GenSap Meeting, June 13-14, Aarhus

Genomic Selection with QTL information

Didier Boichard

13-14/06/2013

Introduction

- Few days ago, Daniel Gianola replied on AnGenMap :

  You seem to be suggesting that the QTL paradigm led to genomic selection and to MAS. I could argue that this is not necessarily so in the sense that genomic selection or MAS is just a prediction problem. I would, instead, say that markers led to genomic selection, and that the "QTL paradigm" was largely inconsequential in this process.

- **Is QTL information worth for genomic selection ?**

- **Is it useful I continue …?**

Genomic evaluation properties

- What properties are we looking for ?
  - **High accuracy** => use of all information: relationships, long range LD, and short distance LD
  - **Persistence** over generations, in order to decrease the need for recurrent update of the reference population
  - **Robustness** to low relationship, in order to evaluate individuals in other populations

    => Need to use short distance LD

    => Reintroduces the concept of genes with individual effects (or QTL)

Genomic evaluation with QTL information

- Model targeting some regions in the genome
  - Location, size of these regions
  - Variance explained by these regions
  - Optimized proxy of the causative mutations (fine tuning)
    - Direct use of causative mutations
    - LD maximization
- Need to also account for the residual polygenic value

Genomic evaluation and QTL mapping

- Bayesian approaches such as Bayes C or Bayes R are efficient in both evaluation and QTL mapping
- Multi-SNP

  => preferential use of small distance LD information

  => lead to reduced mapping interval
- The sum of SNP inclusion probabilities over a given interval can be used to map a QTL
- « Customized » variances in BayesA, BayesB, BayesR

Haplotypes are more informative than individual SNP

- Two alleles, limited information
- Selection on SNP informativity (=> likely more polymorphic than causative variants)
- Their polymorphisms are old (=> on average older than the causative variants)
- Likely incomplete short distance LD with causative variants
- Theoretical advantage to haplotypes,
  - which are more informative
  - likely in higher short distance LD

How to use Haplotypes ?

- To measure relationships
  - IBS haplotypes are more likely IBD than IBS SNP
  - Haplotype-based GBLUP is more accurate than SNP-based GBLUP, through a better measure of relationships
- In a model fitting haplotype effects
  - Many effects to estimate

    => Need for a strong selection => QTL model

$$y_i = \mu + u_i + \sum_{j=1}^{nqtl} (h_{ij}^s + h_{ij}^d) + e_i$$

How to select QTL ?

- By conventional QTL detection ?
  - Low detection power => only the largest QTL are detected
  - Their variance is overestimated
  - They explain only a small proportion of the total genetic variance
- By SNP selection
  - Elastic Net
  - Bayesian methods

How many QTL ?

- A large number to explain a large proportion of the genetic variance
- No general rules
- Usually several hundreds of QTL to explain >50% variance
- Most of them explain a very small variance
- Hard to detect => Most of them are not well defined QTL in the usual sense, but regions with some predictive ability
- Mixture of a few large QTL well characterized, some medium size QTL, many small QTL

What we apply in the French dairy evaluation

- A QTL model, with QTLs and a residual polygenic effect
- With 300-700 QTL per trait
- Each QTL is traced by 3-4 SNP in a <1cM interval
- SNP were selected by EN and then neighbor SNP were grouped into the same haplotype.
- Additional neighbor SNP were added if needed, for a minimum of 3 SNP / haplotype

Efficiency

Correlation between GEBV and DYD in the validation Holstein population

	Milk	Protein	Fat	Prot %	Fat %	Fertility
BLUP	0.38	0.44	0.40	0.47	0.44	0.29
GBLUP	0.56	0.55	0.59	0.73	0.72	0.35
PLS	0.53	0.55	0.58	0.71	0.70	0.33
Elastic Net	0.57	0.57	0.63	0.75	0.80	0.34
QTL-BLUP	0.60	0.57	0.66	0.73	0.81	0.39

Efficiency – Intergenomics data set

- 7041 bulls
- 5 countries: CHE, DEA, FRA, ITA and USA
- 10 traits

	Correlation	Slope deviation from 1	# traits validated by Interbull
GBLUP	0.502	0.182	6.4
B-LASSO	0.533	0.110	8.0
BAYES C Pi	**0.537**	0.109	8.0
QTL-BLUP	0.517	**0.104**	**8.4**

