INRA
SCIENCE & IMPACT

Allice
ÉLEVAGE | INNOVATION | SERVICE

INSTITUT DE L'ELEVAGE

Identification of candidate variants for milk protein composition using sequence data in dairy cattle

Sanchez M.P.[1], Govignon-Gion A.[1], Croiseau P.[1], Barbat A.[1], Gelé M.[2], Fritz S.[3], Miranda G. [1], Martin P.[1], Boussaha M.[1], Brochard M.[2], Boichard D.[1]

[1]INRA, 78350 Jouy en Josas
[2]IDELE, 75012 Paris
[3]ALLICE, 75012 Paris
France

Introduction

***PhénoFinLait** project*
8,080 cows
Milk protein composition
& genotyped 50K beadchip

***1000 bull genomes** project*

whole genome sequences
1,147 bulls (RUN4)

Genome Wide Association Study (**GWAS**)
at the **whole genome sequence** scale

Objectives
Identification of candidate **causal mutations**
for milk protein composition

Material & methods: animals

~ **120,000 cows** with phenotypes
(~ 600,000 test-day milk samples)

8,080 MON, **NOR** & **HOL** cows
genotyped with the 50k Beadchip

2,967	**2,737**	**2,306**
Montbéliardes	Normandes	Holstein
MON	**NOR**	**HOL**

Material & methods: phenotypes

6 major milk proteins:

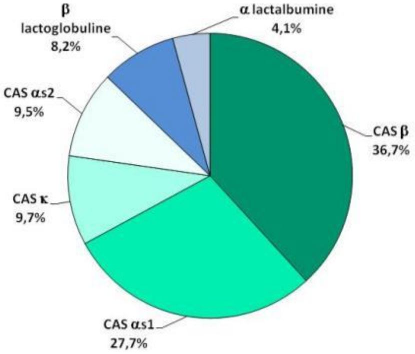
Caseins

αs1, αs2, β & κ

Whey proteins

α lactalbumin

β lactoglobulin



Pre-correction of data for **non genetic** effects

With a mixed model including:

