Next generation sequence data in gene mapping and genomic prediction: opportunities and challenges

Boichard D.

GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy en Josas, France

Introduction

Genomic Prediction (GP) is found to work well with medium density markers

But

- this is true for populations of small effective size (ie a limited number of independent chromosomal segments)
- when candidates are strongly related to the reference population, and there is a rapid loss in persistency when this relationship decreases
- it has a limited efficiency for across breed prediction, and too low for practical purpose, except when populations are strongly related

In conclusion, present methods rely on long range LD with causal variants and/or overall relationship

Whole Genome Sequences will be available at a large scale

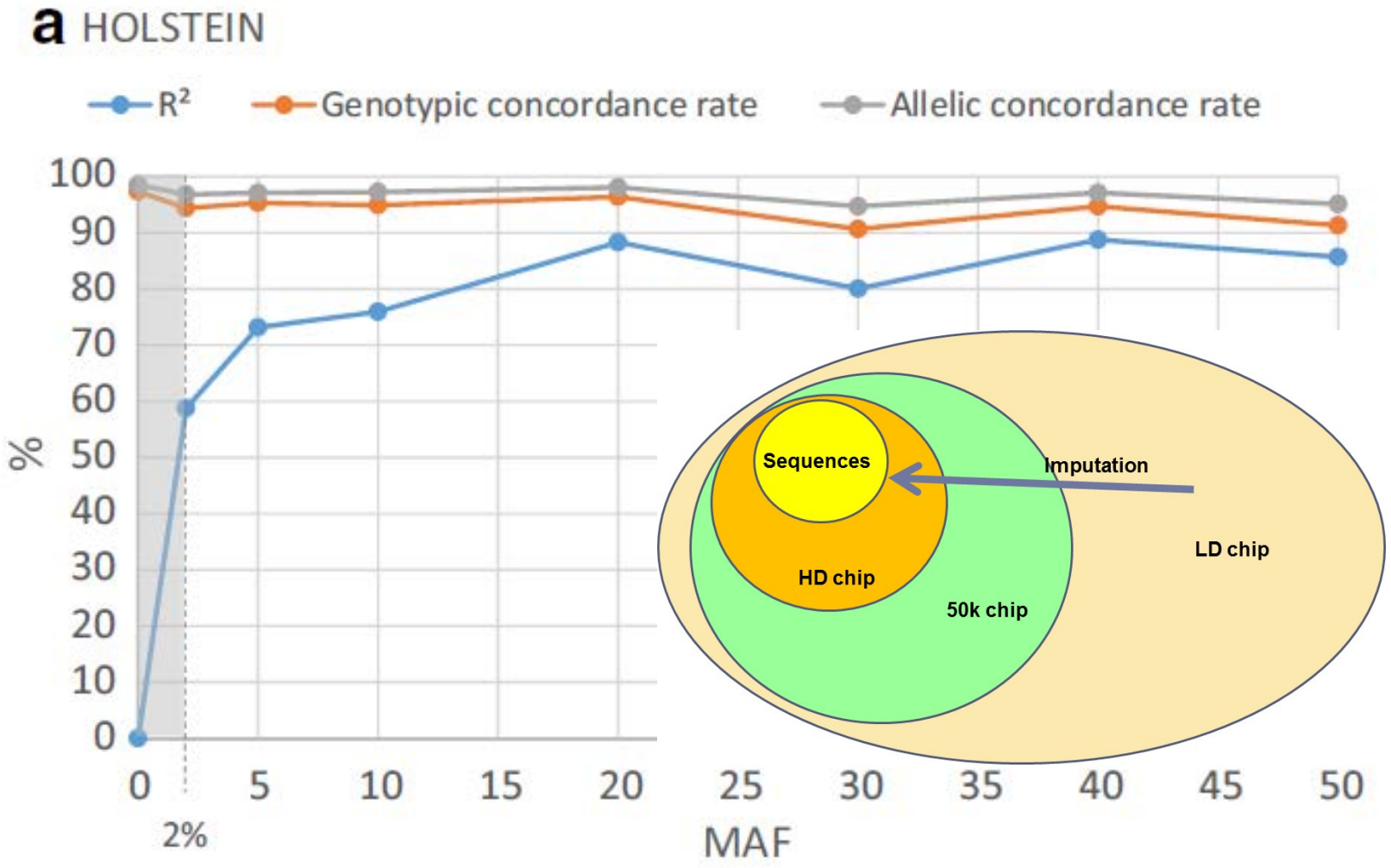
For us, WGS corresponds to a complete genotyping process (eg, ~25M variants vs 50k with a chip)

Sequencing cost is decreasing

Cost still too high for all individuals at medium-high coverage, but low enough for a few thousands of individuals, especially in consortia

But imputation is efficient and cost-effective
(with some limitation, however, with low MAF markers)

Imputation accuracy



Whole Genome Sequences will be available at a large scale

An alternative with low coverage sequencing:
(eg, Hickey et al; NRGene)

1) Deep sequencing **and phasing** for a collection of individuals, describing most haplotypes present in the population
2) Determination of tag-SNPs for each haplotype
3) Low-coverage (0.1 to 1X) of many individuals, therefore at low cost
4) Fast sequence reconstruction based on tag-SNPs

What can we do with Whole Genome Sequences

WGS include all variants

(in fact, not fully exact: calling errors, imputation errors, structural variants…)

Therefore WGS should include causal variants

Assuming they are known, including all causal variants directly in predictions would provide the highest accuracy.

They are there, but as a needle in a hay stack

=> Can we find them ?

WGS are useless without variant selection

No gain in accuracy of sequenced-based GBLUP
over 50k based GBLUP

Too much noise generated by millions of useless variants, to take advantage of the causal variants

Another interpretation: genomic relationship matrices are very similar

Necessity to extract the useful information, ie causal variants or very close variants,
And to discard the other variants which are source of noise.

