Troisième journée NETBIO 2021

eQTLs are key players in the integration of genomic and transcriptomic data for phenotype prediction

Abdou Rahmane WADE

BioForA

21 juin 2021

INRAE

Curriculum

- MSc in Genetics and Plant Breeding

- 3^{rd} year PhD student

- Thesis : Improved genome-based phenotypic predictions with a systems biology approach

- Supervisers : Leopoldo Sanchez Rodriguez and Vincent Segura

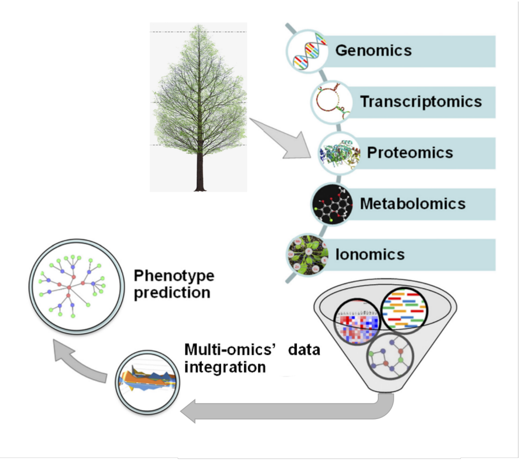
Table of contents

1. **Interest of the muti-omic integration for the prediction**
   - Genomic & Transcriptomic integration
   - Ways of integrating omics data
2. **Research question**
3. **Materials**
   - Phenotypic data
   - Genomic and Transcriptomic data
4. **Methods & Results**
   - Prediction models
   - Prediction results
   - eQTL analyses
   - Behavior analysis of the different variables typologies effects
5. **Conclusions**

Interest of the muti-omic integration for the prediction

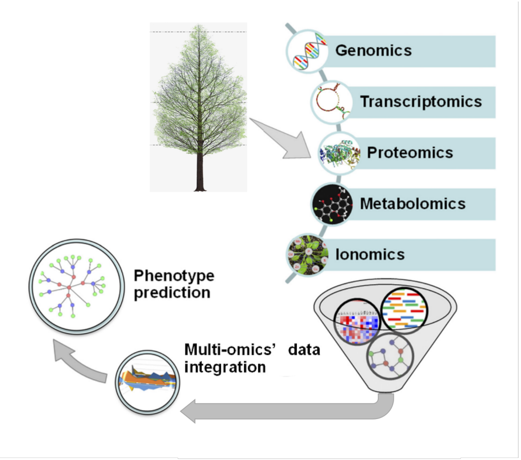
Interest in Multi-omics integration

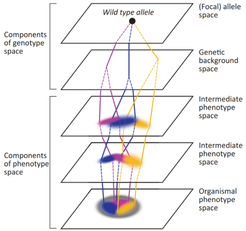
Interest in Multi-omics integration



Modified from Mishra et al. 2018

Interest in Multi-omics integration



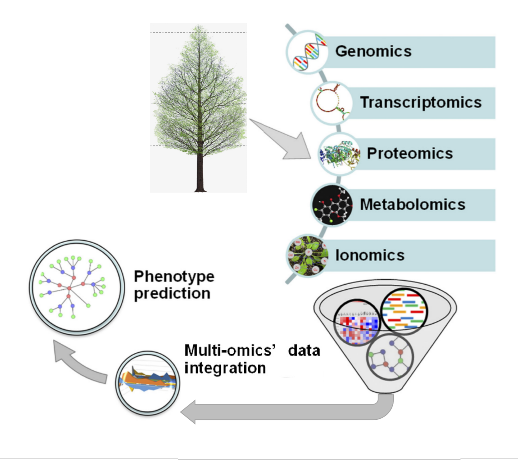


Wild type allele
(Focal) allele space
Components of genotype space
Genetic background space
Intermediate phenotype space
Intermediate phenotype space
Components of phenotype space
Organismal phenotype space

(Chandler et al. Trends in Genetics, 2013)

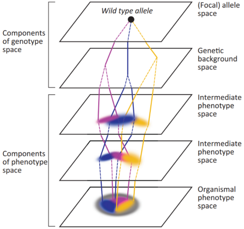
Modified from Mishra et al. 2018

Interest in Multi-omics integration



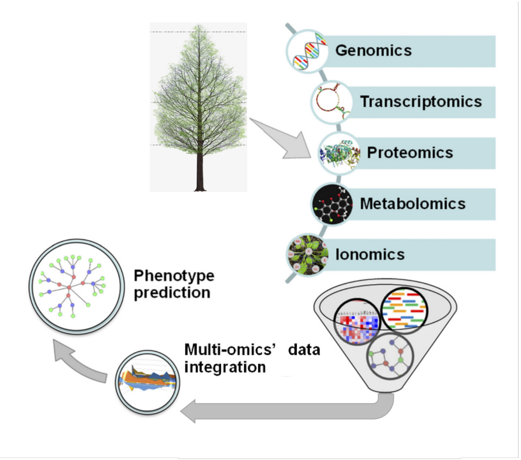
Modified from Mishra et al. 2018

The case of the missing heritability
(Maher, Nature 2008)


Wild type allele
(Focal) allele space
Components of genotype space
Genetic background space
Intermediate phenotype space
Components of phenotype space
Intermediate phenotype space
Organismal phenotype space

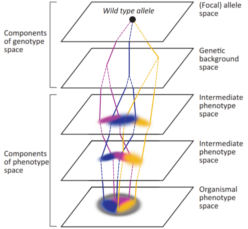
(Chandler et al. Trends in Genetics, 2013)

Interest in Multi-omics integration



Modified from Mishra et al. 2018

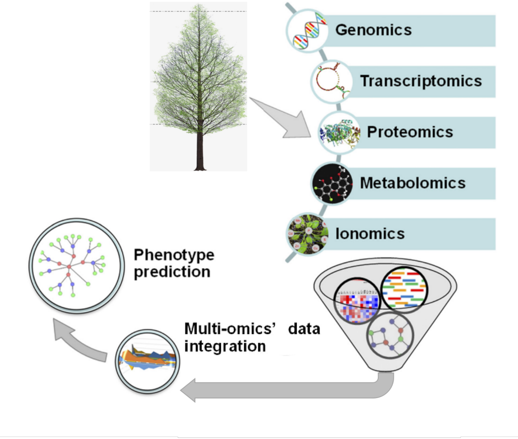
The case of the missing heritability
(Maher, Nature 2008)


Wild type allele
(Focal) allele space
Components of genotype space
Genetic background space
Intermediate phenotype space
Intermediate phenotype space
Components of phenotype space
Organismal phenotype space

(Chandler et al. Trends in Genetics, 2013)

Understanding and predicting complex traits

Interest in Multi-omics integration



Modified from Mishra et al. 2018

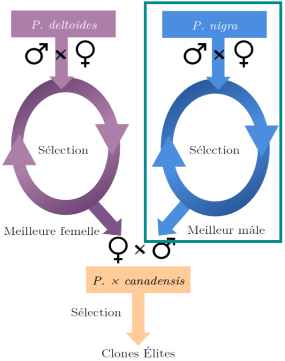
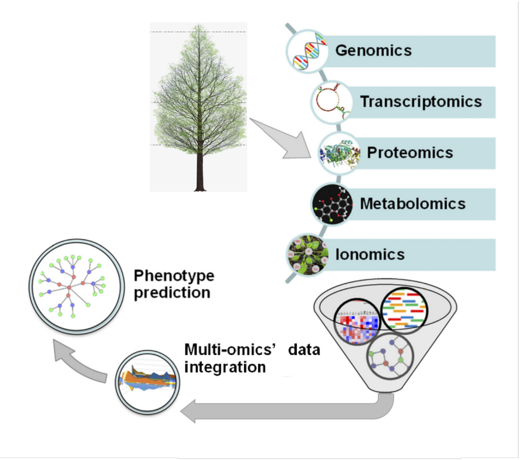