1) Random effects

Animal

Permanent environment

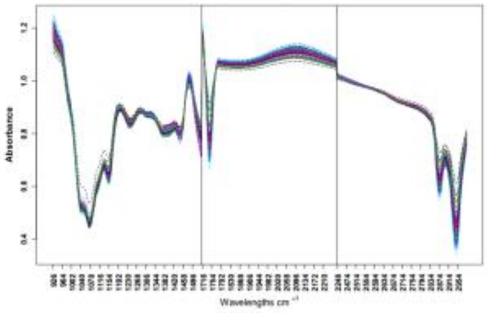
2) Non genetic fixed effects

Herd * test-day

Month * year of calving

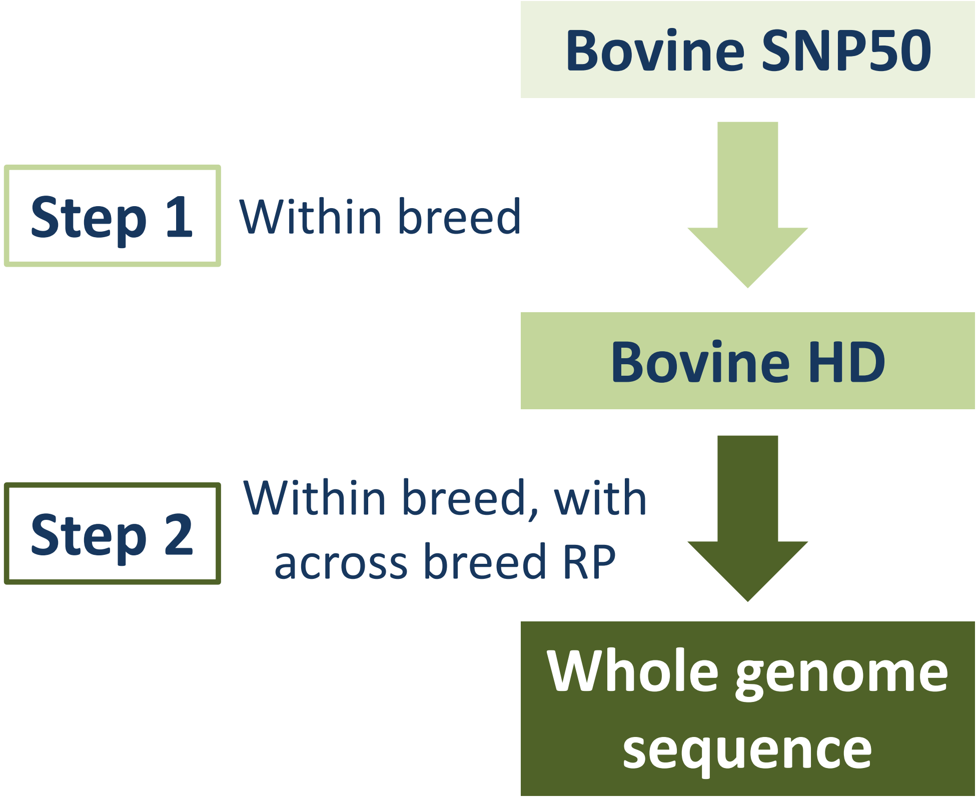
Parity * days in milk

Mid-Infrared **(MIR)** spectra



Material & methods: genotypes & imputation

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



Bovine SNP50

Step 1 Within breed

Bovine HD

Step 2 Within breed, with across breed RP

Whole genome sequence

Reference populations (RP)

3 RP (1 / breed)
1) 522 MON
2) 546 NOR
3) 776 HOL

1 RP for all breeds
= 1,147 bulls

including
28 MON + 24 NOR + 288 HOL

8,080 cows imputed for **27 millions** of sequence variants

GWAS & Bayesian analyses

GWAS *with* ***GCTA*** *(Yang et al., 2011)*

Within breed, **27 millions variants**, **individually** analysed
Polygenic effects of animals with GRM (631,000 HD SNP)

Selection of the most
interesting QTL regions

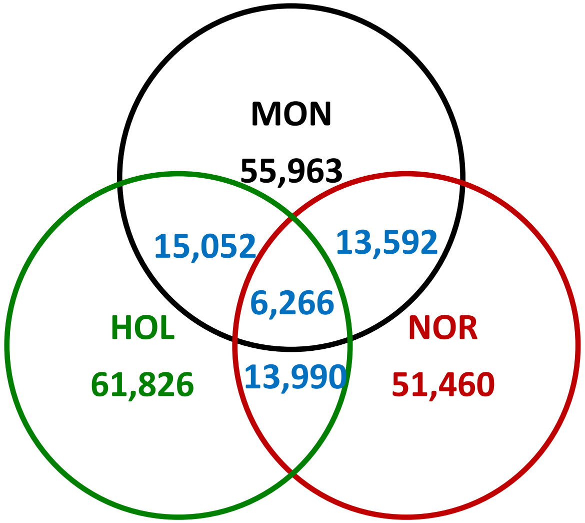
Bayesian analyses *with* ***GS3*** *(Legarra et al., 2013)*

Within breed, **20,000 variants** (~2Mb), **multimarker**
Polygenic effects of animals with pedigree data

Results: GWAS

> 50,000 variants / breed
with relatively low p-value ($< 10^{-6}$)

