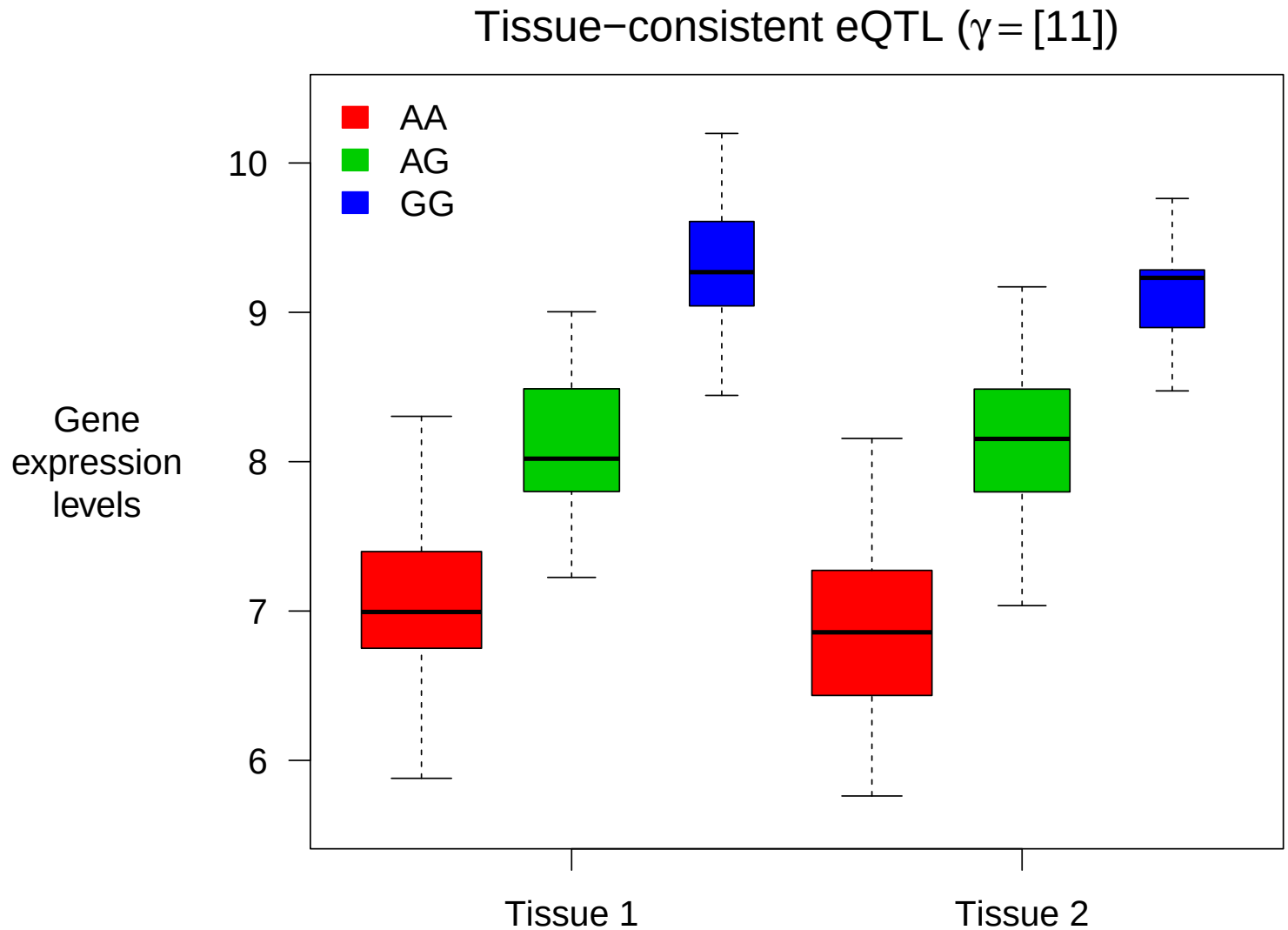
A statistical framework for eQTL analysis among multiple tissues

