Integrating GENomic prediction with GENe regulatory networks to optimize genetic value prediction : biological and statistical challenges

Is « functional understanding » relevant for prediction objectives ?
If it is the case how we take it into consideration ?

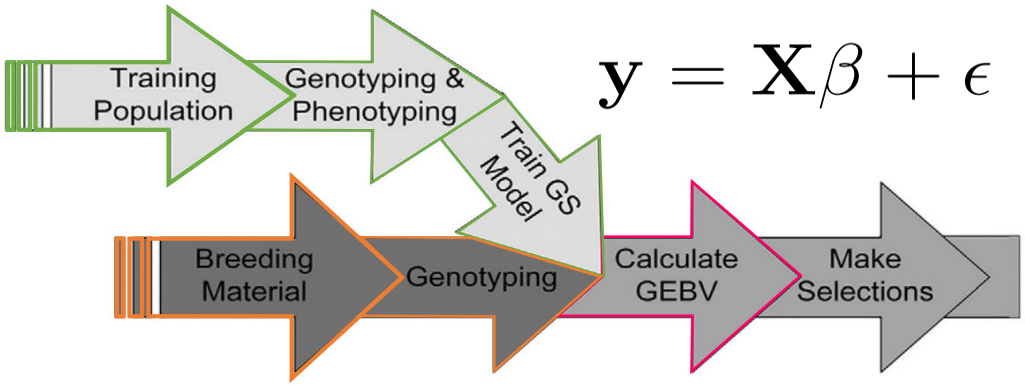
Marie Denis and David Pot

NetBIO Seminar, November 15th 2023

Genomic prediction

Genotype	T1	T2	…	M n
1	10	1.7		
2	12	2.1		
3	15	1.5		
4	14	3.2		
5	16	4.1		
…				
n	20	0.8		

Genotype	M 1	M 2	M 3	M 4	M 5	…	M n
1	A	C	T	C	A		
2	T	C	G	C	G		
3	A	C	G	G	G		
4	A	C	G	G	A		
5	T	A	G	G	A		
…							
n	A	A	T	C	A		



$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Different GP methods : BLUP and Bayes alphabets

Approach	Marker effect	Marker effect distribution	GEBV estimation	The marker effect variance
BLUP alphabets (GBLUP, CBLUP, SBLUP)	All markers are assumed to have effects on the trait variability	Marker effects are assumed to follow normal distribution.	All marker effects are used for estimating GEBV	Common variance for all marker effects
Bayesian alphabets (BayesA, BayesB, BayesC, BL and BRR)	Only a limited number of markers are assumed to have effects on the trait variability	Different prior distributions are considered for different Bayesian models	BayesA, BL, and BRR are shrinkage type models where some of the marker effects are shrunk to zero, and rest of the markers are used for estimating GEBV. In BayesB and BayesC, the markers with non-null effects are used for estimating GEBV	Common variance for BayesC and BRR. Marker specific variances for BayesA, BayesB, and BL