Schéma de sélection de *P. x canadensis*

Interest in Multi-omics integration



Genomics

Transcriptomics

Proteomics

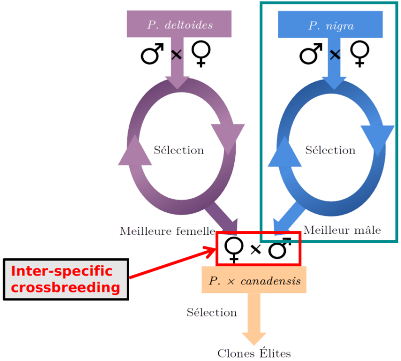
Metabolomics

Ionomics

Phenotype prediction

Multi-omics' data integration

Modified from Mishra et al. 2018



P. deltoides

♂ × ♀

Sélection

P. nigra

♂ × ♀

Sélection

Meilleure femelle

Meilleur mâle

♀ × ♂

Inter-specific crossbreeding

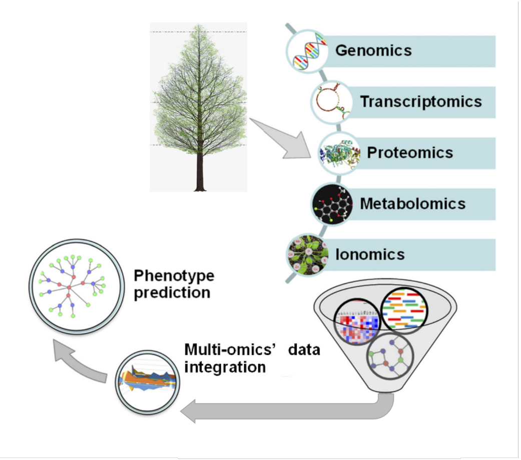
P. × canadensis

Sélection

Clones Élites

Schéma de sélection de *P. x canadensis*

Interest in Multi-omics integration



Modified from Mishra et al. 2018

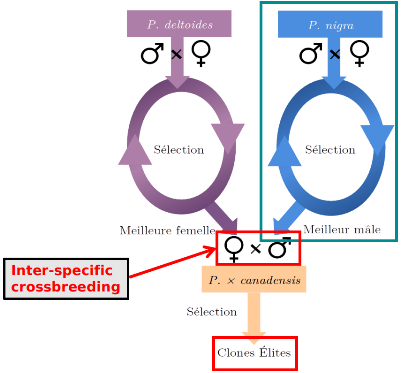
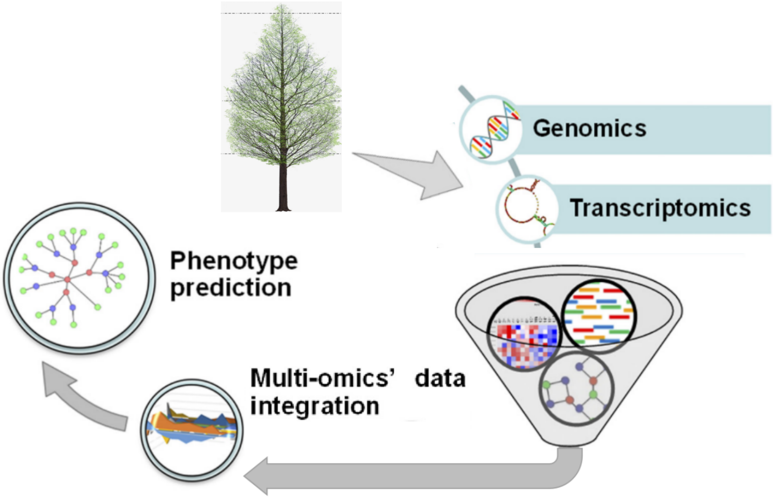


Schéma de sélection de *P. x canadensis*

Interest in Multi-omics integration



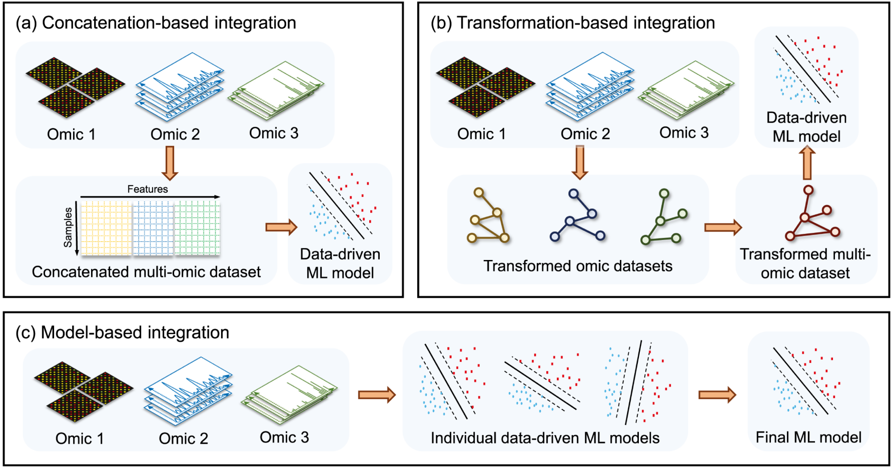
Genomics

Transcriptomics

Phenotype prediction

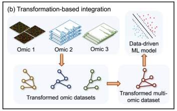
Multi-omics' data integration

Ways of integrating omics data

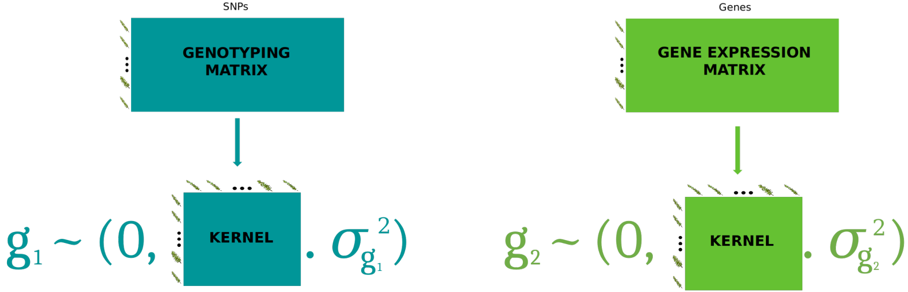


(a) Concatenation-based integration
Omic 1
Omic 2
Omic 3
Features
Samples
Concatenated multi-omic dataset
Data-driven ML model
(b) Transformation-based integration
Omic 1
Omic 2
Omic 3
Data-driven ML model
Transformed omic datasets
Transformed multi-omic dataset
(c) Model-based integration
Omic 1
Omic 2
Omic 3
Individual data-driven ML models
Final ML model

from Zampieri et al 2019


(b) Transformation-based integration
Omic 1
Omic 2
Omic 3
Data-driven ML model
Transformed omic datasets
Transformed multi-omic dataset

Guo et al. 2016 , Weshthues et al. 2017 , Schrag et al. 2018 & Azodi et al. 2020 ; Li et al. 2019 & Morgante et al. 2020



SNPs

GENOTYPING MATRIX

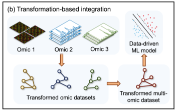
Genes

GENE EXPRESSION MATRIX

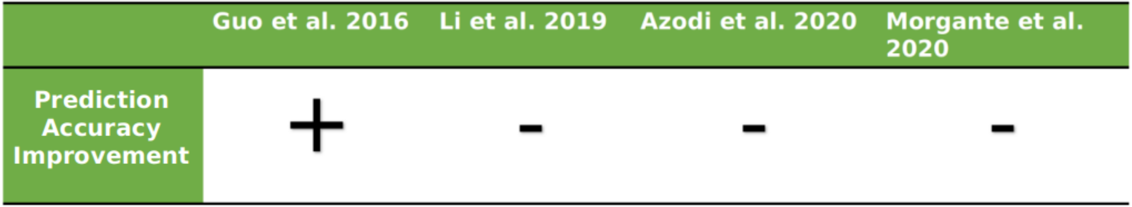
$$g_1 \sim (0, \boxed{\text{KERNEL}} \cdot \sigma_{g_1}^{2})$$

$$g_2 \sim (0, \boxed{\text{KERNEL}} \cdot \sigma_{g_2}^{2})$$

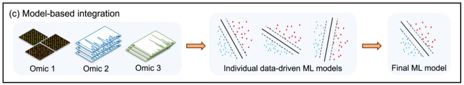
$$y = g_1 c_1 + g_2 c_1 + e$$


(b) Transformation-based integration
Omic 1
Omic 2
Omic 3
Data-driven ML model
Transformed omic datasets
Transformed multi-omic dataset