Timothée Flutre

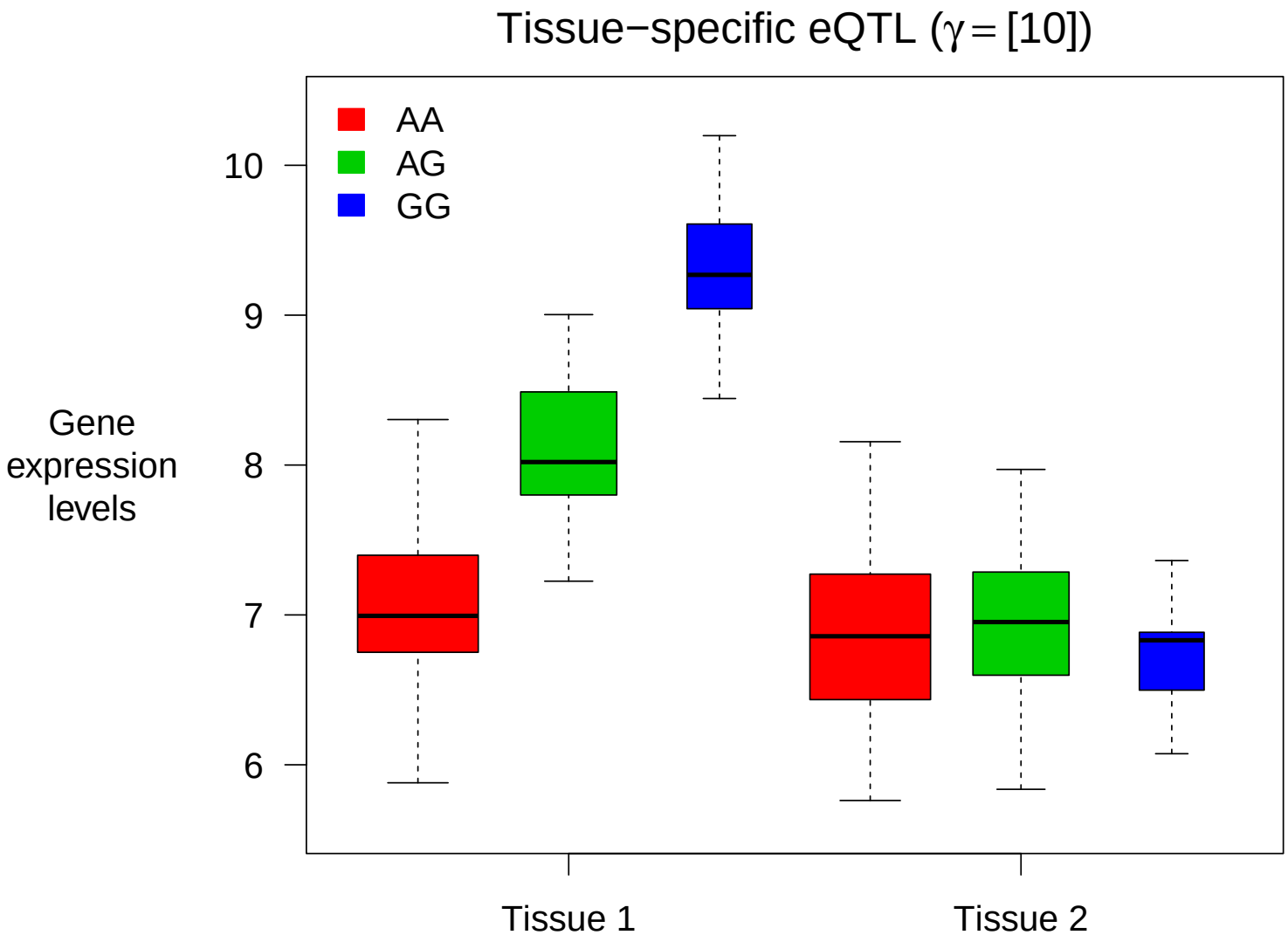
UChicago (Human Genetics) - INRA (Plant Genetics)

November 9, 2012 (ASHG, San Francisco)