Meher et al., 2022

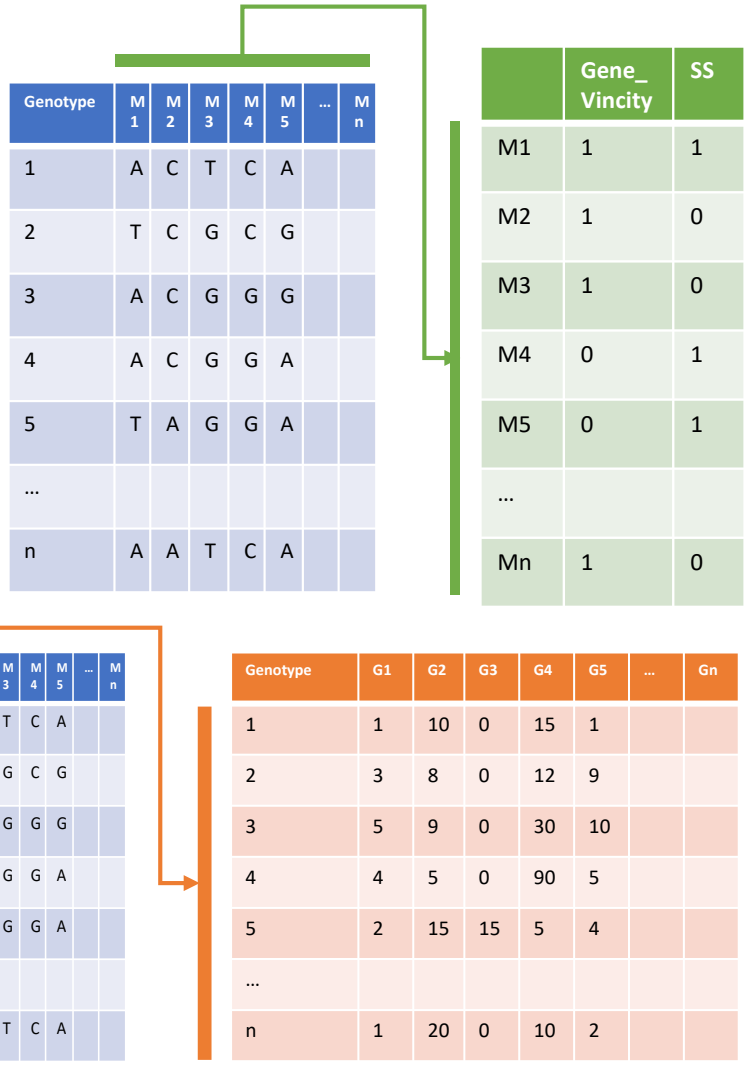
- As expected the efficiency of the different methods vary according to the genetic determinism of the traits (QTL numbers, heritability)
- GBLUP is the less biased method as far as GEBV estimation is concerned
- Functional information is not used at all in these contexts…

Adding functional information in GP

- **Functional information**: information allowing to mobilize biological understanding to optimize prediction accuracy

  - SNP with ANNOTATION :
    - Location in or in the vicinity of a gene
    - Location in GWAS peak
    - Go-Term
    - Location in or in a gene harboring selection signature
    - Expression information :
      - module membership
      - Connectivity level (hub gene or not)
      - Correlations with other genes
    - …

  - Combination of polymorphism information with transcriptomic, proteomic, metabolomic informaton on the same set of genotypes



Genotype	M1	M2	M3	M4	M5	…	Mn
1	A	C	T	C	A		
2	T	C	G	C	G		
3	A	C	G	G	G		
4	A	C	G	G	A		
5	T	A	G	G	A		
…							
n	A	A	T	C	A		

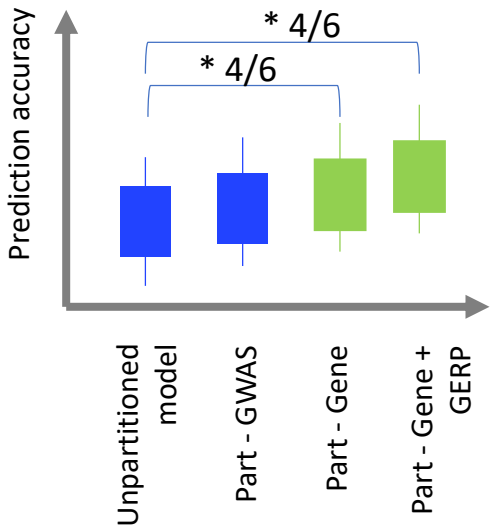
	Gene_ Vincity	SS
M1	1	1
M2	1	0
M3	1	0
M4	0	1
M5	0	1
…		
Mn	1	0

Genotype	M1	M2	M3	M4	M5	…	Mn
1	A	C	T	C	A		
2	T	C	G	C	G		
3	A	C	G	G	G		
4	A	C	G	G	A		
5	T	A	G	G	A		
…							
n	A	A	T	C	A		

Genotype	G1	G2	G3	G4	G5	…	Gn
1	1	10	0	15	1		
2	3	8	0	12	9		
3	5	9	0	30	10		
4	4	5	0	90	5		
5	2	15	15	5	4		
…							
n	1	20	0	10	2		