6,266 variants shared by the 3 breeds



MON 55,963

HOL 61,826

NOR 51,460

15,052

13,592

13,990

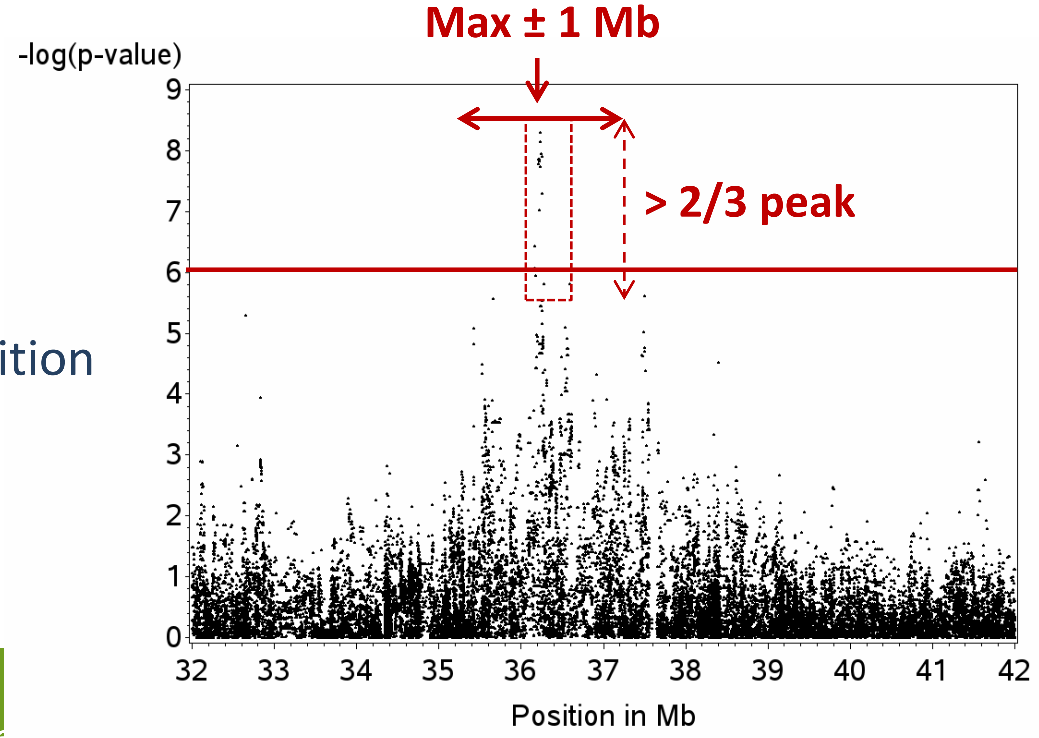
6,266

Definition of QTL regions

$-\log_{10}(\text{p-value}) > 6$

One QTL max in 2 Mb around position of the most sign. variant

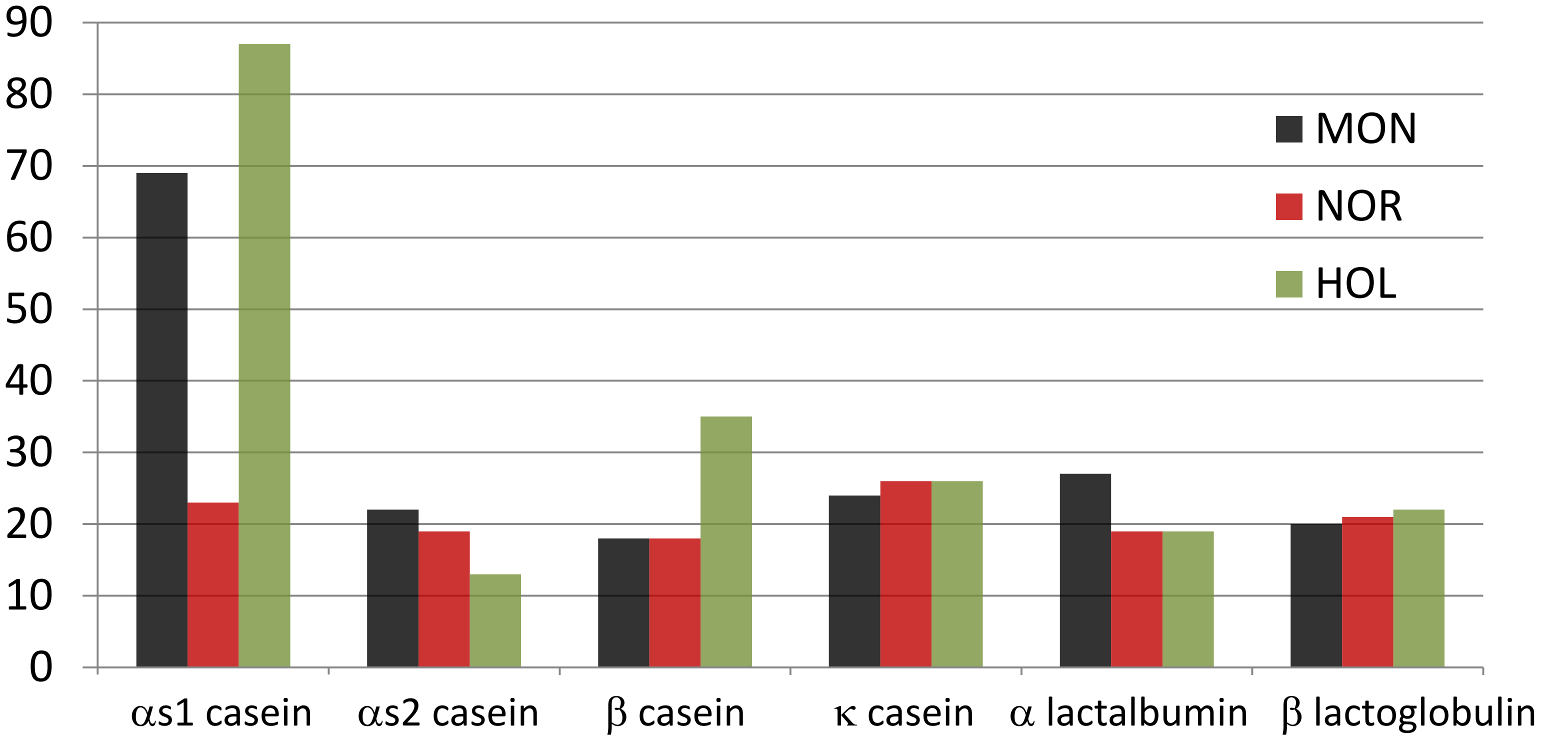
Drop-off value = 1/3 peak



Max ± 1 Mb

> 2/3 peak

Results: GWAS – Number of QTL per trait



From 13 to 87 QTL per trait (max nb for αs1 casein & HOL breed)

GWAS => numerous QTL / breed

Results: GWAS – Most significant QTL

The **most significant** QTL are detected in the 3 breeds