Guo et al. 2016 , Weshthues et al. 2017 , Schrag et al. 2018 & Azodi et al. 2020 ; Li et al. 2019 & Morgante et al. 2020



	Guo et al. 2016	Li et al. 2019	Azodi et al. 2020	Morgante et al. 2020
Prediction Accuracy Improvement	+	-	-	-



(c) Model-based integration

Omic 1

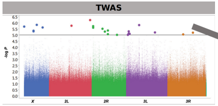
Omic 2

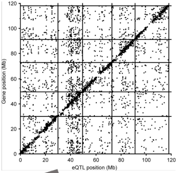
Omic 3

Individual data-driven ML models

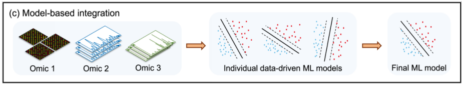
Final ML model

Ye et al. 2020

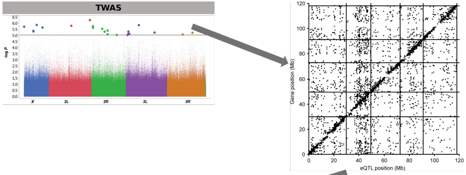




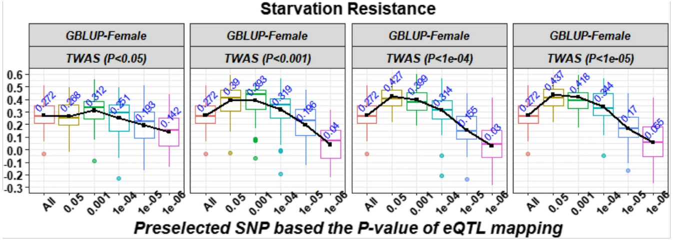
Genomic Prediction (GBLUP)

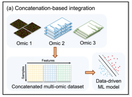


Ye et al. 2020



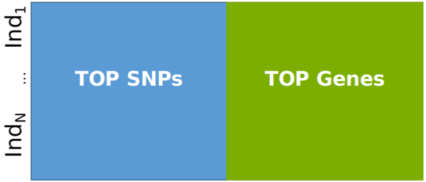
Genomic Prediction (GBLUP)

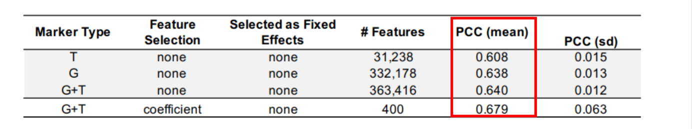



(a) Concatenation-based integration
Omic 1
Omic 2
Omic 3
Features
Samples
Concatenated multi-omic dataset
Data-driven
ML model

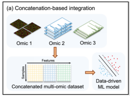
Azodi et al. 2020

Concaténation



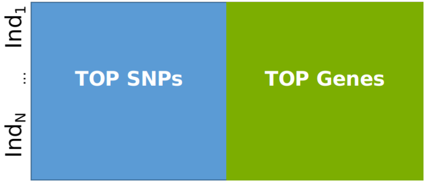


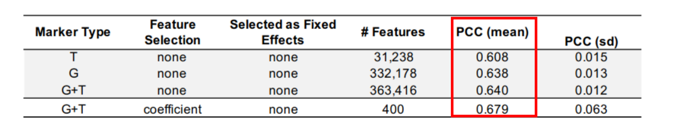
Marker Type	Feature Selection	Selected as Fixed Effects	# Features	PCC (mean)	PCC (sd)
T	none	none	31,238	0.608	0.015
G	none	none	332,178	0.638	0.013
G+T	none	none	363,416	0.640	0.012
G+T	coefficient	none	400	0.679	0.063


(a) Concatenation-based integration
Omic 1
Omic 2
Omic 3
Features
Samples
Concatenated multi-omic dataset
Data-driven ML model

Azodi et al. 2020

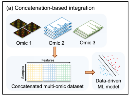
Concaténation





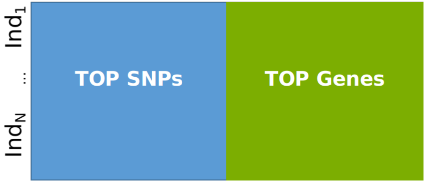
Marker Type	Feature Selection	Selected as Fixed Effects	# Features	PCC (mean)	PCC (sd)
T	none	none	31,238	0.608	0.015
G	none	none	332,178	0.638	0.013
G+T	none	none	363,416	0.640	0.012
G+T	coefficient	none	400	0.679	0.063

1 **Concatenating Top-SNPs and Top-Genes improves prediction accuracy**


(a) Concatenation-based integration
Omic 1
Omic 2
Omic 3
Features
Samples
Concatenated multi-omic dataset
Data-driven ML model

Azodi et al. 2020

Concaténation



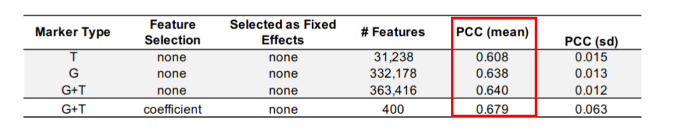
Ind_1

...

Ind_N

TOP SNPs

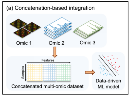
TOP Genes



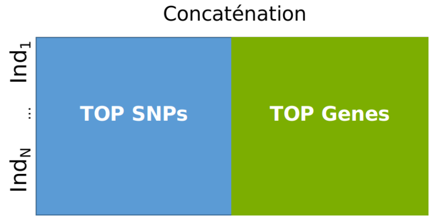
Marker Type	Feature Selection	Selected as Fixed Effects	# Features	PCC (mean)	PCC (sd)
T	none	none	31,238	0.608	0.015
G	none	none	332,178	0.638	0.013
G+T	none	none	363,416	0.640	0.012
G+T	coefficient	none	400	0.679	0.063

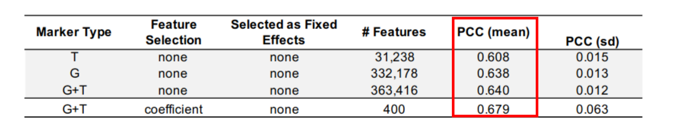
1 **Concatenating Top-SNPs and Top-Genes improves prediction accuracy**

2 **Top-SNPs and Top-Genes are not located in the same gene loci**


(a) Concatenation-based integration
Omic 1
Omic 2
Omic 3
Features
Samples
Concatenated multi-omic dataset
Data-driven ML model

Azodi et al. 2020





Marker Type	Feature Selection	Selected as Fixed Effects	# Features	PCC (mean)	PCC (sd)
T	none	none	31,238	0.608	0.015
G	none	none	332,178	0.638	0.013
G+T	none	none	363,416	0.640	0.012
G+T	coefficient	none	400	0.679	0.063

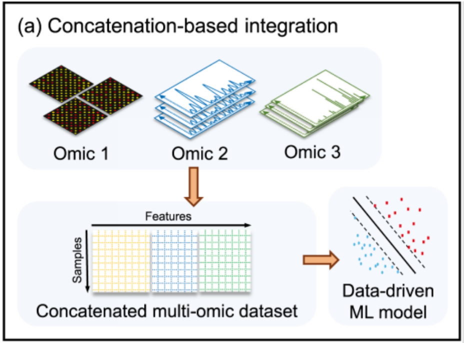
1 **Concatenating Top-SNPs and Top-Genes improves prediction accuracy**