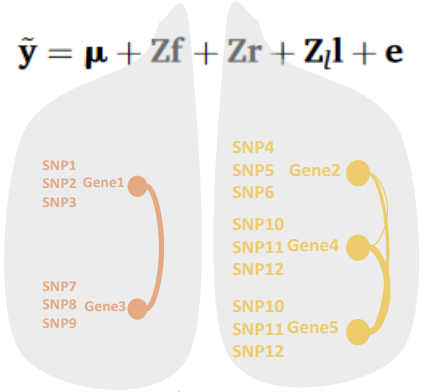
Adding annotations to GP: achieved results

- Ramstein et al., 2020, 2022: 3 traits, 2 cross validation schemes

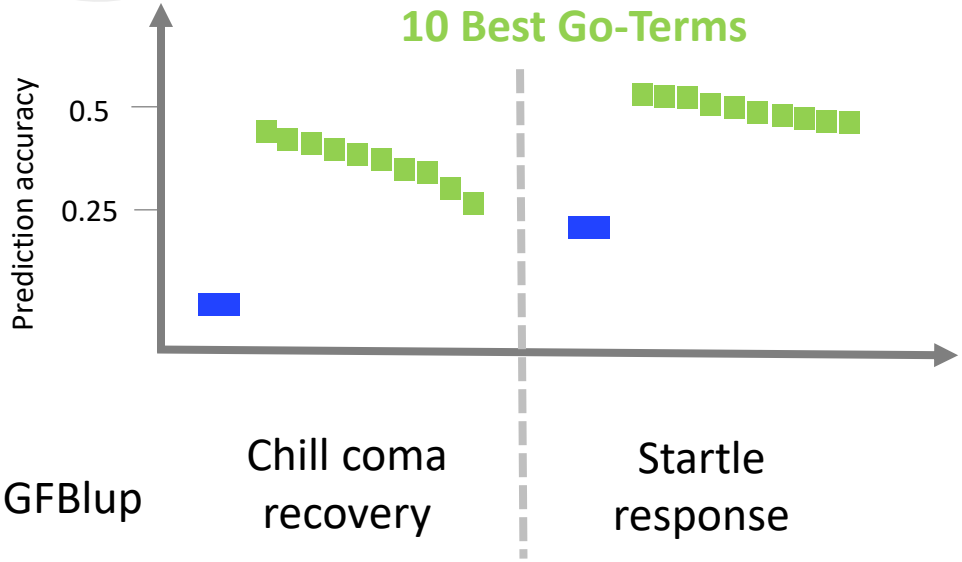


- GFBlup : Genomic Feature BLUP (Edward et al., 2016)