BTA	Trait	Log$_{10}$(1/p) max MON – NOR – HOL	Pos max (kb) MON – NOR – HOL
1	κ casein	**10 – 9 – 13**	**144,402 – 143,555 – 144,471**
2	αs2 casein	**8 – 12 – 7**	**131,809 – 131,807 – 131,711**
6	κ casein	**24 – 22 – 46**	**87,320 – 87,377 – 87,407**
11	β lactoglobulin	**279 – 255 – 226**	**103,289 – 103,301 – 103,298**
20	α lactalbumin	**64 – 44 – 34**	**58,287 – 58,423 – 57,972**
29	αs1 casein	**18 – 8 – 12**	**9,571 – 9,568 – 9,564**

Results: GWAS – Most significant QTL

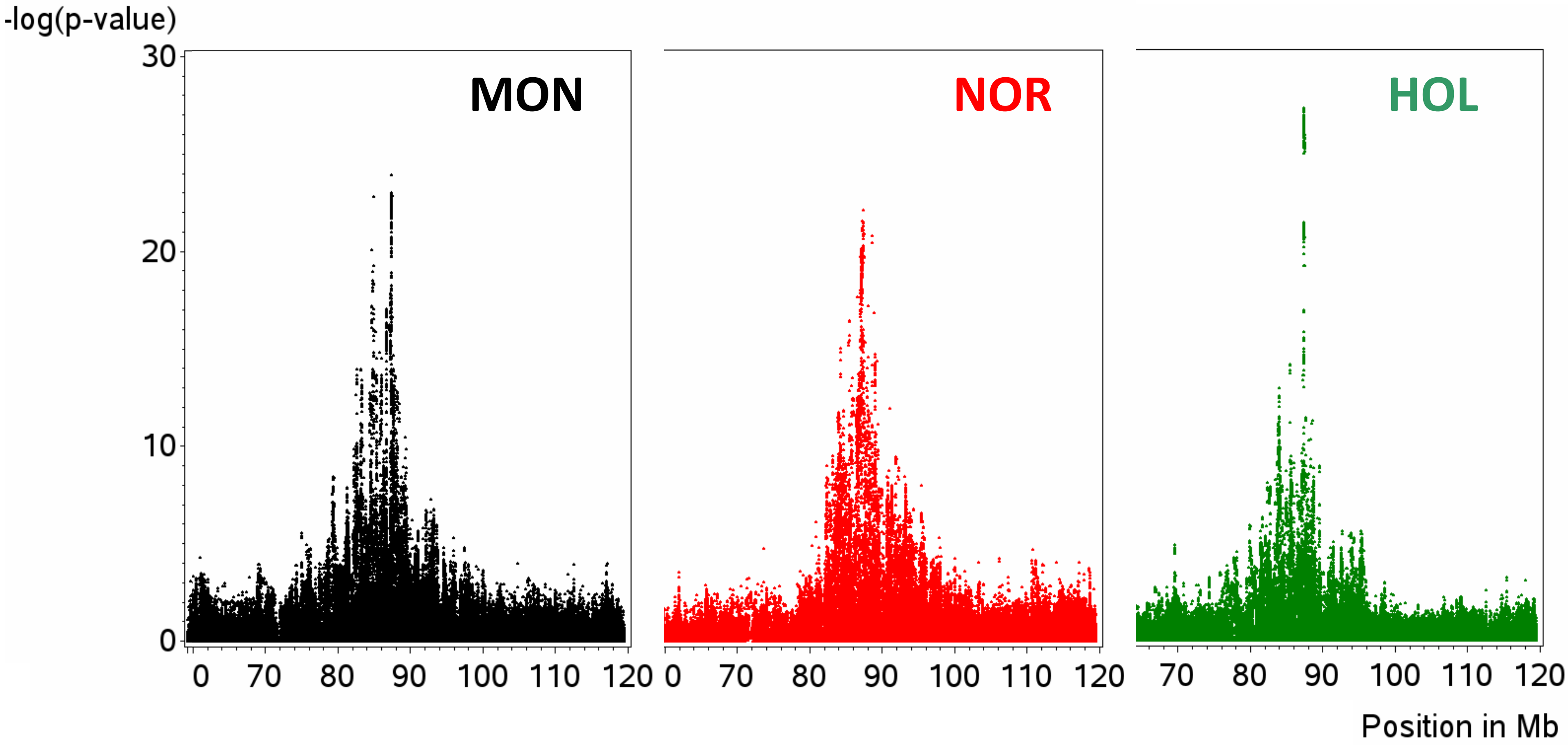
The **most significant** QTL are detected in the 3 breeds