2 **Top-SNPs and Top-Genes are not located in the same gene loci**

3 **Top-SNPs are not eQTLs of Top-genes**

Research question

Research question


(a) Concatenation-based integration
Omic 1
Omic 2
Omic 3
Features
Samples
Concatenated multi-omic dataset
Data-driven
ML model

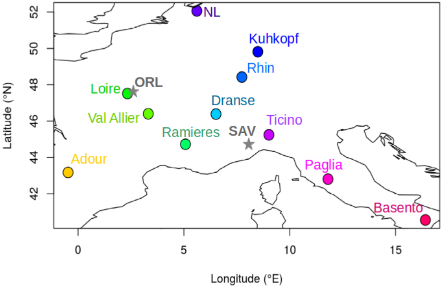
- **How do the different factors, SNPs and Genes, behave during integration?**

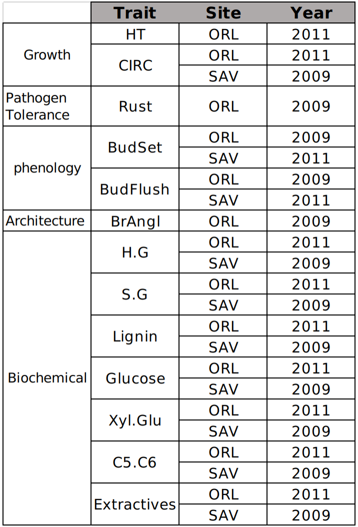
Materials

Phenotypic data

1000 *P. nigra* genotypes from 11 natural populations

- Common garden experiment : 2 sites (Orléans, France / Savigliano, Italy)
- 21 traits

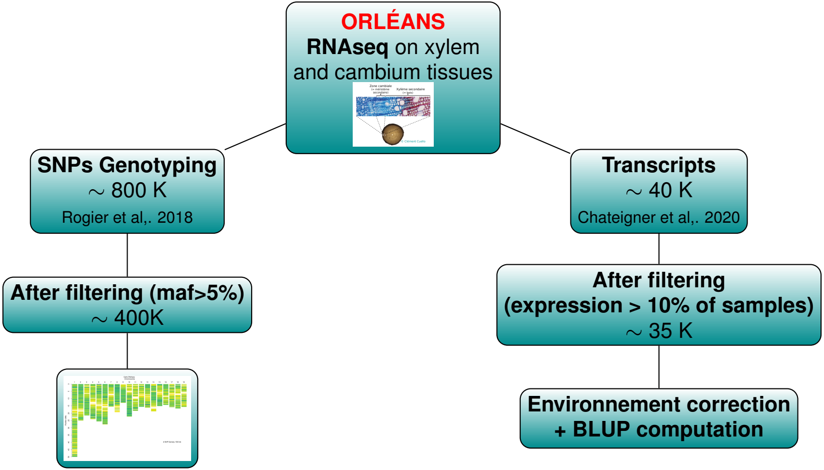




	Trait	Site	Year
Growth	HT	ORL	2011
	CIRC	ORL	2011
		SAV	2009
Pathogen Tolerance	Rust	ORL	2009
phenology	BudSet	ORL	2009
		SAV	2011
	BudFlush	ORL	2009
		SAV	2011
Architecture	BrAngl	ORL	2009
Biochemical	H.G	ORL	2011
		SAV	2009
	S.G	ORL	2011
		SAV	2009
	Lignin	ORL	2011
		SAV	2009
	Glucose	ORL	2011
		SAV	2009
	Xyl.Glu	ORL	2011
		SAV	2009
	C5.C6	ORL	2011
		SAV	2009
	Extractives	ORL	2011
		SAV	2009

Genomic and Transcriptomic data

241 **genotypes representing the genetic diversity of the 1000 phenotyped individuals**



ORLÉANS
RNAseq on xylem and cambium tissues

Zone cambiale (= méristème secondaire)
Xylème secondaire (= bois)
© Clément Cuello

SNPs Genotyping
~ 800 K
Rogier et al,. 2018

After filtering (maf>5%)
~ 400K

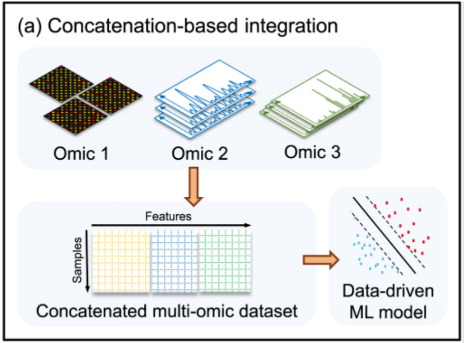
Transcripts
~ 40 K
Chateigner et al,. 2020

After filtering (expression > 10% of samples)
~ 35 K

Environnement correction + BLUP computation

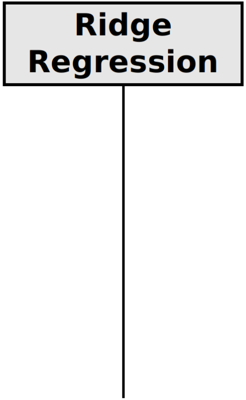
Methods & Results

Research question


(a) Concatenation-based integration
Omic 1
Omic 2
Omic 3
Features
Samples
Concatenated multi-omic dataset
Data-driven ML model

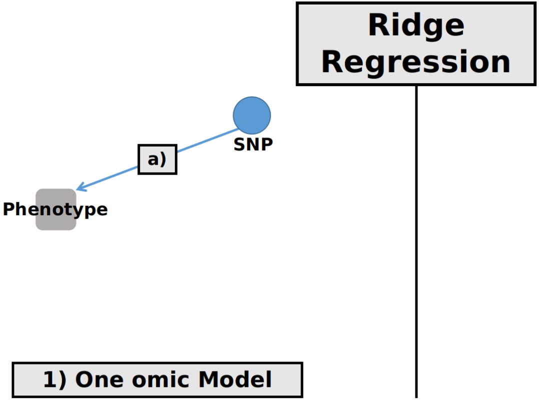
- **How do the different factors, SNPs and Genes, behave during integration?**

Prediction models



Ridge Regression

Prediction models



Ridge Regression

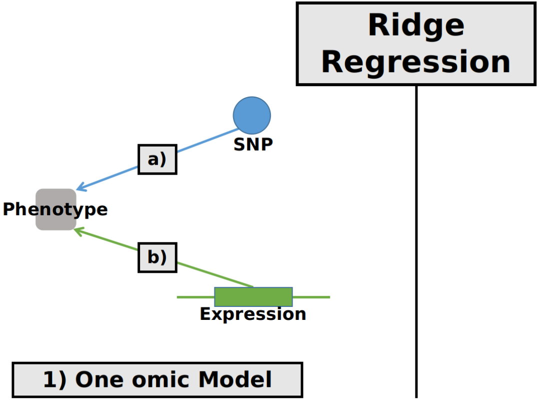
SNP

a)

Phenotype

1) One omic Model

Prediction models



Ridge Regression

SNP

a)

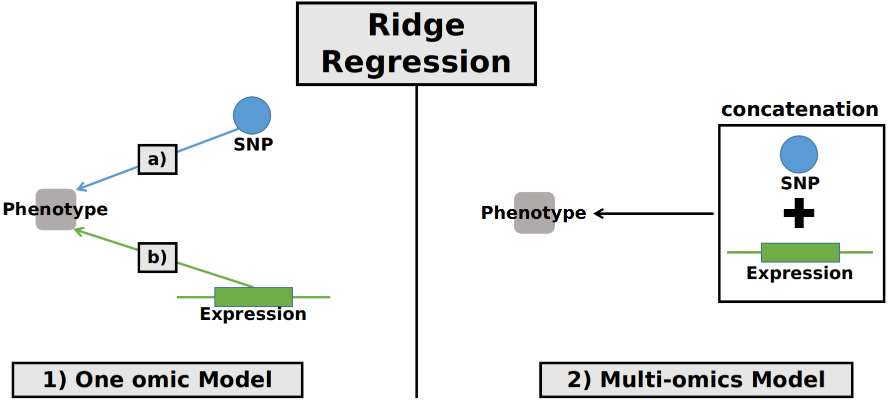
Phenotype

b)