$$\tilde{y} = \mu + Zf + Zr + Z_l l + e$$



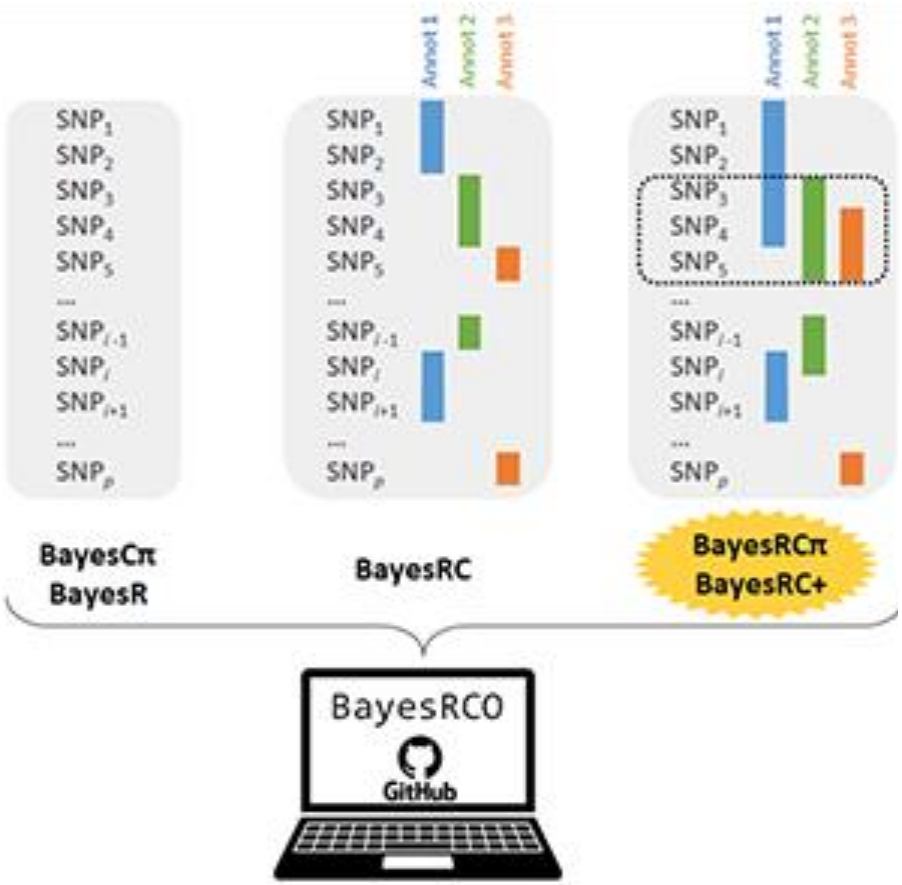
Feature = Go-Term



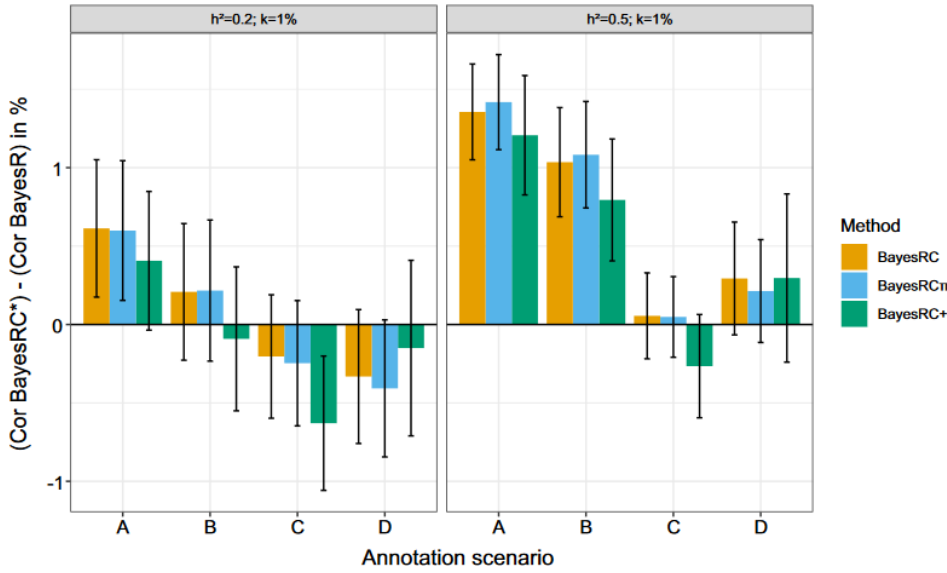
- GBLUP partitioned based models : only one feature is « partitionned »

- Renaud Rincent and Ali Baber also tested GFBlup on maize with lower success (Pers Com)

Adding annotations to GP: achieved results



	Annotation enrichment			
	Strongly	**Moderately**	**Weakly**	**Unenriched**
SNP effect class				
Large	5	2	–	–
Medium	300	100	20	–
Low/null	150	300	400	450
Scenario				
A	1	1	–	–
B	1	1	1	1
C	–	2	1	1
D	2	2	3	2