BTA	Trait	$\text{Log}_{10}(1/p)$ max MON – NOR – HOL	Pos max (kb) MON – NOR – HOL
1	κ casein	10 – 9 – 13	144,402 – 143,555 – 144,471
2	αs2 casein	8 – 12 – 7	131,809 – 131,807 – 131,711
6	κ casein	**24 – 22 – 46**	**87,320 – 87,377 – 87,407**
11	β lactoglobulin	**279 – 255 – 226**	**103,289 – 103,301 – 103,298**
20	α lactalbumin	**64 – 44 – 34**	**58,287 – 58,423 – 57,972**
29	αs1 casein	18 – 8 – 12	9,571 – 9,568 – 9,564

3 QTL highly significant in the 3 breeds

Results: GWAS – Most significant QTL detected in 3 breeds

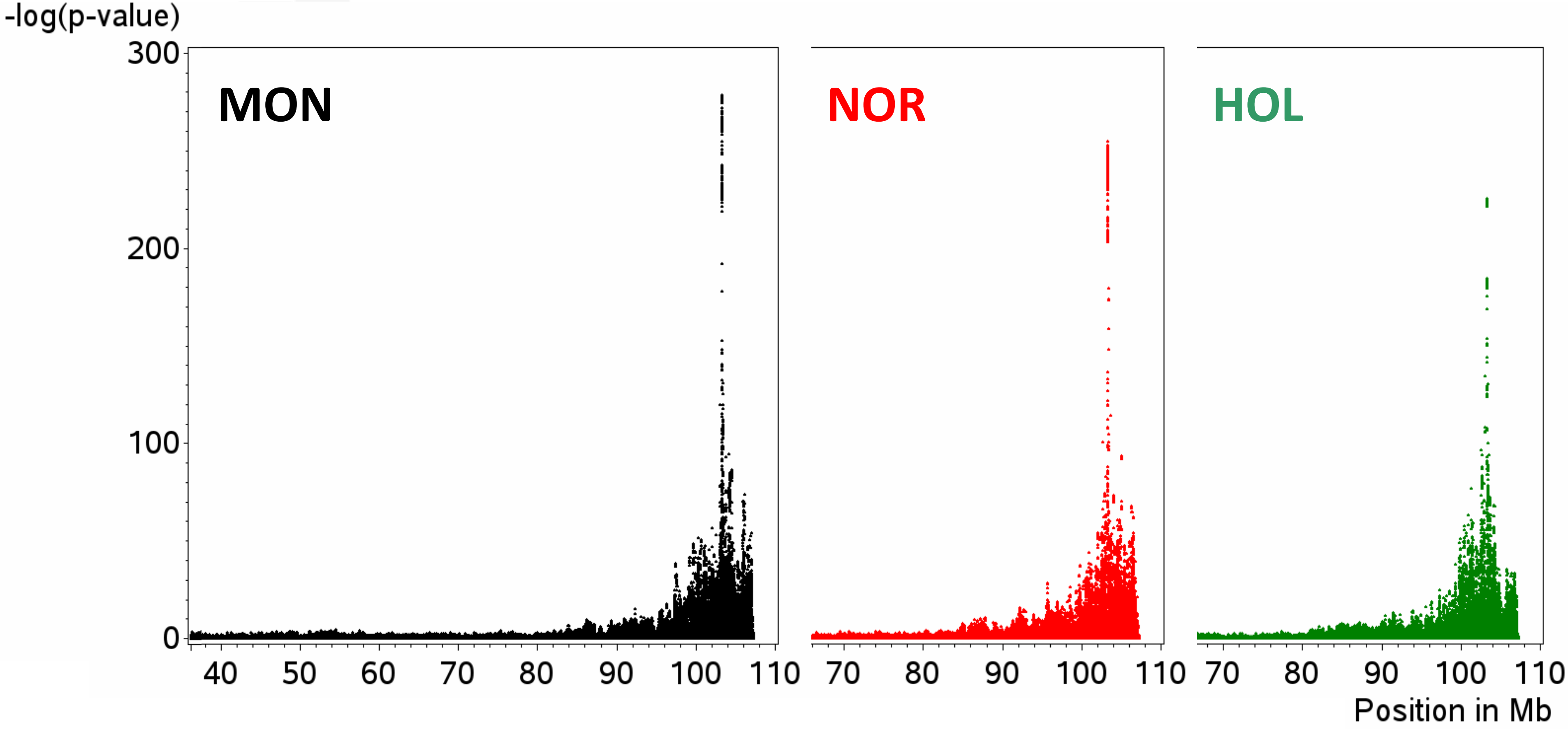