Expression

1) One omic Model

Prediction models

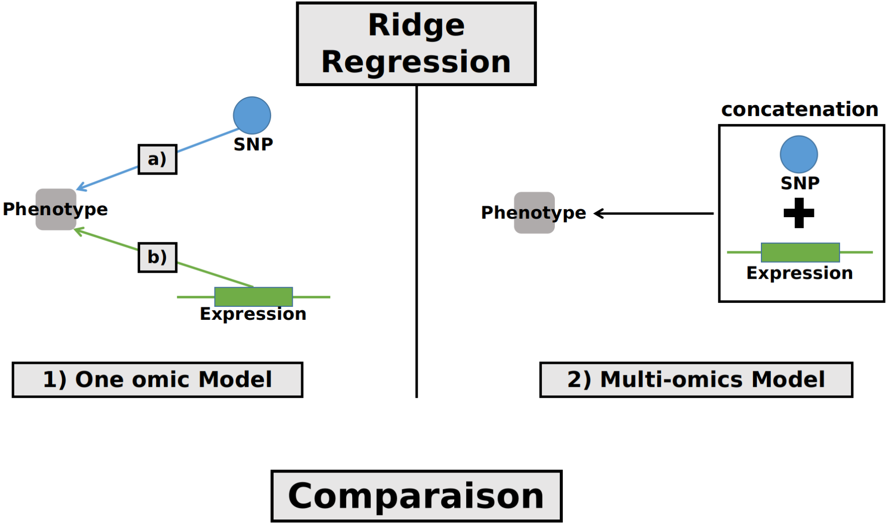


Ridge Regression

1) One omic Model

2) Multi-omics Model

Prediction models



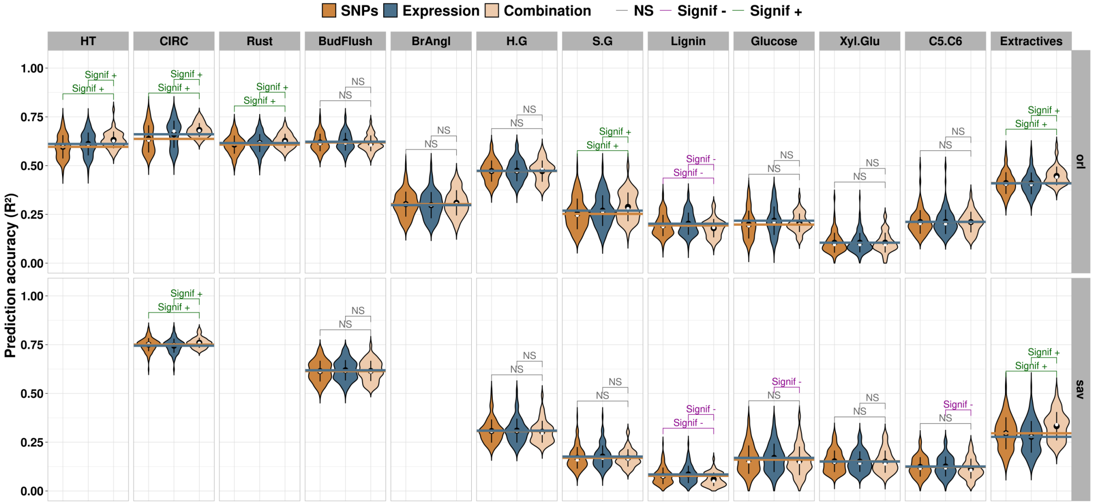
Ridge Regression

1) One omic Model

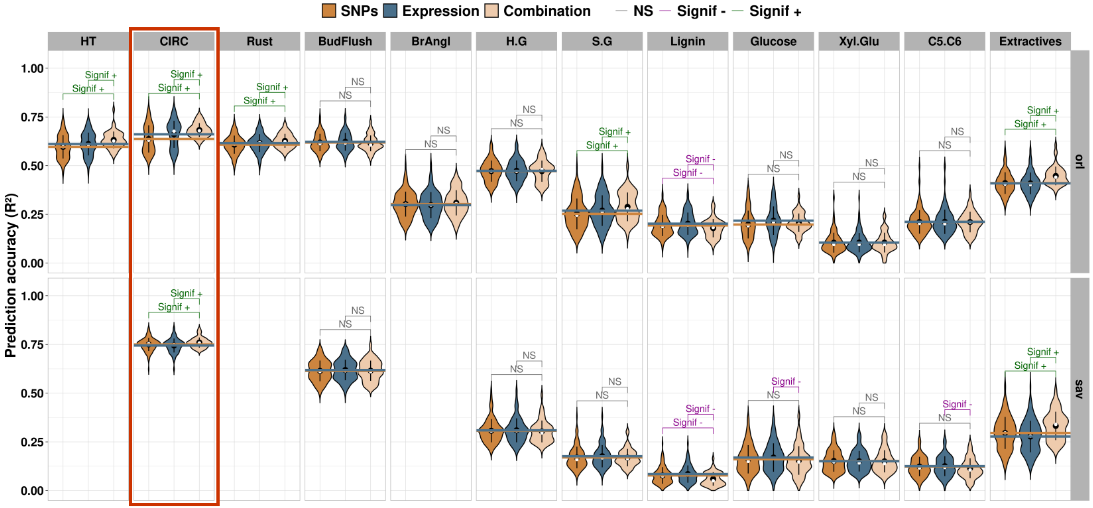
2) Multi-omics Model

Comparaison

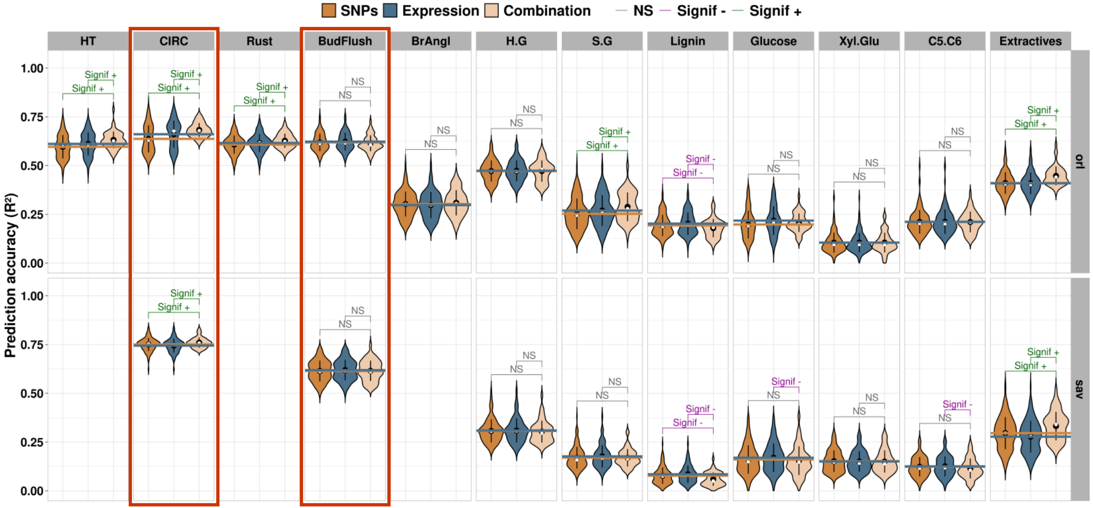
Prediction accuracies



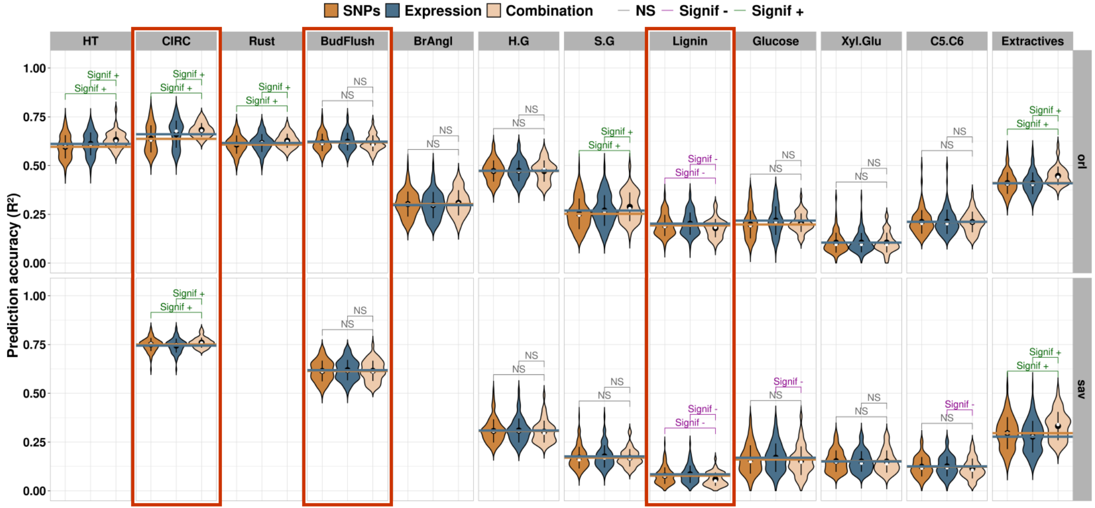
Prediction accuracies



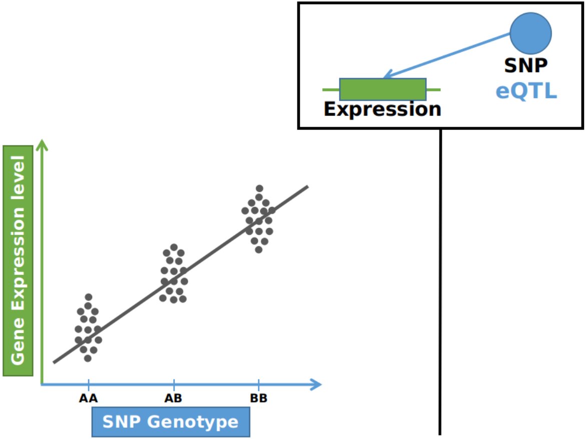
Prediction accuracies



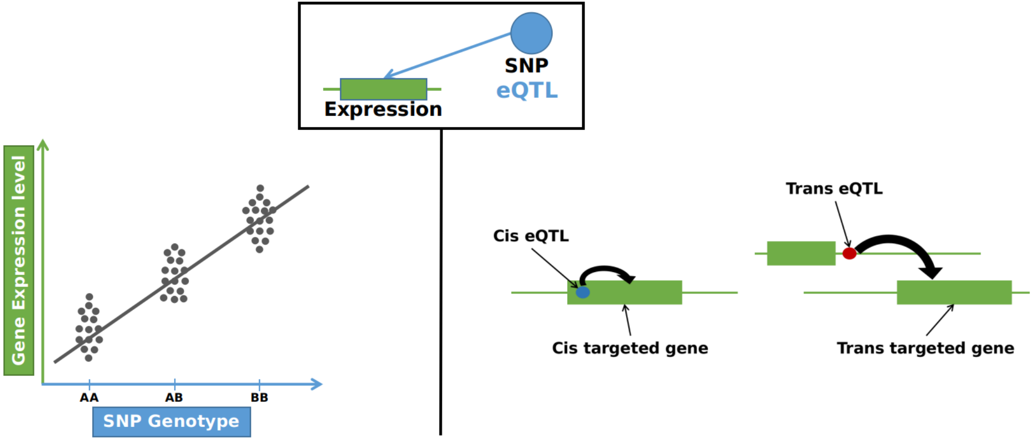
Prediction accuracies



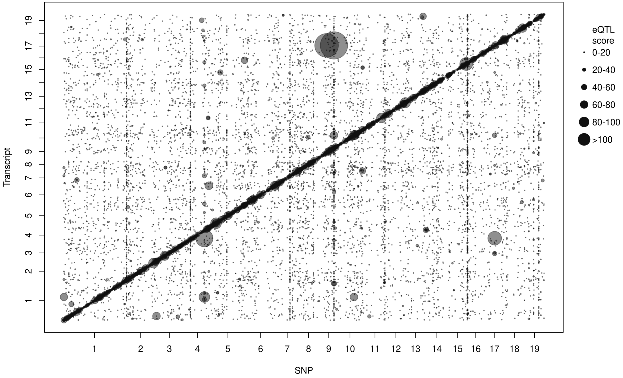
eQTLs → potential redundancy

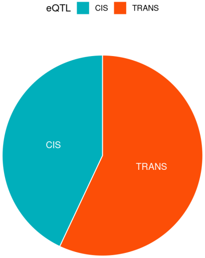


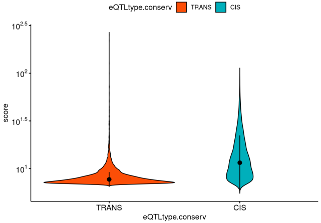
eQTLs → potential redundancy



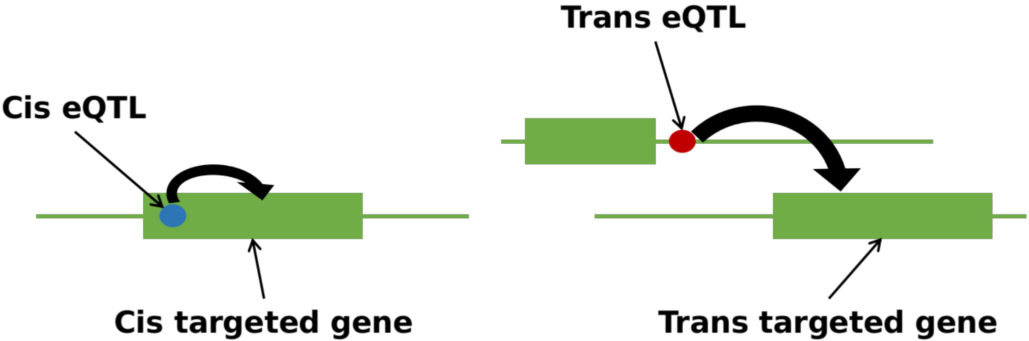
Analyses eQTL





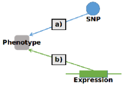