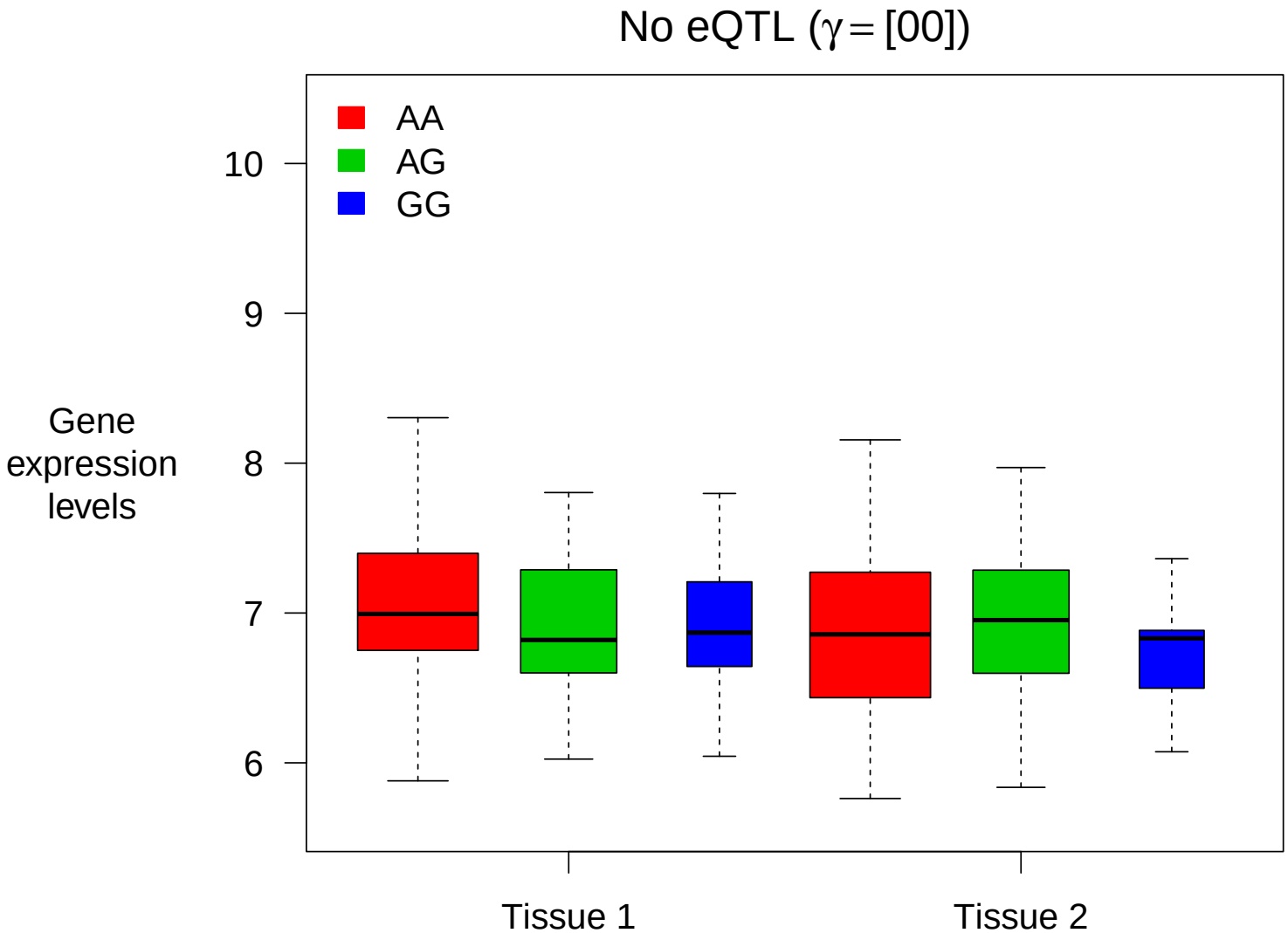
Examples of eQTLs with two tissues



Examples of eQTLs with two tissues



Examples of eQTLs with two tissues



Intuitions to analyse eQTLs in multiple tissues

Each tissue separately:

- fails to leverage commonalities between tissues
- seems easier to investigate heterogeneity

Intuitions to analyse eQTLs in multiple tissues

Each tissue separately:

- fails to leverage commonalities between tissues
- seems easier to investigate heterogeneity

All tissues jointly:

- allows to borrow information across tissues
- seems harder to identify tissue-specific eQTLs

Intuitions to analyse eQTLs in multiple tissues

Each tissue separately:

- fails to leverage commonalities between tissues
- seems easier to investigate heterogeneity

All tissues jointly:

- allows to borrow information across tissues
- seems harder to identify tissue-specific eQTLs

Trade-off depends on the amount of tissue-specific eQTLs and noise.

Intuitions to analyse eQTLs in multiple tissues

Each tissue separately:

- fails to leverage commonalities between tissues
- seems easier to investigate heterogeneity

All tissues jointly:

- allows to borrow information across tissues
- seems harder to identify tissue-specific eQTLs

Trade-off depends on the amount of tissue-specific eQTLs and noise.

Two main goals:

- detect eQTLs in any tissue (hypothesis testing)
- identify in which tissue(s) they are active (model comparison)

Linear regression and configurations

For each gene-SNP pair, in tissue s of individual i:

- $y_{si} = \mu_s + \beta_s g_i + \epsilon_{si}$ with $\epsilon_{si} \sim \mathcal{N}(0, \sigma_s^2)$
- errors allowed to be correlated between tissues

Linear regression and configurations

For each gene-SNP pair, in tissue s of individual i:

- $y_{si} = \mu_s + \beta_s g_i + \epsilon_{si}$ with $\epsilon_{si} \sim \mathcal{N}(0, \sigma_s^2)$
- errors allowed to be correlated between tissues

Configurations represent tissue consistency/specificity:

- "1" for active eQTL ($\beta_s \neq 0$), "0" otherwise
- $\gamma =$[110] corresponds to an eQTL in the first two tissues