A Rau

P Croiseau,

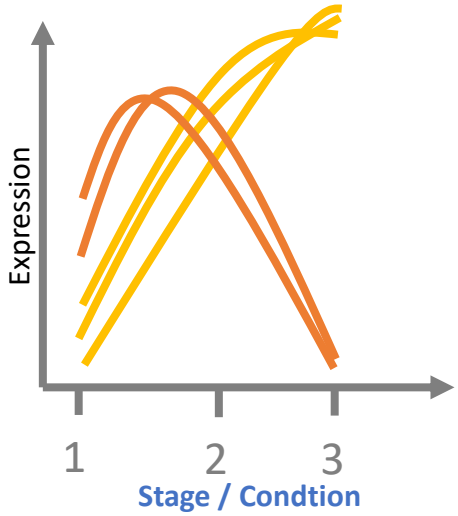
Mollandin et al., 2022

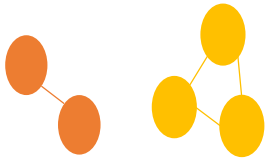
- Significant gains (when relevant annotation are used) but reltatively limited

Functionnal information : Gene Co-expression NetworkS (1)

- Different « types » of networks
  - Developmental Networks : Genotype = 1, Tissue = 1, Conditions >1

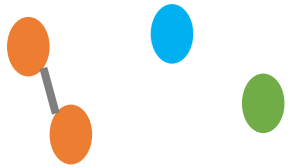
Genotype	Treat	Gene1	Gene2	Gene3	Gene4	Gene5
1	1	1	2	3	7	6
1	2	10	15	14	15	13
1	3	20	25	24	0	1



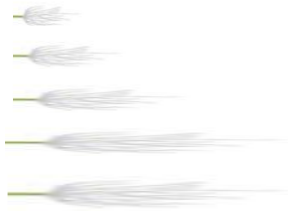


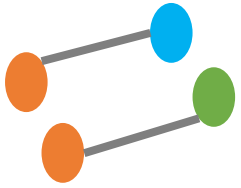
Developmental netwoks are variable depending on…

Genotype	Treat	Gene1	Gene2	Gene3	Gene4
1	1	1	10	1000	0
1	2	10	100	100	0
1	3	20	200	500	100
1	4	30	300	50	200
1	5	40	400	120	300
1	6	50	500	400	1000
1	7	60	600	900	2000
1	8	70	700	1000	0
1	9	80	800	1500	2
1	10	90	900	10	1



Genotype	Treat	Gene1	Gene2	Gene3	Gene4
1	1	1	10	100	100
1	2	1	20	200	101
1	3	2	30	300	103
1	4	3	40	400	105
1	5	0	50	500	100





Developmental network vary according to the tissues / conditions. But these variabilities are relevant ! They will probably be of variable interests according to the target trait to predict