Expected gain in accuracy within population

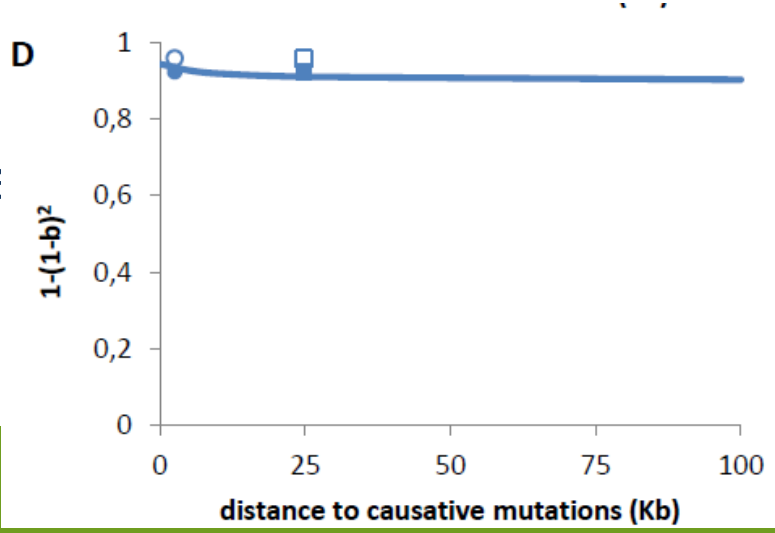
Studies by simulation, assuming causal variants are known

Accuracy is (obviously) maximum with causal variants

But little is lost with neighbor variants, even a some distance
Markers capture most of genetic variance due to long range LD
No/little improvement over one generation

Van den Berg, G3
250 variants

But a better persiste bination
(any publication?)



Expected gain in accuracy across populations

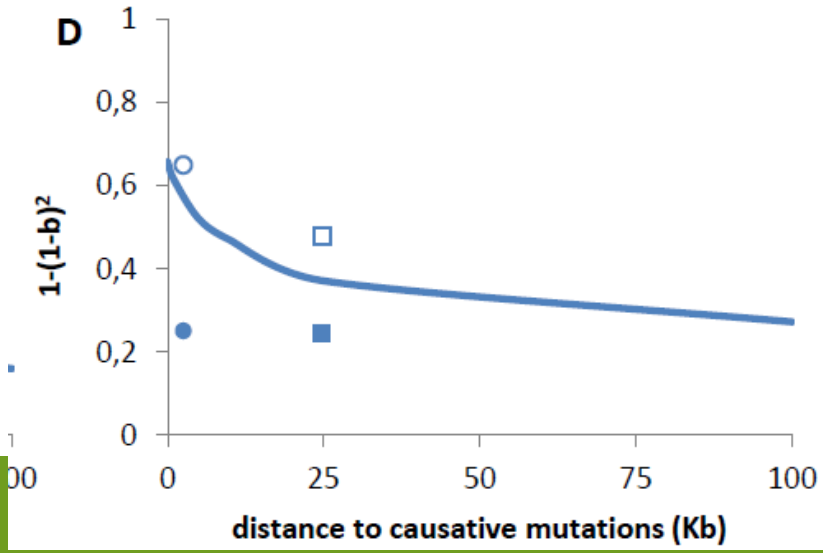
Accuracy is maximum with causal variants

It is less than 1, due to fixed causal variants

Rapid decay in accuracy with distance due to LD decay

(NB: Optimistic situation: constant and additive only effects)

Van den Berg, G3
250 variants



Question

Are we able to identify or at least to accurately map the causal variants?

Strategy for mapping

1. Through GWAS

- Imputation at the sequence level
- GWAS at the sequence level on large populations
- Joint analysis to test for the lack of residual effect in the region
- Across breed GWAS or meta-analysis to improve resolution
- Functional annotation to orient variant selection
- Possibly, haplotypic analysis

Example

Sanchez *et al. Genet Sel Evol (2017) 49:68*
DOI 10.1186/s12711-017-0344-z

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle

Marie-Pierre Sanchez[1*], Armelle Govignon-Gion[1,2], Pascal Croiseau[1], Sébastien Fritz[1,3], Chris Hozé[1,3], Guy Miranda[1], Patrice Martin[1], Anne Barbat-Leterrier[1], Rabia Letaïef[1], Dominique Rocha[1], Mickaël Brochard[2], Mekki Boussaha[1] and Didier Boichard[1]

Material & methods: animals

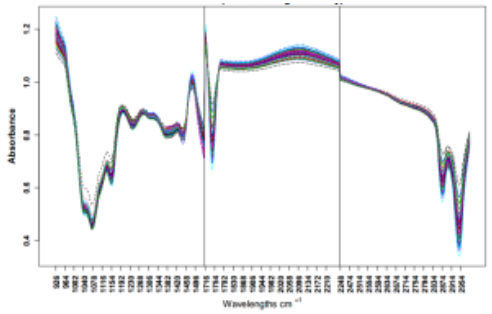
PhénoFinlait

8,752 cows genotyped with the 50k Beadchip and with milk composition phenotypes

2,967	**2,737**	**3,048**
Montbéliardes	Normandes	Holstein
MON	**NOR**	**HOL**

MIR Prediction

PhénoFinlait



N=450

Trait	Relative estimation error (%)	R^2
C14:0	4	0.97
β Casein	4	0.90
C18:0	10	0.89
α_{s1} Casein	6	0.81
Omega 3	13	0.89

23 Fatty acids and 6 proteins

UNSAT
MONO
C18:1cis9
C18:1cis12
C18:1t11t10
TotC18:1
TotC18:1cis
TotC18:1trans
POLY
C18:2cis9trans11
C18:2cis9cis12
C18:3n3
TotC18:3
Omega 3
Omega 6