Some constraints to use QTL and haplotypes

- Selection work for each trait in each population
- Variance estimation
- No missing marker => imputation
- Known phases (less limiting since LD chip use and imputation)

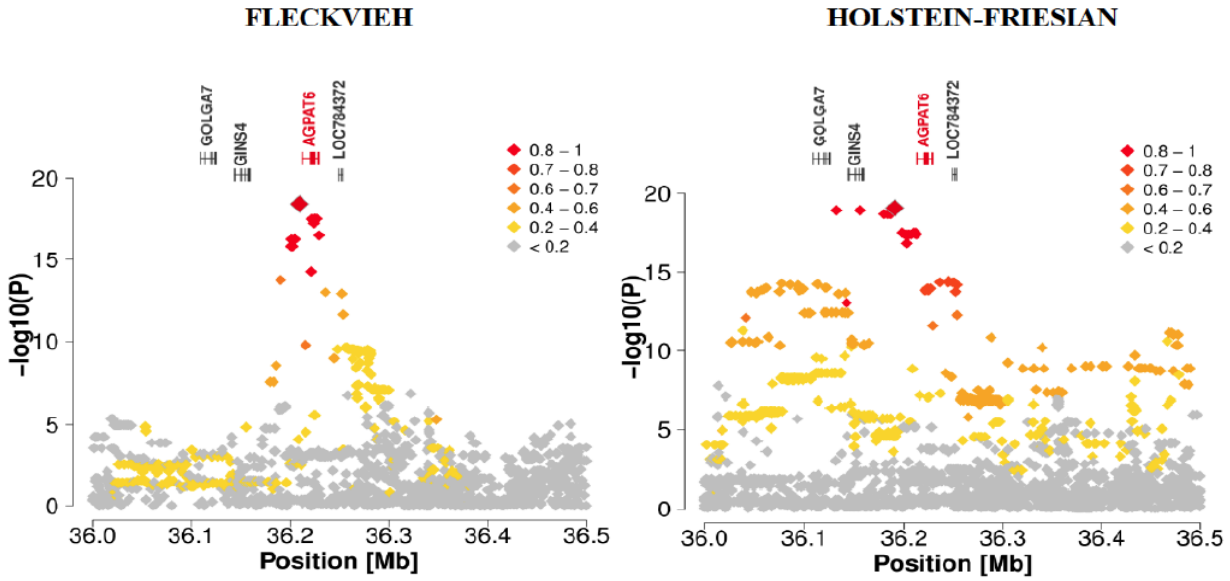
Across breed evaluation

- G BLUP cannot accurately predict a candidate of breed A from a reference population from breed B – no long distance LD
- Use of High Density chip (HD=777k, 1 SNP every 4 kb)
- Length of conserved segments across breeds: 10-20 kb
- Idea:
  - Apply a QTL model with haplotypes of 2-4 markers
  - Define two QTL categories: within breed or shared across breeds (because not all QTL are shared across populations)

$$y_{ij} = \mu + u_{ij} + \sum_{q=1}^{nqs} (h^s_{iq} + h^d_{iq}) + \sum_{b=1}^{nbreed} \sum_{q=1}^{nqw} (h^s_{biq} + h^d_{biq}) + e_i$$

Discovery of causative mutations

- Imputation of reference populations up to the sequence
- GWAS on real or imputed genotypes
- We can expect a large number of candidate causative mutations in the near future



(Fries et al)

Use of causative mutations

- Obtain the genotypes of candidates
  - Either through imputation
  - Or by direct genotyping with a custom chip

    The example of the EuroG10k Illumina chip,

    with ~150 candidate mutations presently,

    updated every 6 months to incorporate new discoveries
- Confirm the effect of these mutations with large scale female reference populations
- Include them into the model – straightforward with a QTL model
  - LD is maximized !
  - More persistent effect across populations and backgrounds
  - More easily allows for more complex modeling (interaction effects) ?

Conclusion : some feelings

- A QTL-based model allows for fine tuning to account for individual regions of importance
- It is more efficient with haplotypes than with single SNP to take advantage of short distance LD
- It is likely less efficient to account for residual polygenic effects
- It is more adapted to across populations evaluations than GBLUP
- It can easily incorporate causative mutations

THANK YOU FOR YOUR ATTENTION !