BTA6 & κ-casein



At about **87 Mb** => **casein genes**

Results: GWAS – Most significant QTL detected in 3 breeds

BTA11 & β-lactoglobulin

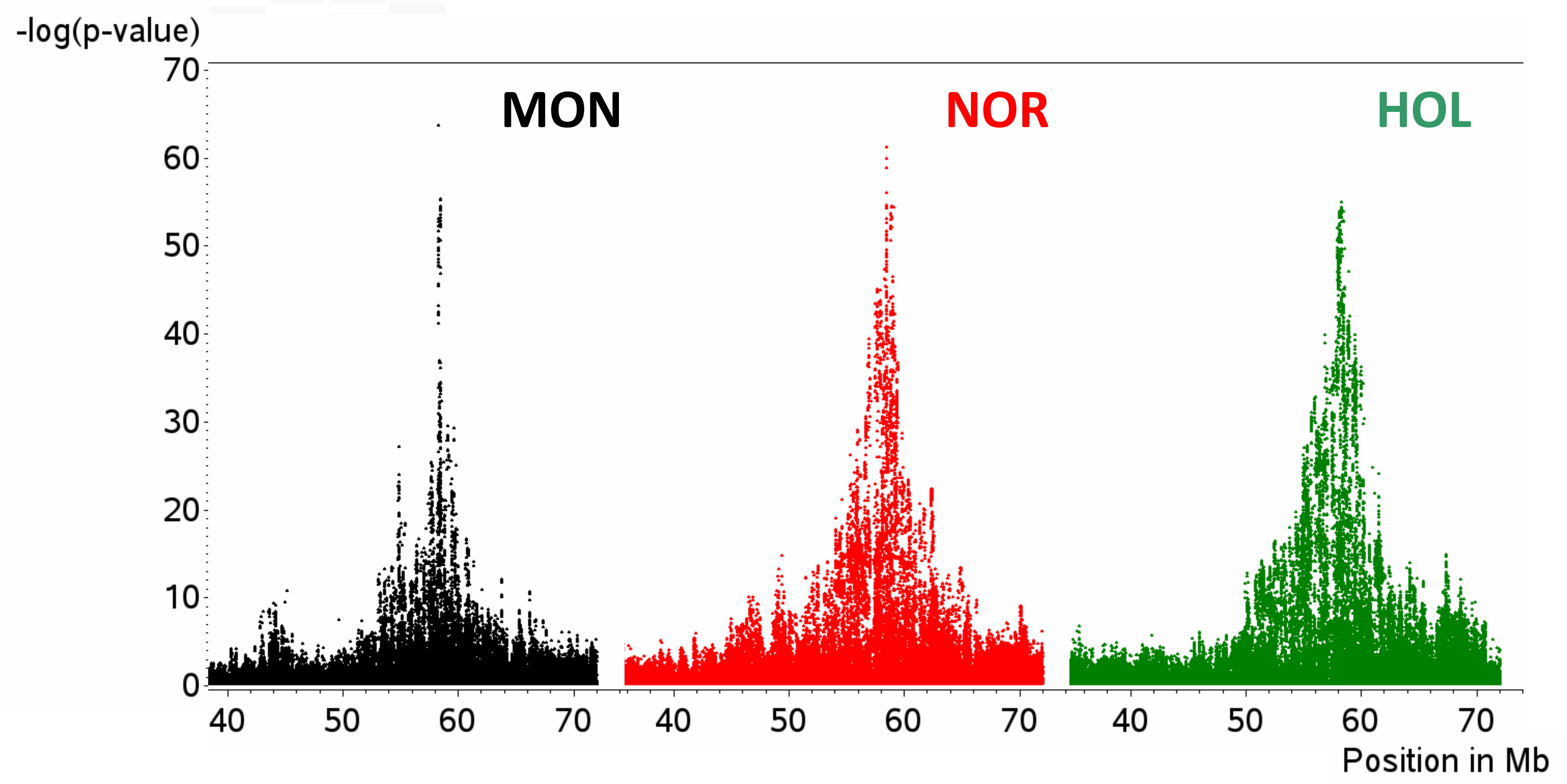


At about **103 Mb** => **LGB gene**

2 known mutations (Ganai et al., 2009) not the most significant

Results: GWAS – Most significant QTL detected in 3 breeds

BTA20 & α-lactalbumin



At about **58 Mb**

Results: GWAS – Most significant QTL

The **most significant** QTL are detected in the 3 breeds

BTA	Trait	Log$_{10}$(1/p) max MON – NOR – HOL	Pos max (kb) MON – NOR – HOL