SAT
C4:0
C6:0
C8:0
C10:0
C12:0
C14:0
C16:0
C18:0

αs1 casein
αs2 casein
β casein
κ casein
α lactalbumin
β lactoglobulin

Material & methods: genotypes & imputation

***Imputation in two steps with FImpute** (Sargolzaei et al., 2014)*

Bovine SNP50

Step 1 Within breed imputation
Within breed RP

Bovine HD

Step 2 Within breed imputation
One across breed RP

Whole genome sequence

Reference populations (RP)

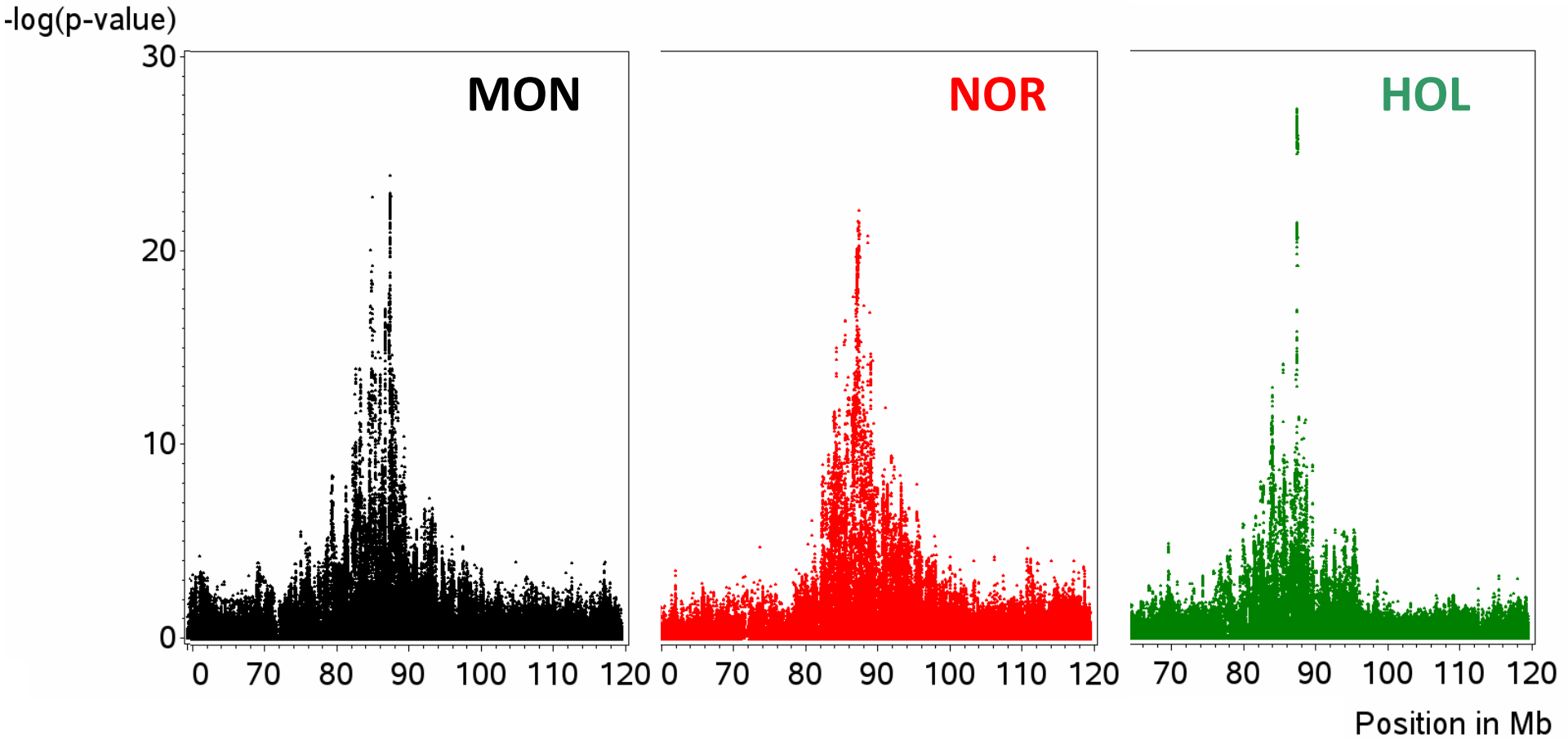
Within breed,
HD genotyped bulls
522 MON
546 NOR
776 HOL

Across breed,
WGS of 1147 bulls
(«1000 Bull Genomes», RUN4)
including 28 MON + 24 NOR + 288 HOL