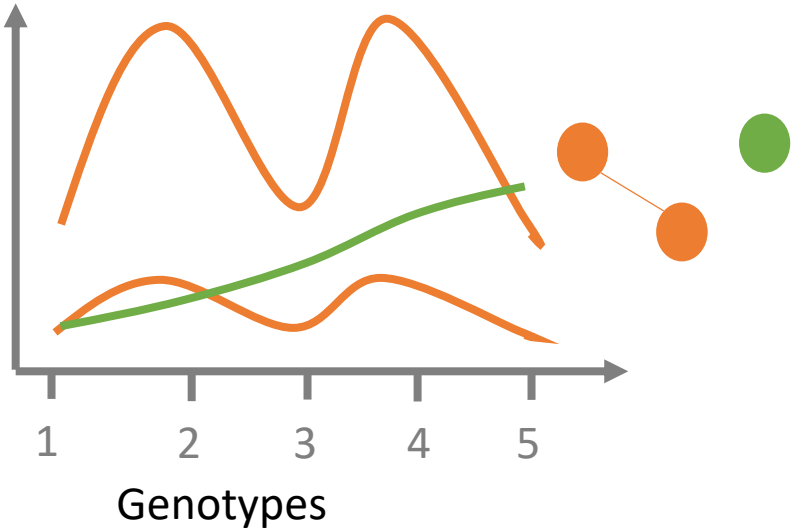
Gene variability networks

- Genetic variability Networks : Genotype >1, Tissue / Condition = 1

Genotype
1
2
3
4
5

Treat
1
1
1
1
1

Gene1	Gene2	Gene3
10	100	10
20	200	15
10	100	20
20	200	30
10	100	40



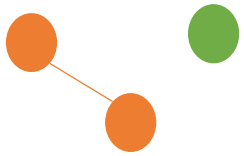
- There are strategies to :
  - Identfy the edges that are stable over several genotypes
    - Leave 1 outstrategy: Maud Fagny (personnal com)
  - Or to develop « genotype based » GCN even if only one tissue was sampled
    - Kruijer et al 2019

Gene variability x developmental networks

- Genetic variability Networks : Genotype >1, **Tissues / Conditions >1**

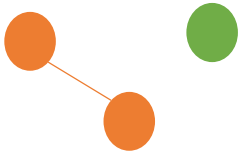
Genotype	Treat
1	1
1	2
1	3

Gene1	Gene2	Gene3
10	100	10
20	200	15
10	100	20



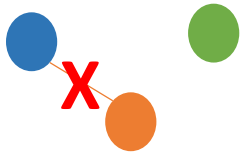
Genotype	Treat
2	1
2	2
2	3

Gene1	Gene2	Gene3
10	100	10
20	200	15
10	100	20



Genotype	Treat
3	1
3	2
3	3

Gene1	Gene2	Gene3
10	10	10
20	10	15
10	10	20



GCN: a growing set of information likely to be useful ?

- Different types of Gene Co-Expression Networks are available
- They contain different types of information that reveal real biological mechanisms
  - f(condition),
  - f(genotype),
  - f(genotype x condition)
  - **and a lot of noise (ok)…**
- Each day their number is growing
- We can argue that they are :
  - Unstable
  - Noisy
  - Not designed initially to develop prediction models
  - That they are most of the time not relevant for the target trait
  - **But still…**
    - **They contain molecular and genetic information**
    - **And there are available « for free » ! (As signature of selection…)**

« The geneticist » point of view : mmhh, my point of view…

- **Stable parts of the CGN are relevant**: they allow to identify clear links between genes that will be « always » connected
  - If a polymrophism in gene 1 will impact a trait, it is likely that its « linked-gene » will also contribute to the target trait (expression variability, metabolic flux…)

- **Unstable parts of CGN are relevant :**
  - Between tissues / conditions, they inform on the plasticity of the transcriptome according to organogenesis, ontogeny, environmental effects
  - Between genotypes, they inform on the beauty of genetic diversity that is a key information to predict genetic variability at the integrative phenotype level

- **Difficult to engage biologist, geneticists, evolutionists, statisticians, breeders in the same « game »:**
  - Biologist : « your CGN is not based on the relevant tissues / stages to predict your trait, you should work with mutants ! »
  - Geneticist / breeders : « CGN – Go-terms… are only useful to understand but not to predict and even less to breed ! »
  - Statistician : » you need to really understand what are the links between your genes, reduce the number of actors, exploring the whole « landscape » is too « expensive »… »

A dream (or a nightmare) or just a stupid idea

- The end-use traits breeders want to predict depends on **Genotypes**, **Ontogeny**, **Environment**, and a lot of **interactions between these main factors**
- Depending on the ontogenic stage, environments different metabolic pathway will be key (height depend on cell division, then on cell growth, then on meristem behavior…)
- Signature of selection, gene co-expression networks describe these dynamic processes
- The challenge is to identify in a large set of environmental contexts (availability of Environmental Covariates) what are the components of the CGN that are systematically important and the ones that are specific to particular contexts
- Using CGN modules (that are already available) as priors to identify sets of polymorphisms that are likely to impact the traits does not seem too stupid…
- The challenge is probably to **agregate systematically CGN (and other annotations) and be able to optimize the model recurrently**. Exploring this growing « space » is probably extremely expensive (but probably less than a purely blind approach ?)

Let’s dive in more precise statistical aspects