Linear regression and configurations

For each gene-SNP pair, in tissue s of individual i:

- $y_{si} = \mu_s + \beta_s g_i + \epsilon_{si}$ with $\epsilon_{si} \sim \mathcal{N}(0, \sigma_s^2)$
- errors allowed to be correlated between tissues

Configurations represent tissue consistency/specificity:

- "1" for active eQTL ($\beta_s \neq 0$), "0" otherwise
- $\gamma =$[110] corresponds to an eQTL in the first two tissues

References: Wen & Stephens (2011, arXiv), Wen (2012, arXiv), Han & Eskin (AJHG, 2011)

Bayesian Model Averaging and hierarchical modeling

- Bayes Factor as support for an eQTL in configuration γ:

$$\text{BF}_\gamma = \frac{\text{P(data} \mid \text{eQTL in configuration } \gamma)}{\text{P(data} \mid \text{no eQTL in any tissue)}}$$

Bayesian Model Averaging and hierarchical modeling

- Bayes Factor as support for an eQTL in configuration γ:

$$\text{BF}_\gamma = \frac{\text{P(data} \mid \text{eQTL in configuration } \gamma)}{\text{P(data} \mid \text{no eQTL in any tissue)}}$$

- Measure overall evidence against the global null hypothesis:

$$\text{BMA} = \sum_\gamma \eta_\gamma \text{BF}_\gamma$$

Bayesian Model Averaging and hierarchical modeling

- Bayes Factor as support for an eQTL in configuration γ:

$$\text{BF}_\gamma = \frac{\text{P(data} \mid \text{eQTL in configuration } \gamma)}{\text{P(data} \mid \text{no eQTL in any tissue)}}$$

- Measure overall evidence against the global null hypothesis:

$$\text{BMA} = \sum_\gamma \eta_\gamma \text{BF}_\gamma$$

- Estimate configuration proportions η_γ with a hierarchical model which borrows information across genes (pooling).

Bayesian Model Averaging and hierarchical modeling

- Bayes Factor as support for an eQTL in configuration γ:

  $$\text{BF}_\gamma = \frac{\text{P(data} \mid \text{eQTL in configuration } \gamma)}{\text{P(data} \mid \text{no eQTL in any tissue)}}$$

- Measure overall evidence against the global null hypothesis:

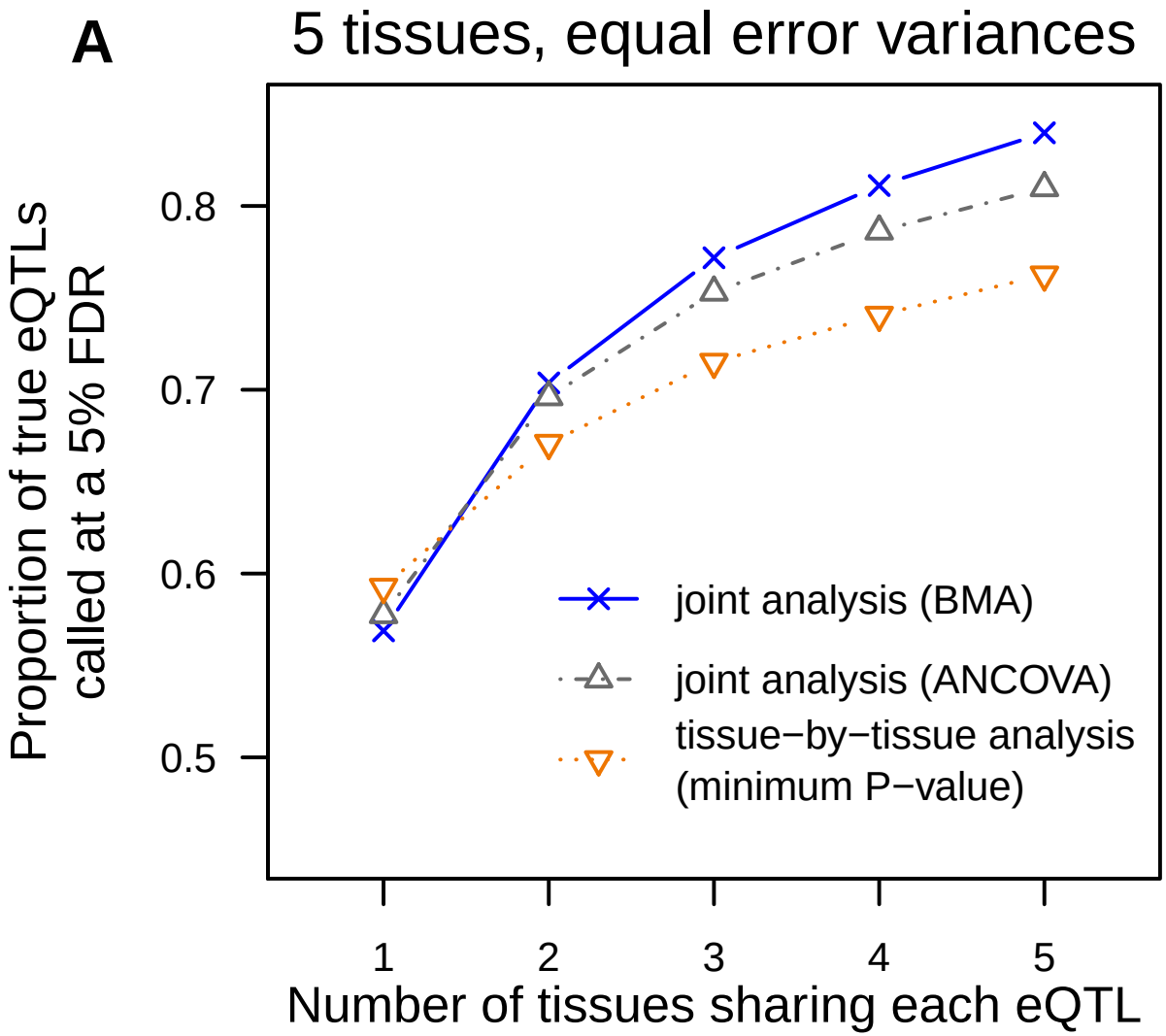
  $$\text{BMA} = \sum_\gamma \eta_\gamma \text{BF}_\gamma$$