27 millions of sequence variants imputed for **8752** cows

GWAS results

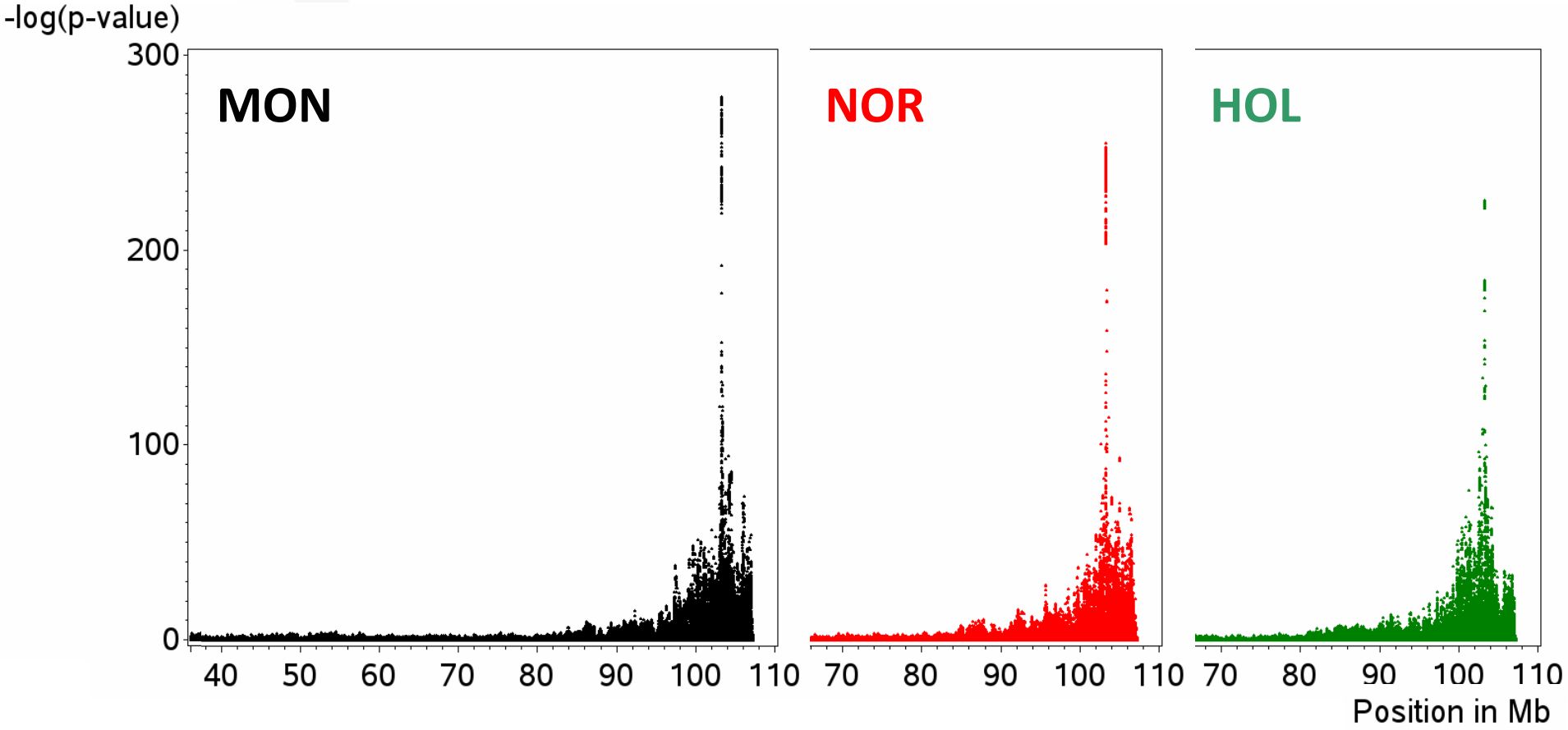
BTA 6 – κ casein



~87 Mb => in the casein genes cluster

GWAS results

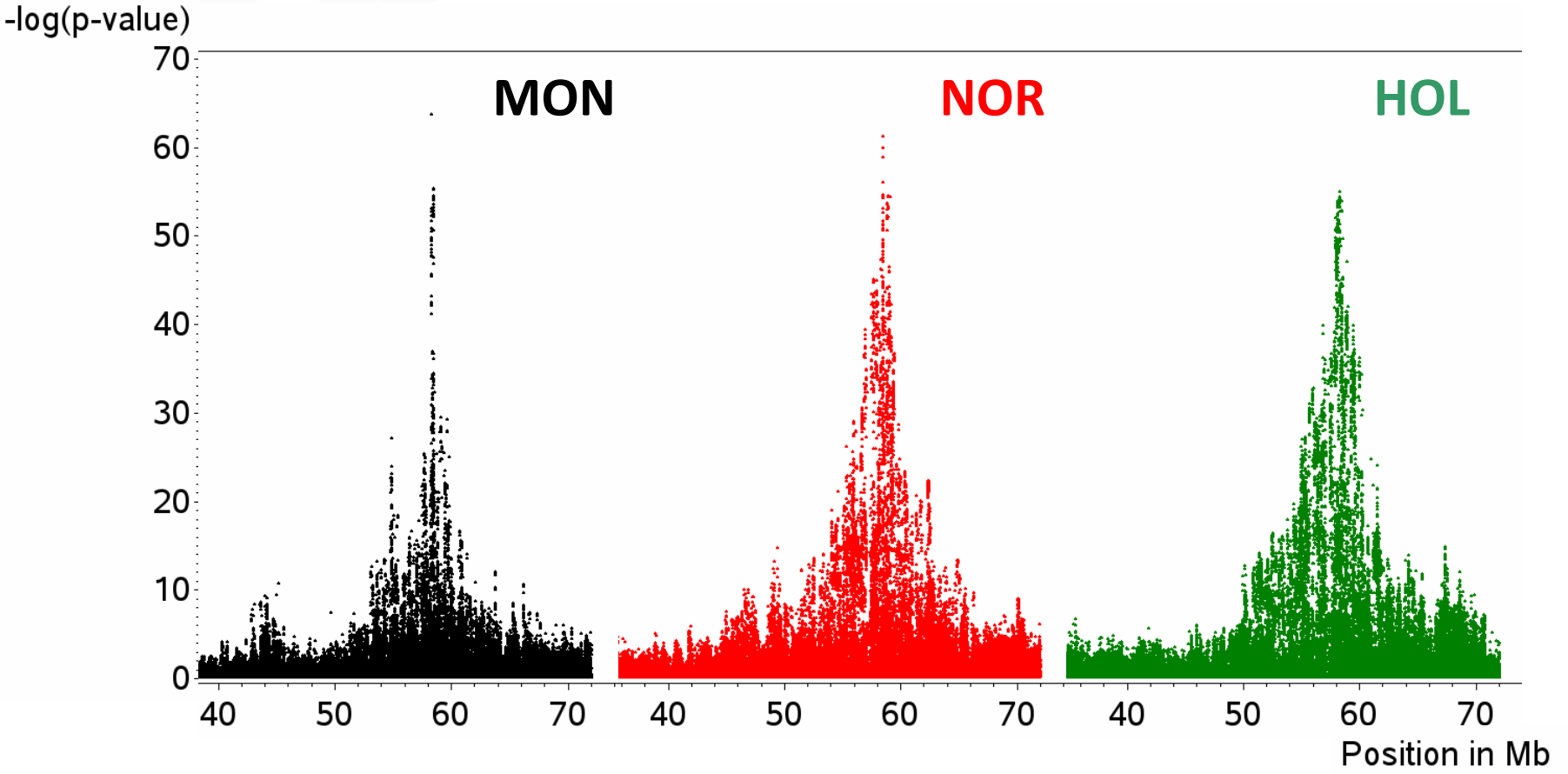
BTA 11 – β lactoglobulin



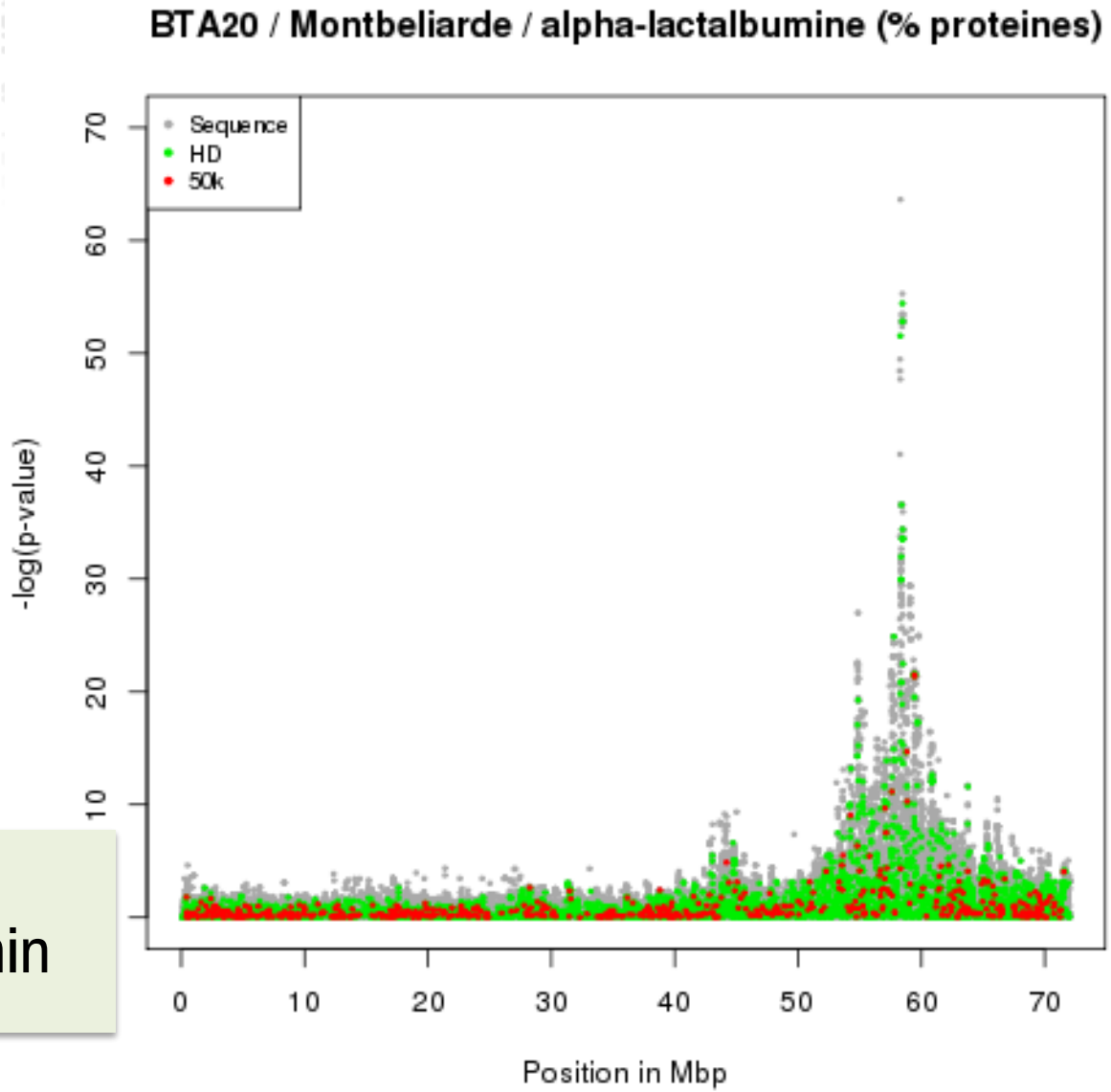
~103 Mb => close to LGB (PAEP) gene

GWAS results

BTA 20 – α lactalbumin



~58 Mb => ANKH gene



BTA 20
α lactalbumin

Results

Nearly all major peaks are on genes

SLC37A1, MGST1, ABCG2, CSN1S1, CSN2, CSN1S2, CSN3, PAEP, DGAT1, AGPAT6, ALPL, ANKH, PICALM affect milk protein

In majority out of the coding sequences

Strategy for mapping

2. Multi-marker analysis with Bayesian methods

2.A. At the full sequence level, with BayesR

- A computational challenge
- Some improvements made by the Australian colleagues (T Wang, I Van Den Berg)
  - Burn-in replaced by EM
  - SNP discarding (remove if not selected after N iterates)

2.B. At the chromosomal region level

- Account for the rest of the genome (pedigree, 50k)
- BayesC over 1-2 Mb including a QTL