1	κ casein	**10 – 9 – 13**	**144,402 – 143,555 – 144,471**
2	αs2 casein	**8 – 12 – 7**	**131,809 – 131,807 – 131,711**
6	κ casein	**24 – 22 – 46**	**87,320 – 87,377 – 87,407**
11	β lactoglobulin	**279 – 255 – 226**	**103,289 – 103,301 – 103,298**
20	α lactalbumin	**64 – 44 – 34**	**58,287 – 58,423 – 57,972**
29	αs1 casein	**18 – 8 – 12**	**9,571 – 9,568 – 9,564**

The most significant variants distant from 8 kb to ~ 1 Mb / breed

Question : **is it possible to refine locations of QTL ?**

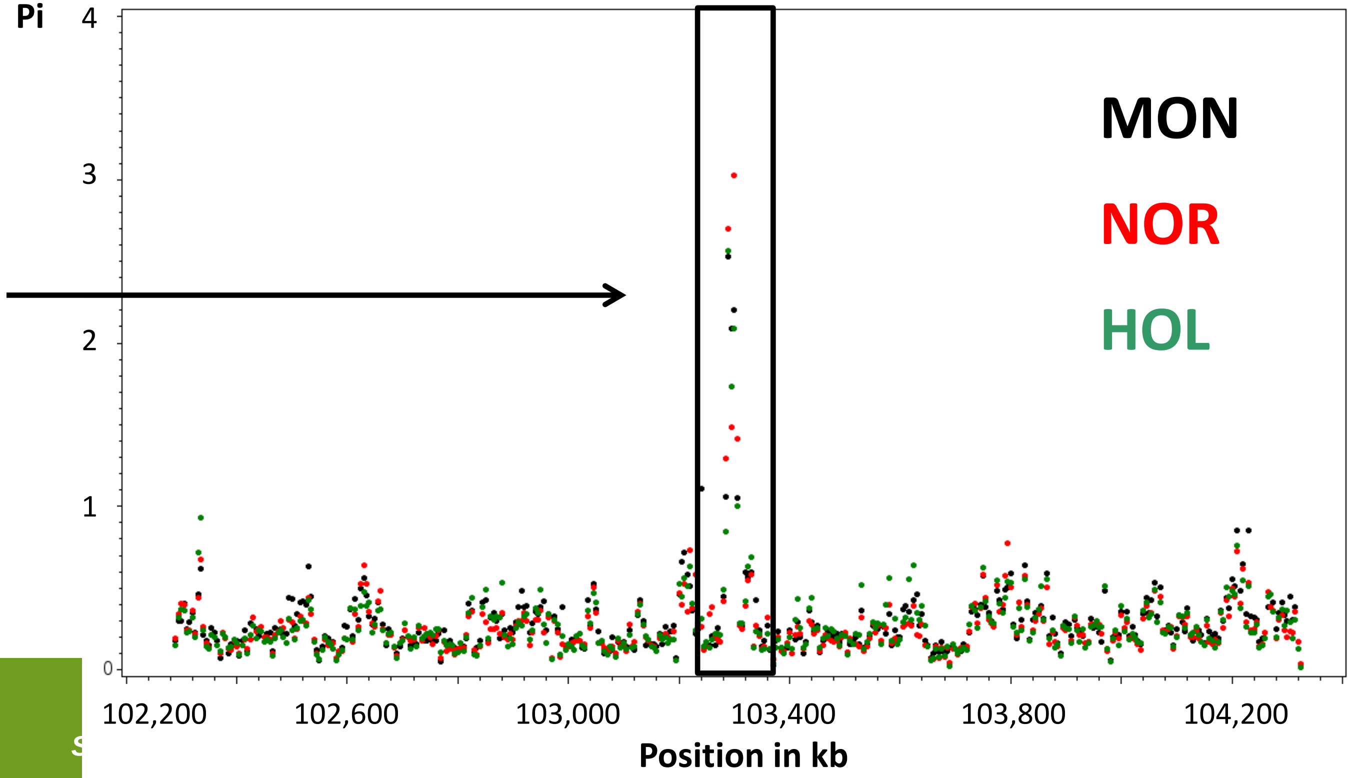
Results: Bayesian analyses to refine position of QTL