- Estimate configuration proportions η_γ with a hierarchical model which borrows information across genes (pooling).

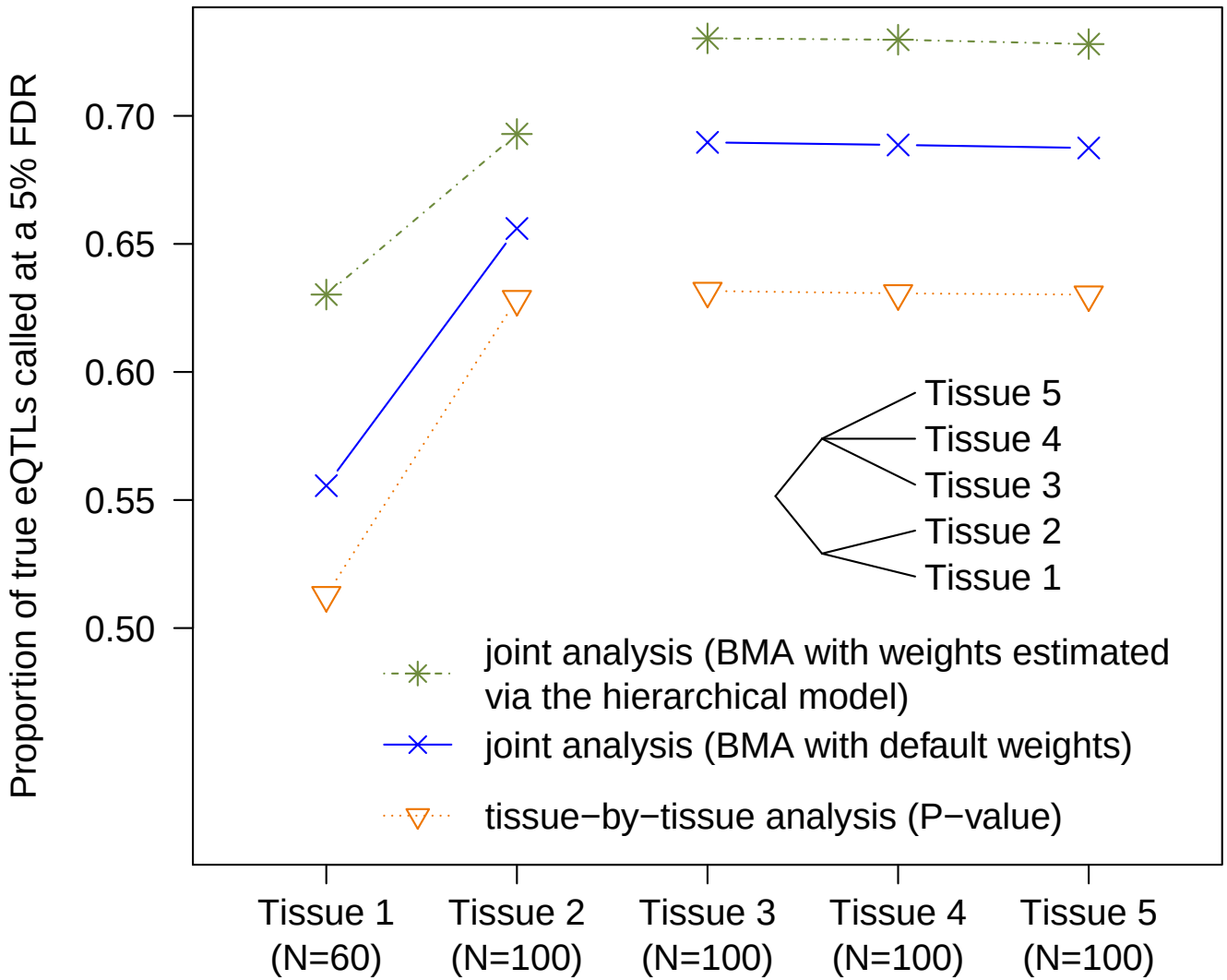
- Posterior probability to interpret the associations:

  $$\text{P(SNP is in configuration } \gamma \mid \text{data, SNP is eQTL}) = \frac{\eta_\gamma \text{BF}_\gamma}{\sum_\gamma \eta_\gamma \text{BF}_\gamma}$$

Simulations - power gain and borrowing of information



Simulations - power gain and borrowing of information



Analysis of the data set from Dimas *et al.* (2009)

- 3 cell types: Fibroblasts, LCLs and T-cells

Analysis of the data set from Dimas *et al.* (2009)

- 3 cell types: Fibroblasts, LCLs and T-cells
- 75 unrelated individuals (GenCord project)

Analysis of the data set from Dimas *et al.* (2009)

- 3 cell types: Fibroblasts, LCLs and T-cells
- 75 unrelated individuals (GenCord project)
- $\approx$ 400,000 SNPs

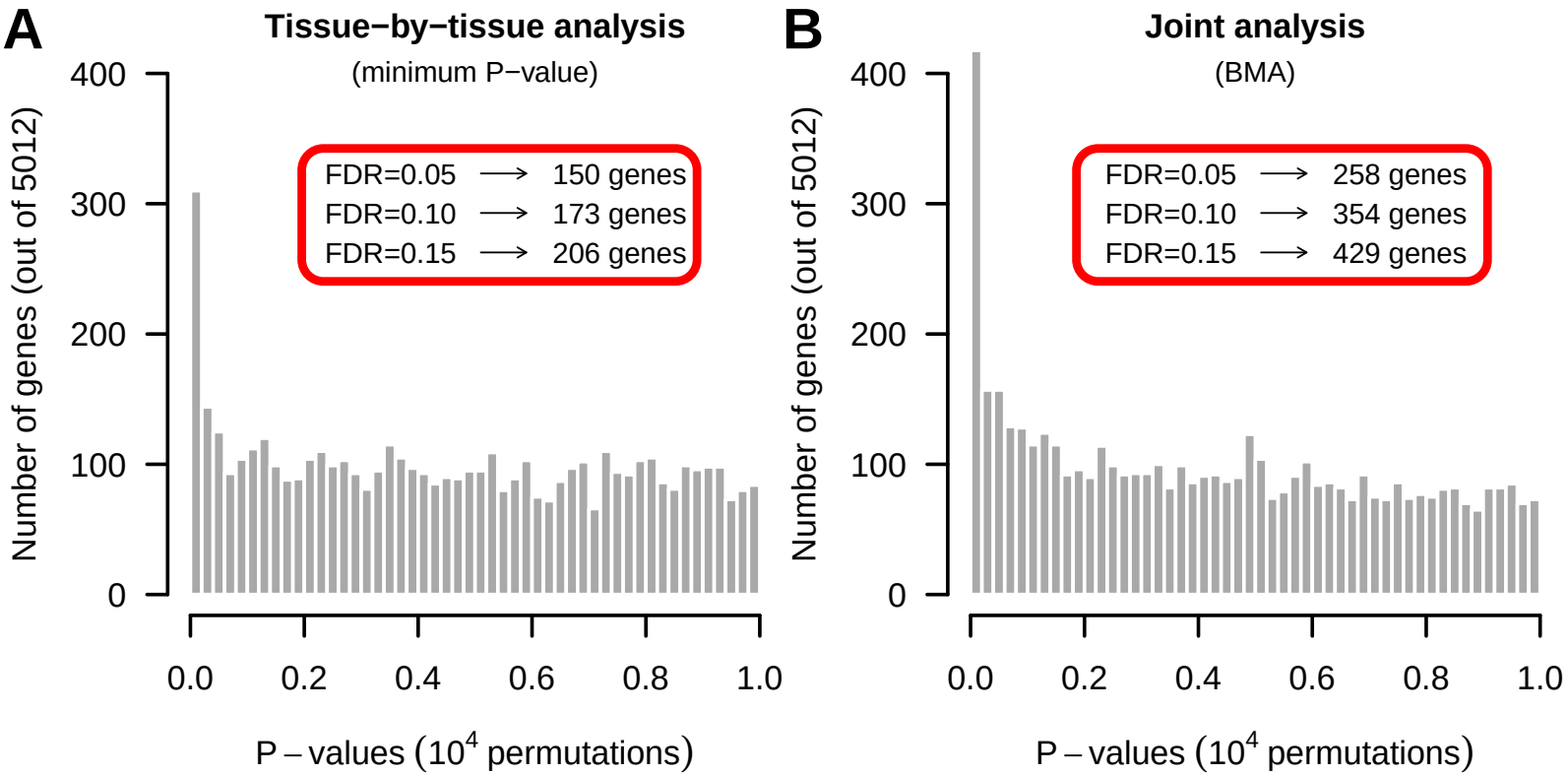
Analysis of the data set from Dimas *et al.* (2009)