Some GWAS QTL results *analyzed in detail*

Proteins

BTA	Trait	$Log_{10}(1/p)$ max MON – NOR – HOL	Bounds of the region analyzed (Mb)
1	κ casein	**10** – **9** – **13**	143 – 145
2	αs2 casein	**8** – **12** – **7**	130.5 – 132.5
6	κ casein	**24** – **22** – **46**	86.5 – 88.5
11	β lactoglobulin	**279** – **255** – **226**	102 – 104
20	α lactalbumin	**64** – **44** – **34**	57 – 59

QTL shared between breeds

Bayesian analyses

Candidate variants were selected according to their **probability of inclusion** (based on 100,000 iterations, burn-in=20,000, thin = 50)

A **difficulty**: due to **very high LD**, inclusion probability of a region may be distributed over many linked variants,
and can be low for individual variants

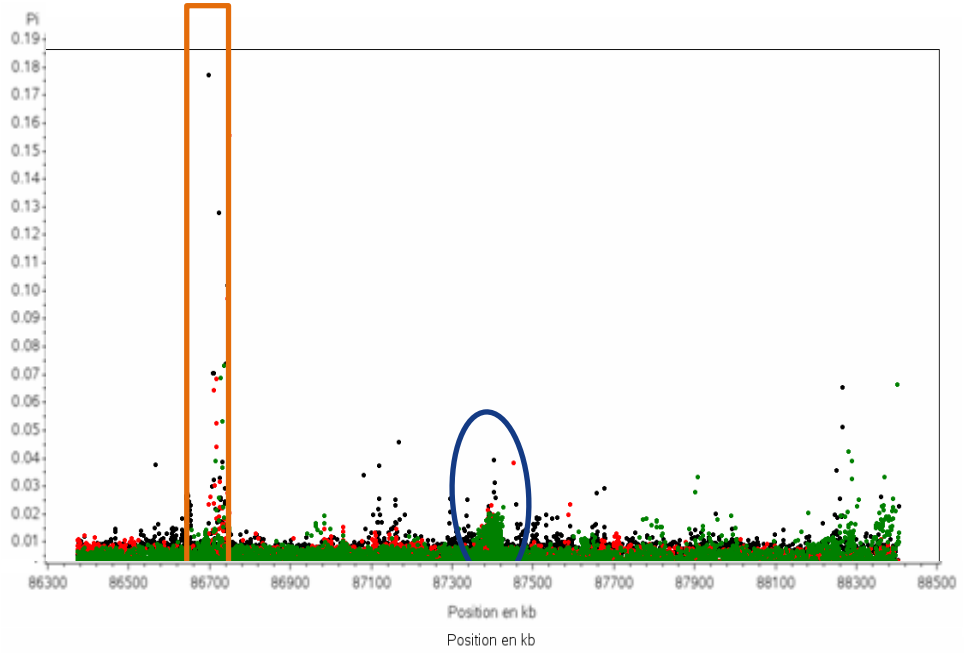
Inclusion probabilities were **summed over 5kb windows** to detect the largest signals, and
candidate variants were searched within the best windows