Multimarker Bayesian analyses
on the 6 **most significant** QTL regions of ~2 Mb

Candidate variants selected / inclusion probability (Pi)
Pi of a region distributed over linked variants and can be low for individual variants => summed over **5kb-windows**

BTA11
& β lactoglobulin

Selection of finer peaks



Results: Bayesian analyses to refine position of QTL

BTA	Bounds of peak (kb)	Peak in kb	Nb variants Pi > 0.01
1	144,395-144,405	10	39
2	131,810-131,835	25	45
6	87,390-87,410	20	92
11	103,285-103,315	30	101
20	58,410-58,440	30	85
29	9,565-9,580	15	10

Sizes of peaks ranged from **10 to 30 kb**
with a **limited** number of significant variants (10 to 101)

Results: Bayesian analyses to refine position of QTL

BTA	Bounds of peak (kb)	Peak in kb	Nb variants Pi > 0.01	Nb variants in genes	Genes	Annotation of variants in genes & highest Pi
1	144,395-144,405	10	39	30	SLC37A1	**30 intronic**
2	131,810-131,835	25	45	24	ALPL	**1 intronic**
6	87,390-87,410	20	92	40	CSN3	**10 in regulatory regions**
11	103,285-103,315	30	101	45	LGB	**1 missense** (Ganai et al, 2009) **19 in regulatory regions**
20	58,410-58,440	30	85	70	ANKH	**10 intronic**
29	9,565-9,580	15	10	0	-	-

In 5 of the 6 regions, variants with highest Pi located in genes

= **good candidates** for milk protein composition

Results: annotation of variants

+ Candidate causal mutations in **2 genes** encoding κ casein and β lactoglobulin milk proteins

CSN3 10 mutations in regulatory regions

LGB 1 missense mutation (Ganai et al. 2009) + 19 mutations in regulatory regions

+ 3 mutations in intronic regions of ***GPSM1*** (*G-protein signaling modulator 1*) located 500kb-downstream of *LGB*

+ Candidate causal mutations in intronic regions of **3 genes** with function in **milk synthesis** or over expressed in **mammary tissue**

SLC37A1 (30 mut) *glucose 6-phosphate transporter* (Kemper et al., 2015)

ALPL (1 mut) encodes an *alkaline phosphatase* that can dephosphorylate caseins

ANKH (10 mut) *inorganic pyrophosphate transport regulator* (Kemper et al., 2015)

Conclusion

GWAS on imputed **whole genome sequences**
+ **Bayesian analyses**
⇒ Limited number of candidate variants located in genes
It seems a **good approach** to pinpoint **causal mutations**

Our study : **serious candidate mutations** identified
in 5 QTL regions for **milk protein composition**

- ➢ they can be imputed / genotyped to be selected by **genomic selection** in order to improve **techno-functional properties** of milk (cheese yield, milk coagulation time…)

Aknowledgements

To the **PhénoFinlait** consortium

To the **1000 bull genomes** project partners

French sequencing was funded by the French National Agency for Research (**ANR - Cartoseq**) and **Apisgene**

ANR AGENCE NATIONALE DE LA RECHERCHE

APIS-GENE

Thank you for your attention