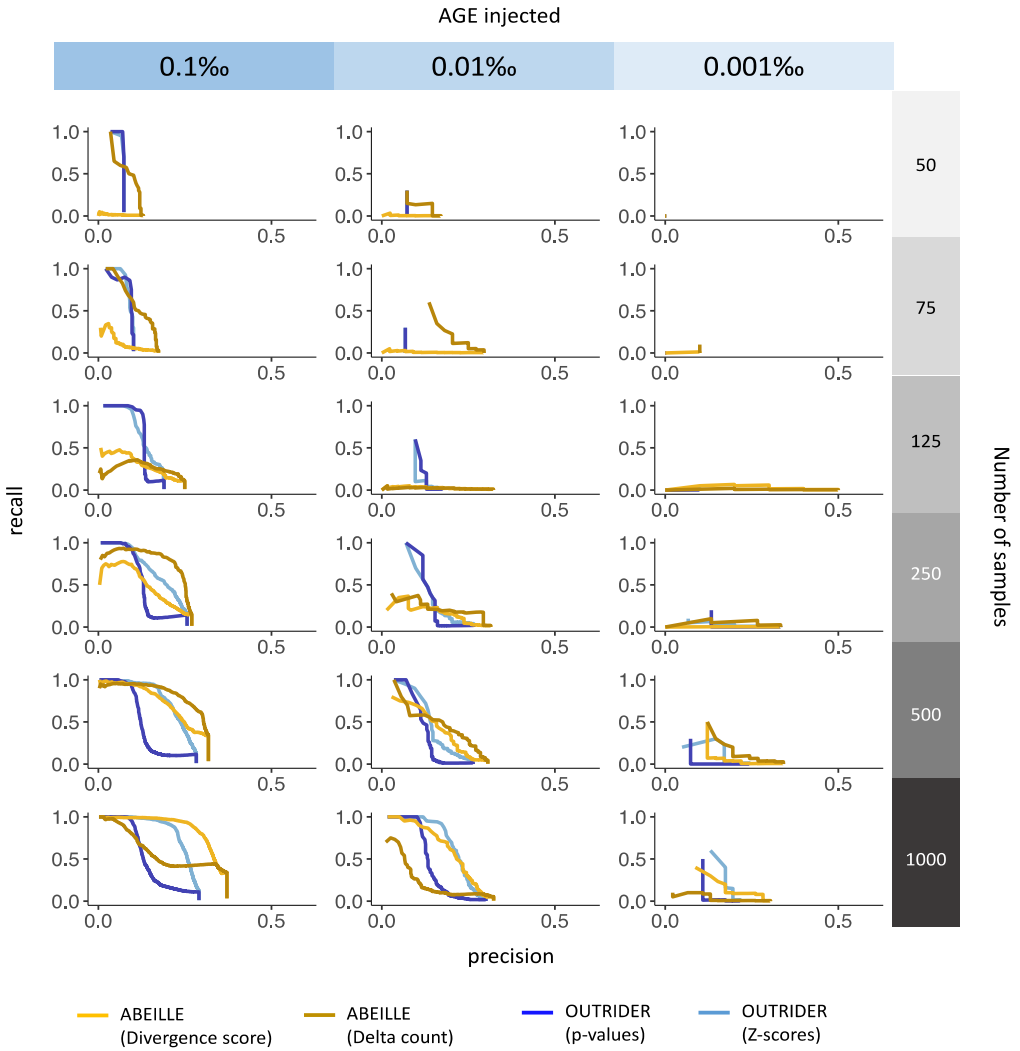


Datasets with no AGE

Computational AGE generation



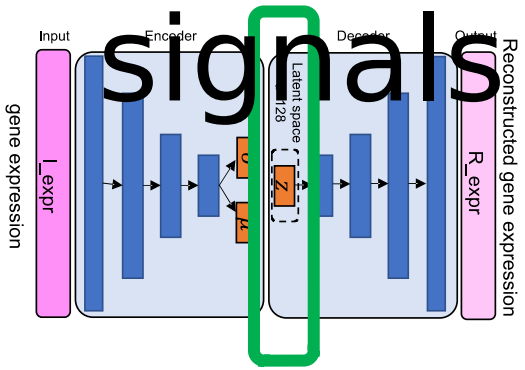
Datasets with synthetic AGE

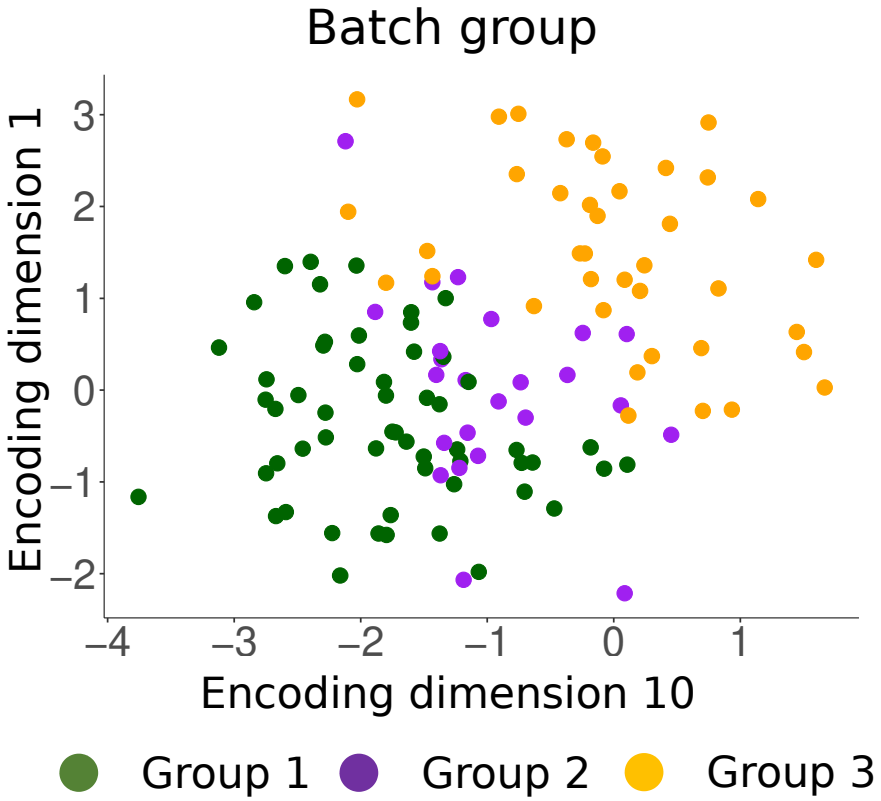


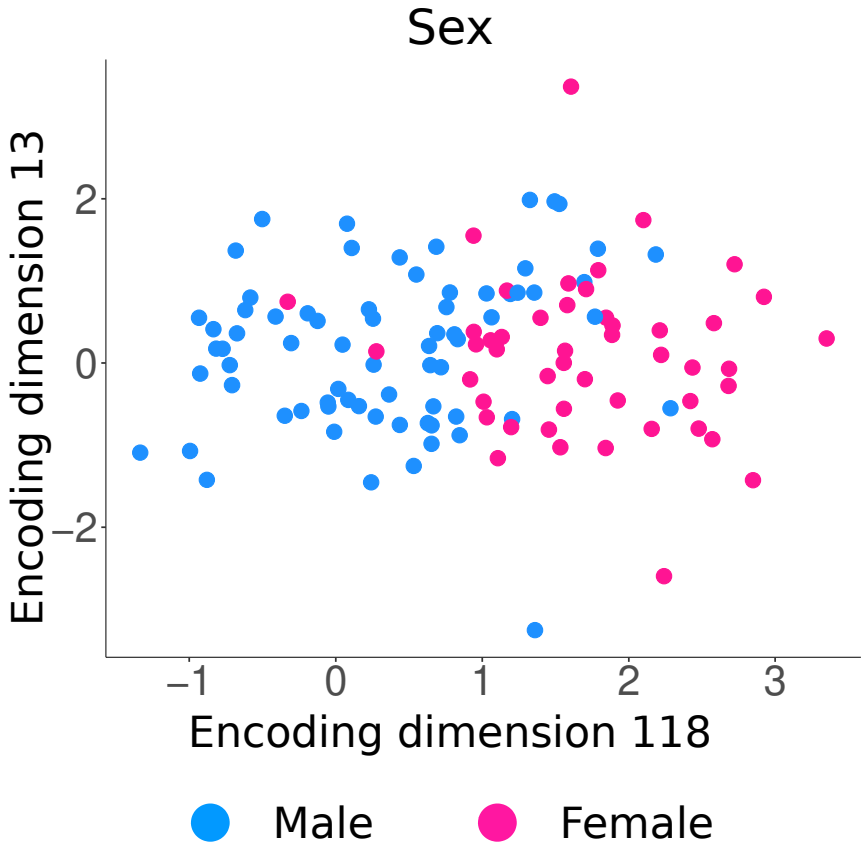
- On datasets with 0.1‰ AGEs injected, ABEILLE ranking by Delta Count showed the higher performances than the ranking by Divergence Score
- When the percentage of injected AGEs diminish, the ranking by Divergence Score yielded better results for ABEILLE.
- OUTRIDER ranking by p-values are slightly better than by Z-score

 The performances of the tools depend on the score used to rank the AGEs

ABEILLE VAE features captures biological signals







Conclusion

ADVANTAGES

- ABEILLE identifies AGEs from RNA-seq data without the need of replicates and without assumption on the distribution
- ABEILLE showed good performances on small datasets and datasets with few AGEs

DRAWBACKS

- The decision tree must be trained for each different omic
- ABEILLE doesn’t use a flexible model to do multi-omics integration and analysis

Thank you!

https://github.com/UCA-MSI/

Contact: **justine.labory@etu.univ-**

Operational director : Silvia BOTTINI

Engineers:
Djampa Kozlowski
Marco Milanesio

Master students:
Marlize de Villeris
Evariste Njomgue
Mame Seynabou Fall
Gauthier Marcovich

Former members/students:
Fanny Simoes
David Pratella
Gwendal Le Bideau
Morgane Fiervielle
Loubna El-Hami
Paola Porracciolo