Bayesian analysis results: *BTA6, κ casein*

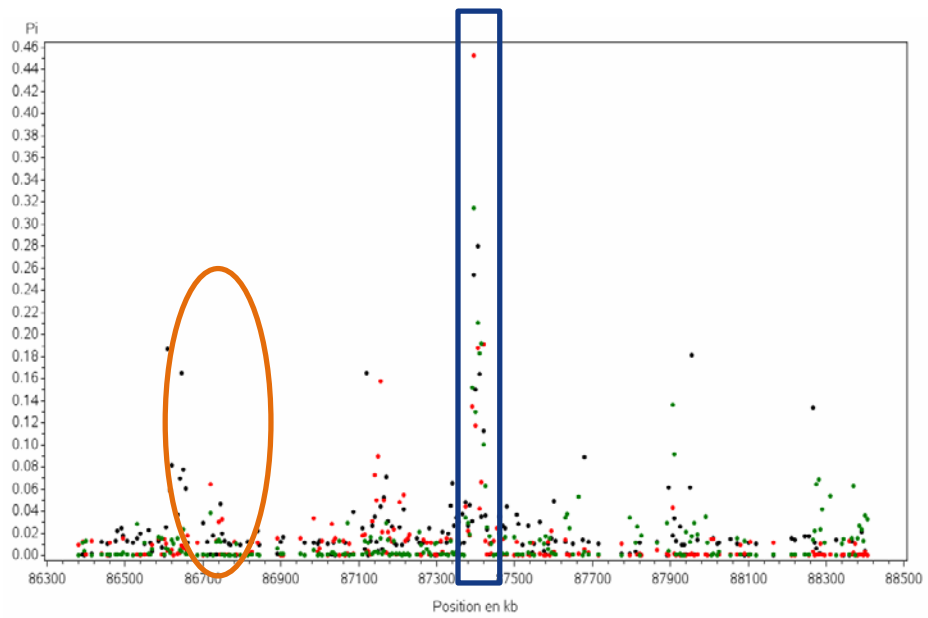
BTA6 – 86 – 88.5 Mb

Use of sums of Pi per 5 kb intervals to detect the strongest signals

Individual SNPs



Sum of Pi over 5kb windows

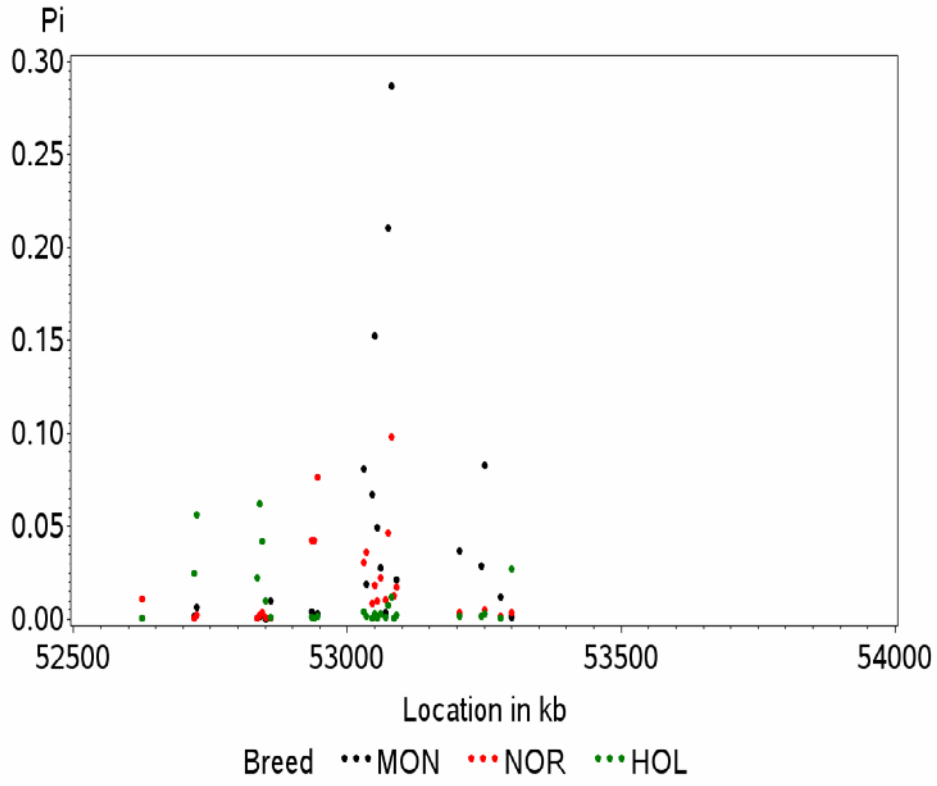


Breed ••• MON ••• NOR ••• HOL

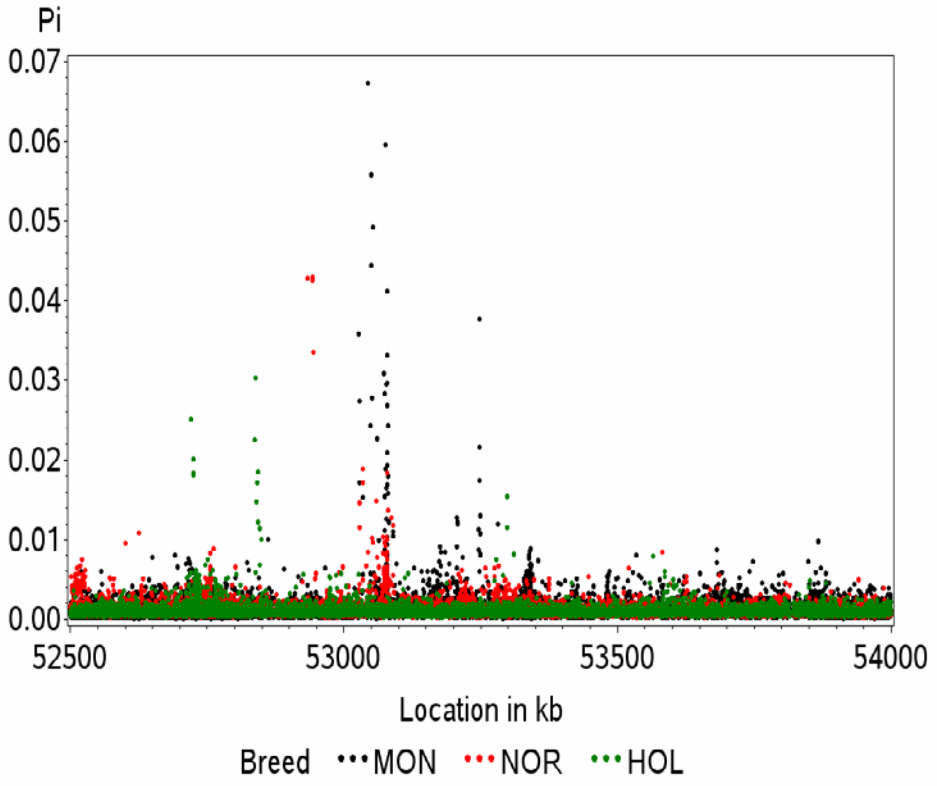
40 variants in the strongest regions between 87,390 and 87,410 Mb, including **10 in the regulatory region of CSN3 gene**.

Bayesian analysis results: *BTA17, C4:0*

Sum over 5kb windows



Inclusion probability for each SNP



25 markers in very high LD, with similar probabilities, in the **BRI3BP** gene (all intronic)

Summary of results (fatty acids)

BTA	Bounds of peak (kb)	Trait	Candidate variants	Genes	Annotation of variants in genes
5	93,940-93,955	SAT	4	MGST1	**Upstream**
	1,620-1,625	SAT, POLY	1	GPT	**3'UTR**
14	1,790-1,870	SAT	4	DGAT1	**Various**
	2,700-2,720	POLY	4	CYP11B1	**Upstream / Downstream**
17	53,075-53,085	C4:0	22	BRI3BP	**Intronic**
19	51,360-51,385	C12:0	6	FASN	**Upstream**
27	36,205-36,220	C16:0	4	AGPAT6	**Upstream**

Summary of results (proteins)

BTA	Bounds of peak (kb)	Trait	Candidate variants	Genes	Annotation of variants in genes
1	144,395-144,405	κ casein	30	SLC37A1	**intronic**
2	131,810-131,835	αs2 casein	1	ALPL	**intronic**
6	87,390-87,410	κ casein	10	CSN3	**regulatory regions**
11	103,285-103,315	β lactoglobulin	20	LGB	**1 missense** (Ganai et al, 2009) **19 in regulatory regions**
20	58,410-58,440	α lactalbumin	10	ANKH	**10 intronic**

A nice confirmation tool, the custom chip

- Chips can be customized
- Large scale use for genomic selection (eg > 200 000/y in France)
- Add-on with candidate variants => true genotypes with very limited error rate
- Estimation of allelic frequencies in different breeds
- Very nice tool for major genes/genetic defects/embryonic lethals…, but it is another subject
- Unfortunately, young animals do not have phenotypes
    - Wait several years for performances
    - Backward imputation of animals genotyped with another chip => Low imputation error rate, because of the large reference population
- Confirmation GWAS: do candidate variants explain the QTL?