- 3 cell types: Fibroblasts, LCLs and T-cells
- 75 unrelated individuals (GenCord project)
- $\approx$ 400,000 SNPs
- $\approx$ 5000 genes deemed expressed in all three cell types

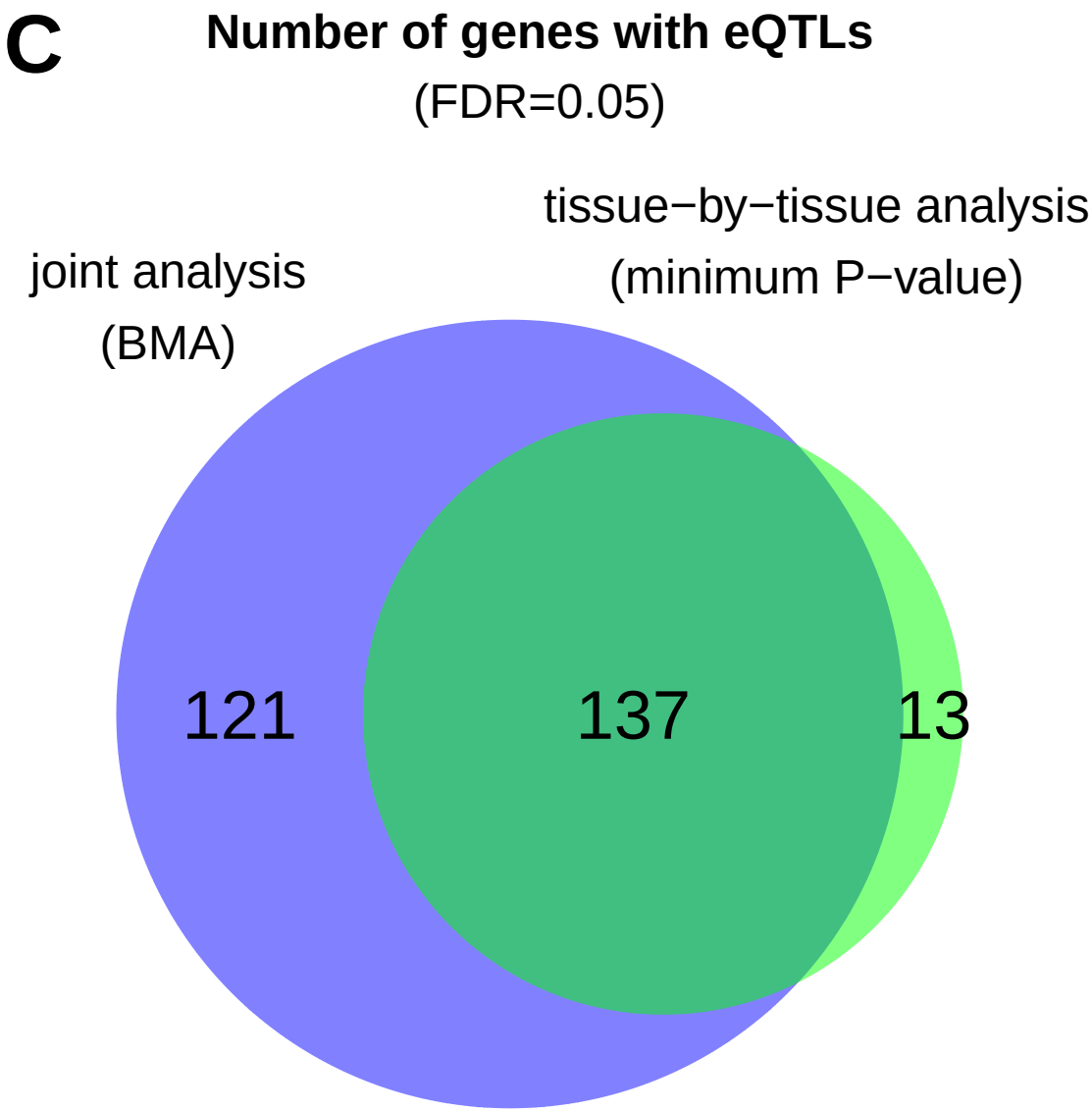
Analysis of the data set from Dimas *et al.* (2009)

