Using whole genome sequences to identify candidate mutations affecting Milk Fatty Acids in dairy cattle

Govignon-Gion A.¹, Sanchez M.P.¹ Croiseau P.¹, Barbat A.¹, Fritz S.², Boussaha M.¹, Brochard M.³, Boichard D.¹

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

Complex traits are influenced by many QTLs, each explaining a small part of their variability.

QTL full characterization is especially challenging and only a few QTLs have been identified so far, in spite of large efforts.

New tools have become available:
- High throughput genotyping of large populations
- Whole genome sequencing

Is it now possible to re-address the question of QTL identification in a more efficient way?
Introduction

Objectives
Identification of candidate causal mutations for milk fat composition

PhénoFinLait project
8746 cows with milk fat composition & 50k genotypes

« 1000 bull genomes » project
1147 bulls with whole genome sequences (RUN4)

Genome Wide Association Study (GWAS) at the full sequence level
**Material & methods:** 23 Fatty acids estimated by MIR

Mid-Infrared (MIR) spectra

**Pre-correction** of data for non genetic effects
- Herd * test-day
- Month * year of calving
- Parity * days in milk

<table>
<thead>
<tr>
<th>Total SAT</th>
<th>Total UNSAT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C4:0</strong></td>
<td><strong>Total MONO</strong></td>
</tr>
<tr>
<td><strong>C6:0</strong></td>
<td>C18:1cis9</td>
</tr>
<tr>
<td><strong>C8:0</strong></td>
<td>C18:2cis9</td>
</tr>
<tr>
<td><strong>C10:0</strong></td>
<td>C18:2cis9trans11</td>
</tr>
<tr>
<td><strong>C12:0</strong></td>
<td>C18:2cis9cis12</td>
</tr>
<tr>
<td><strong>C14:0</strong></td>
<td>C18:2cis9cis12</td>
</tr>
<tr>
<td><strong>C16:0</strong></td>
<td>C18:3n3</td>
</tr>
<tr>
<td><strong>C18:0</strong></td>
<td>TotC18:1</td>
</tr>
<tr>
<td></td>
<td>TotC18:1cis</td>
</tr>
<tr>
<td></td>
<td>TotC18:1trans</td>
</tr>
<tr>
<td></td>
<td>Omega 3</td>
</tr>
<tr>
<td></td>
<td>Omega 6</td>
</tr>
</tbody>
</table>
Material & methods: animals

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

8746 cows genotyped with the 50k Beadchip

2882 Montbéliardes MON
2816 Normandes NOR
3048 Holstein HOL
Imputation in two steps with *FlImpute* (Sargolzaei et al., 2014)

**Step 1**

Within breed

- Bovine SNP50

**Step 2**

Within breed, with across breed reference

- Bovine HD

Whole genome sequence

**Reference populations**

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

- 1 Reference population
  - = 1147 bulls from « 1000 Bull Genomes »
  - including
  - 28 MON + 24 NOR + 288 HOL

27 millions of sequence variants imputed for 8746 cows
GWAS & Bayesian analyses

Within breed single marker **GWAS with GCTA** (Yang et al., 2011)
**27 millions variants**, analyzed one at a time
Polygenic effects of animals, GRM calculated from HD 631,000 SNP

Selection of the most interesting QTL regions

**Bayesian analyses (BayesC) with GS3** (Legarra et al., 2013)
Within breed, Multimarker (up to 30,000 markers)
Includes also a pedigree-based polygenic effect
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 is distributed over many linked variants, and can be low for individual variants

Inclusion probabilities were summed over 5kb windows to detect the largest signals

Candidate variants were searched within the best windows

Complementary information: (1) Across breed comparison (2) Variant Annotation (1000 bull genomes)
GWAS results

Number of QTLs

- $\log_{10}(p_{\text{value}}) > 6$
- 1 QTL maximum in 2 Mb
- Drop-off value = $\max(2, \frac{2}{3} \text{ peak})$

<table>
<thead>
<tr>
<th>Trait</th>
<th>MON</th>
<th>NOR</th>
<th>HOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4:0</td>
<td>24</td>
<td>30</td>
<td>28</td>
</tr>
<tr>
<td>C12:0</td>
<td>24</td>
<td>18</td>
<td>31</td>
</tr>
<tr>
<td>SAT</td>
<td>23</td>
<td>11</td>
<td>26</td>
</tr>
<tr>
<td>MONO</td>
<td>22</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>$\omega_3$</td>
<td>25</td>
<td>21</td>
<td>44</td>
</tr>
</tbody>
</table>
GWAS results: status of known mutations