Content of the custom chip

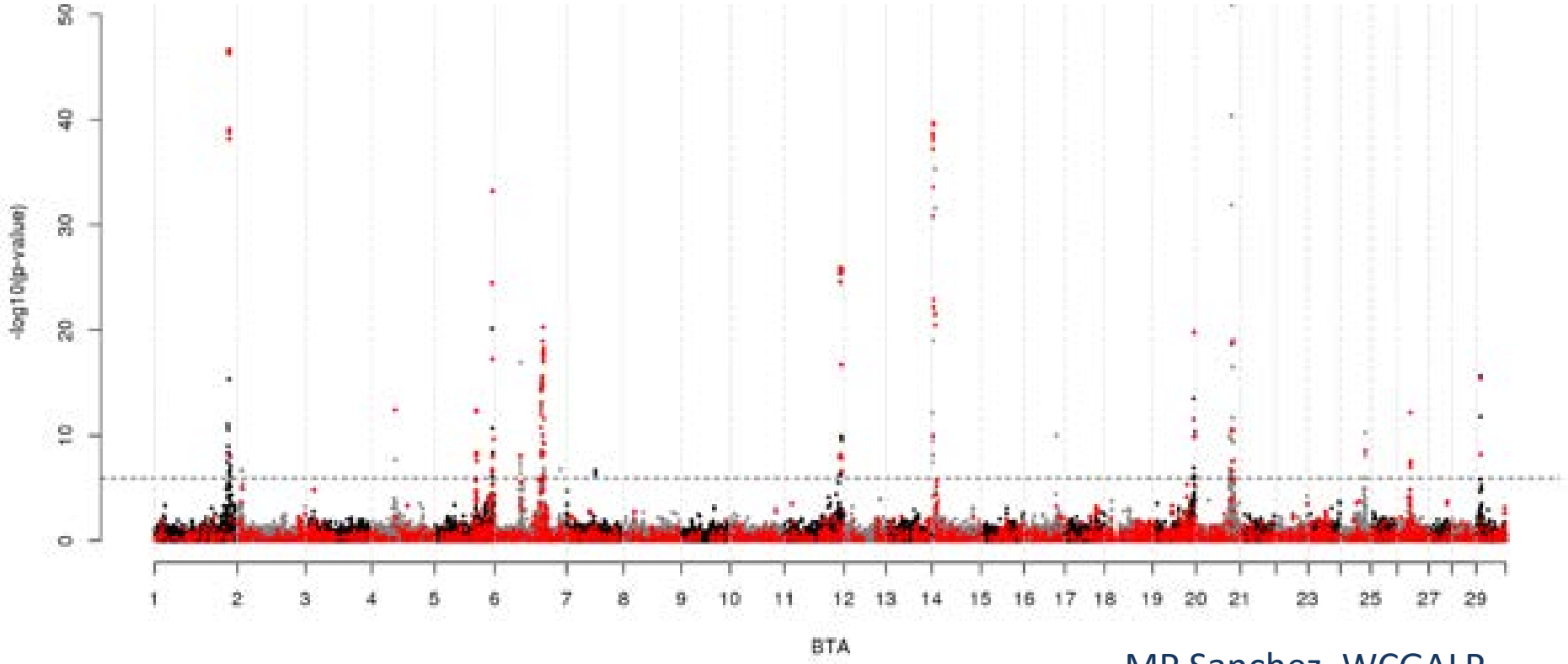
Version 7 presently

- 10k for standard genomic prediction => imputation to 50k
- A number of single gene tests
- A research part for confirmation studies
  - Variants with strong deleterious annotation
  - Variants believed to have a regulatory function
  - Breakpoints of structural variants
  - Peaks of GWAS (~2000 SNP)
    50% from 4 main traits (protein, SCS, fertility, locomotion)
    50% from other traits

- Straightforward switch from research status to production status

Confirmation of SNP

19675 Montbéliarde cows with α_{s1}-casein prediction from MIR spectra



MP Sanchez, WCGALP

Do we improve genomic predictions with these variants

- Tested only within breed, preliminary results on production, udder health and fertility traits
- Not fantastic !
- Some gain with 50k + add-on
- Limited results explained by
  - Good efficiency already achieved
  - Limited number of QTL

Five simulated Scenarios (Croiseau, Wcgalp)

- PEAK: 43,800 most significant SNP, each selected in windows of 300 SNP
- COVER: The genome was divided in 43,800 segments. In each segment, the SNP with a MAF higher than 1% and the lowest p-value was retained.
- COVER2: Same as COVER but the SNP was required to must be located in a gene (if any).
- OPT_QTL: Potential replacement of a 50k marker by the best marker (selected on p-value, MAF) in the same 1Mb-interval.
- Bottom-Up: Replacement of the 50k markers with zero effect by the best sequence SNP

=> **No clear improvement** (Croiseau, WCGALP)

Conclusion

- WGS are here
- Mapping methods give access to the most important variants
- Prediction methods still need to be improved to take full advantage of this information

Acknowledgements

ANR AGENCE NATIONALE DE LA RECHERCHE

APIS-GENE

INRA SCIENCE & IMPACT

EGS ABG

VALOGENE

GenSAP