Typologie des variants

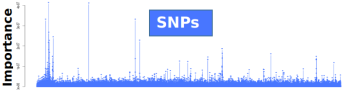


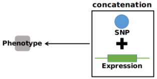
Behavior analysis of the different types of variables effects between the one omic models and the multi-omics model

Change in predictors importance

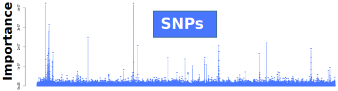


1) One omic Model

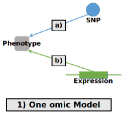




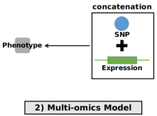
2) Multi-omics Model

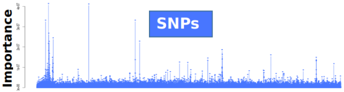


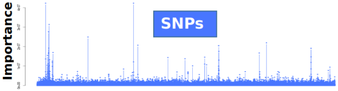
Change in predictors importance



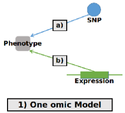
Ranking (ascending order)



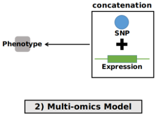


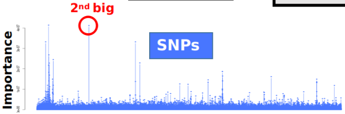


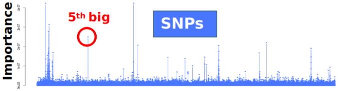
Change in predictors importance



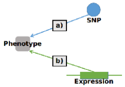
Ranking (ascending order)