**BTA14** DGAT1 HOL
C18:0
1 802 266

**BTA26** – SCD - MON
C12:0
21 144 708

Difficult to conclude:
- Known mutations are not always observed at top
- A mutation may be causative without being at top
  (test = f(â, var(â) )

C18:3n-3 - LGB
103 304 757
103 303 475
## Results: BayesC

### The regions studied

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Région (Mb)</th>
<th>Trait</th>
<th>GWAS test Log$_{10}(1/p)$ max MON – NOR – HOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>92.5-94.5</td>
<td>SAT</td>
<td>14.2 – 13.2 – 24.4</td>
</tr>
<tr>
<td>14</td>
<td>1.3-3.8</td>
<td>SAT</td>
<td>34.4 – 79.9 – 169.8</td>
</tr>
<tr>
<td>17</td>
<td>52.5-55.0</td>
<td>C4:0</td>
<td>30.2 – 47.8 – 12.3</td>
</tr>
<tr>
<td>19</td>
<td>50.0-53.0</td>
<td>C12:0</td>
<td>28.2 – 15.2 – 38.8</td>
</tr>
<tr>
<td>27</td>
<td>36.0-36.5</td>
<td>C16:0</td>
<td>16.8 – 9.3 – 9.7</td>
</tr>
</tbody>
</table>
Results: Chromosome 5, SAT

Sum over 5kb windows

Inclusion probability for each SNP

MON: intergenic

NOR: intergenic

HOL: 4 variants with PI >5% upstream of MGST1

Mon: NOR

HOL: 4 variants with PI >5%
**Results: Chromosome 14, SAT**

**Sum over 5kb windows**

**Inclusion probability for each SNP**

Other mutations than K232A suspected: QTL in populations fixed for K232A and in bulls homozygous for K232A.

This region seems to be rich, with other mutations in DGAT1 (4 ?) and in 2 other genes.
Results: Chromosome 17, 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)
Results: Chromosome 19, C12:0

Sum over 5kb windows

Inclusion probability for each SNP

Several genes involved
6 candidates in the upstream regulatory region of FASN
**Results: Chromosome 27, C16:0**

Sum over 5kb windows

Inclusion probability for each SNP

12 variants very close to each other present PI between 2 and 5% in two breeds. 4 with the highest probabilities are upstream of AGPAT6 and are the best candidates.
## Results: Summary of results

<table>
<thead>
<tr>
<th>BTA</th>
<th>Bounds of peak (kb)</th>
<th>Trait</th>
<th>Candidate variants</th>
<th>Genes</th>
<th>Annotation of variants in genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>93 940-93 955</td>
<td>SAT</td>
<td>4</td>
<td>MGST1</td>
<td>Upstream</td>
</tr>
<tr>
<td>1620-1625</td>
<td>SAT, POLY</td>
<td>1</td>
<td>GPT</td>
<td>3'UTR</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1790-1870</td>
<td>SAT</td>
<td>4</td>
<td>DGAT1</td>
<td>Various</td>
</tr>
<tr>
<td>2700-2720</td>
<td>POLY</td>
<td>4</td>
<td>CYP11B1</td>
<td>Upstream / Downstream</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>53 075-53 085</td>
<td>C4:0</td>
<td>22</td>
<td>BRI3BP</td>
<td>Intronic</td>
</tr>
<tr>
<td>19</td>
<td>51 360-51 385</td>
<td>C12:0</td>
<td>6</td>
<td>FASN</td>
<td>Upstream</td>
</tr>
<tr>
<td>27</td>
<td>36 205-36 220</td>
<td>C16:0</td>
<td>4</td>
<td>AGPAT6</td>
<td>Upstream</td>
</tr>
</tbody>
</table>
Conclusion

BayesC, used to analyze targeted regions, is a good tool to select candidate mutations, in combination with functional annotation.

Across breeds, when QTL co-localize, we observed that the same genes are involved.

But across breed information is weaker than expected to target candidates. A majority of candidate mutations seem to be breed specific.

Acknowledgements

« The 1000 bull genomes » consortium