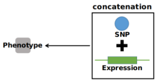


Change in predictors importance

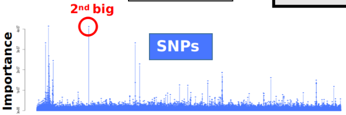


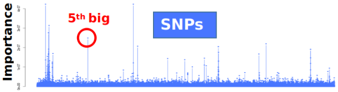
1) One omic Model

Ranking (ascending order)



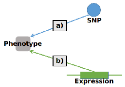
2) Multi-omics Model





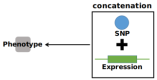
Change in predictors importance = **Rank : Multi-omic Model** − **Rank : One omic Model**

Change in predictors importance

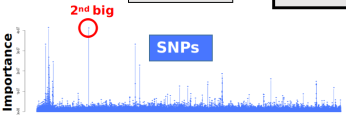


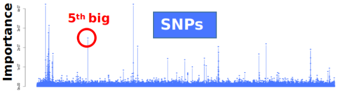
1) One omic Model

Ranking (ascending order)



2) Multi-omics Model

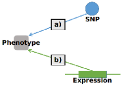




Change in predictors importance = **Rank : Multi-omic Model** − **Rank : One omic Model**

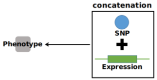
Toy SNP : = -3

Change in predictors importance

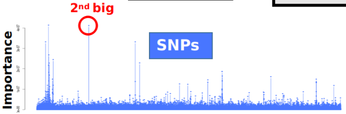


1) One omic Model

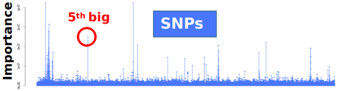
Ranking (ascending order)



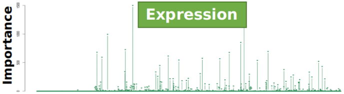
2) Multi-omics Model

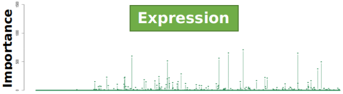


2nd big



5th big

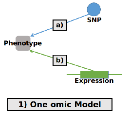




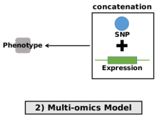
Change in predictors importance = **Rank : Multi-omic Model** − **Rank : One omic Model**

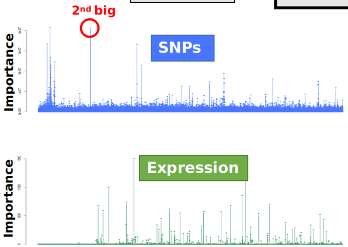
Toy SNP : = -3

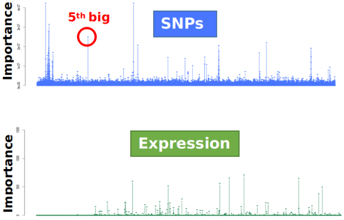
Change in predictors importance



Ranking (ascending order)





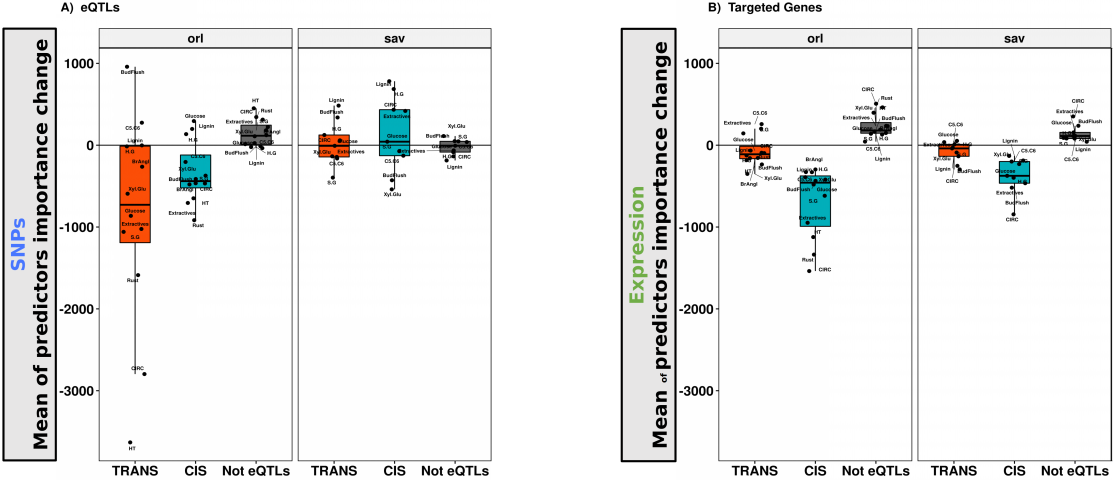


Change in predictors importance = **Rank : Multi-omic Model** − **Rank : One omic Model**

Toy SNP : = -3

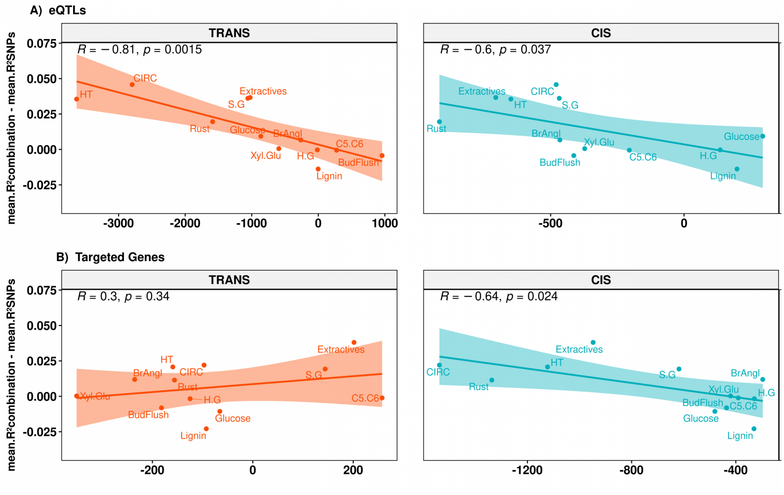
For each trait : mean of predictors importance change

Change in predictors importance



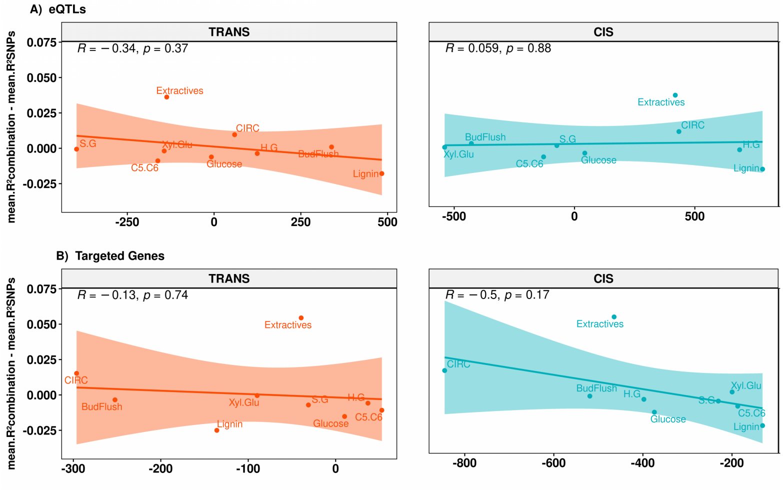
Change in predictors importance *VS* Multi-omics model Gain

Site : Orléans



Change in predictors importance *VS* Multi-omics model Gain

Site : Savigliano



Conclusions

Conclusions

1. The integration advantage varies depending on the trait.

Conclusions

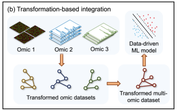
1. The integration advantage varies depending on the trait.
2. The traits that benefit most from integration → **Change in predictor importance for eQTL TRANS effects and CIS regulated transcripts.**

Conclusions

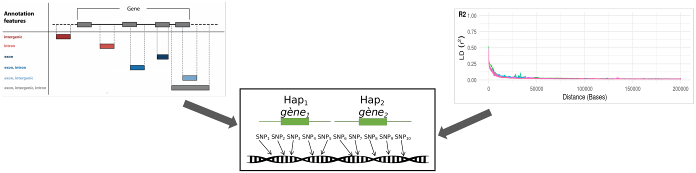
1. The integration advantage varies depending on the trait.

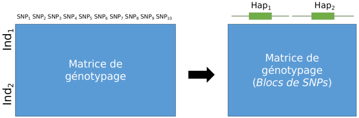
2. The traits that benefit most from integration → **Change in predictor importance for eQTL TRANS effects and CIS regulated transcripts.**

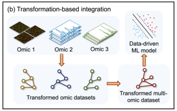
3. The integration advantage → **minimizing the redundancy between predictors**.


(b) Transformation-based integration
Omic 1
Omic 2
Omic 3
Data-driven ML model
Transformed omic datasets
Transformed multi-omic dataset

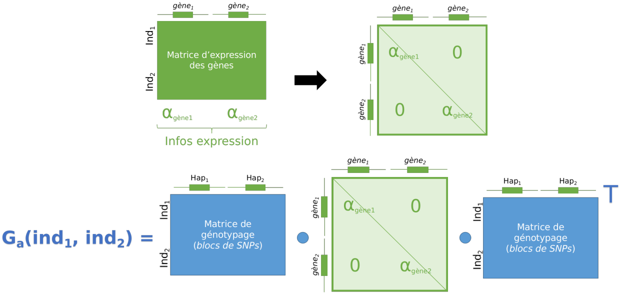
Estimate the *"genes loci"* effects according to their genotypes and the expression level of the corresponding genes.

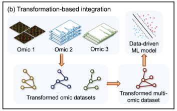



SNP$_1$ SNP$_2$ SNP$_3$ SNP$_4$ SNP$_5$ SNP$_6$ SNP$_7$ SNP$_8$ SNP$_9$ SNP$_{10}$
Ind$_1$
Ind$_2$
Matrice de génotypage
Hap$_1$
Hap$_2$
Matrice de génotypage (*Blocs de SNPs*)

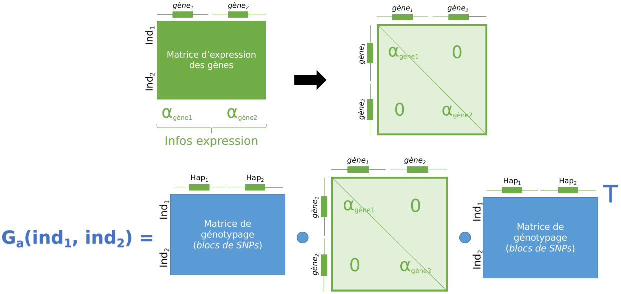

(b) Transformation-based integration
Omic 1
Omic 2
Omic 3
Data-driven ML model
Transformed omic datasets
Transformed multi-omic dataset

Estimate the *"genes loci"* effects according to their genotypes and the expression level of the corresponding genes.




(b) Transformation-based integration
Omic 1
Omic 2
Omic 3
Data-driven ML model
Transformed omic datasets
Transformed multi-omic dataset

Estimate the *"genes loci"* effects according to their genotypes and the expression level of the corresponding genes.



1 **no effect of gene expression level weighting**

Conclusions

1. The integration advantage varies depending on the trait.

2. The traits that benefit most from integration → **Change in predictor importance for eQTL TRANS effects and CIS regulated transcripts**.

3. The integration advantage → **minimizing the redundancy between predictors**.

Conclusions

1. The integration advantage varies depending on the trait.

2. The traits that benefit most from integration → **Change in predictor importance for eQTL TRANS effects and CIS regulated transcripts.**

3. The integration advantage → **minimizing the redundancy between predictors.**

4. **Such relationship was mainly observed for the traits evaluated in the site of transcriptomic sampling (Orléans)**

Conclusions

1. The integration advantage varies depending on the trait.
2. The traits that benefit most from integration → **Change in predictor importance for eQTL TRANS effects and CIS regulated transcripts.**
3. The integration advantage → **minimizing the redundancy between predictors.**
4. **Such relationship was mainly observed for the traits evaluated in the site of transcriptomic sampling (Orléans)**
5. **These results constitute a promising way to explore data integration for multi-omics through differential weighting of features.**

ACKNOWLEDGMENTS

Encadrants :
Leopoldo Sanchez Rodriguez
Vincent Segura

Harold Duruflé

EPINET
Métaprogramme SelGen

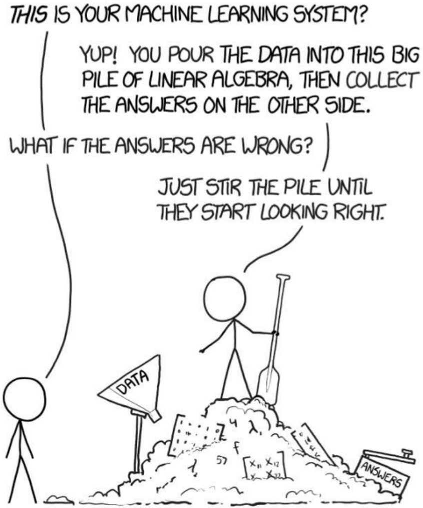
INRA SCIENCE & IMPACT SelGen

B4est

BioForA

INRAE

THANK YOU !


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

REFERENCES

Azodi, Christina B., Jeremy Pardo, Robert VanBuren, Gustavo de los Campos, et Shin-Han Shiu. « Transcriptome-Based Prediction of Complex Traits in Maize ». The Plant Cell 32, no 1 (1 janvier 2020): 139-51. https://doi.org/10.1105/tpc.19.00332.

Guo, Zhigang, Michael M. Magwire, Christopher J. Basten, Zhanyou Xu, et Daolong Wang. « Evaluation of the Utility of Gene Expression and Metabolic Information for Genomic Prediction in Maize ». Theoretical and Applied Genetics 129, no 12 (1 décembre 2016): 2413-27. https://doi.org/10.1007/s00122-016-2780-5.

Li, Zhengcao, Ning Gao, Johannes W. R. Martini, et Henner Simianer. « Integrating Gene Expression Data Into Genomic Prediction ». Frontiers in Genetics 10 (2019). https://doi.org/10.3389/fgene.2019.00126.

Morgante, Fabio, Wen Huang, Peter Sørensen, Christian Maltecca, et Trudy F C Mackay. « Leveraging Multiple Layers of Data To Predict Drosophila Complex Traits ». G3 Genes|Genomes|Genetics 10, no 12 (1 décembre 2020): 4599-4613. https://doi.org/10.1534/g3.120.401847.

Schrag, Tobias A, Matthias Westhues, Wolfgang Schipprack, Felix Seifert, Alexander Thiemann, Stefan Scholten, et Albrecht E Melchinger. « Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize ». Genetics 208, no 4 (1 avril 2018): 1373-85. https://doi.org/10.1534/genetics.117.300374.

Westhues, Matthias, Tobias A. Schrag, Claas Heuer, Georg Thaller, H. Friedrich Utz, Wolfgang Schipprack, Alexander Thiemann, et al. « Omics-Based Hybrid Prediction in Maize ». Theoretical and Applied Genetics 130, no 9 (1 septembre 2017): 1927-39. https://doi.org/10.1007/s00122-017-2934-0.

Ye, Shaopan, Jiaqi Li, et Zhe Zhang. « Multi-Omics-Data-Assisted Genomic Feature Markers Preselection Improves the Accuracy of Genomic Prediction ». Journal of Animal Science and Biotechnology 11, no 1 (1 décembre 2020): 109. https://doi.org/10.1186/s40104-020-00515-5.