- 3 cell types: Fibroblasts, LCLs and T-cells
- 75 unrelated individuals (GenCord project)
- $\approx$ 400,000 SNPs
- $\approx$ 5000 genes deemed expressed in all three cell types
- *cis* region: $\pm$ 1 Mb from the TSS

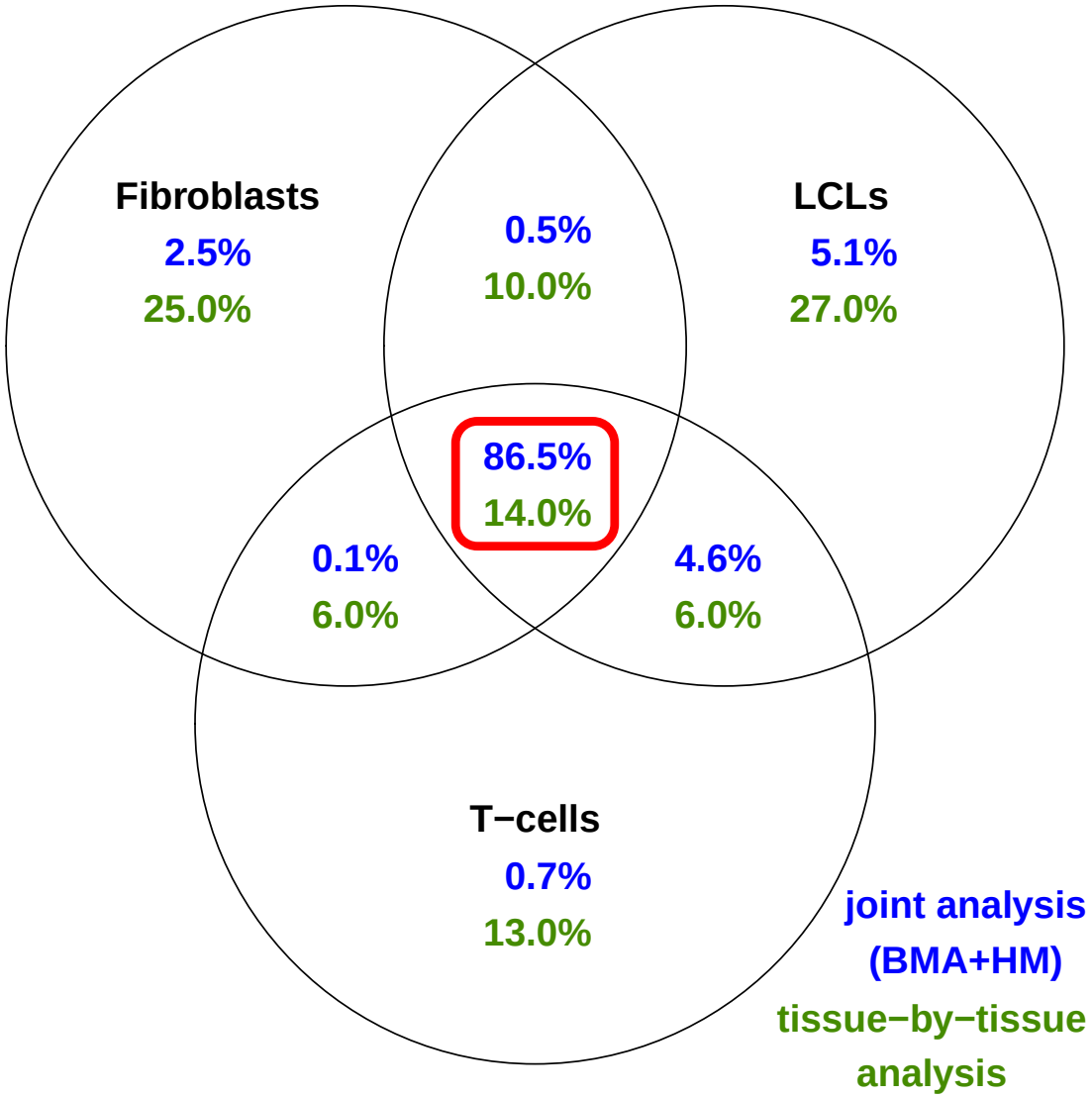
Gain in power from the joint analysis



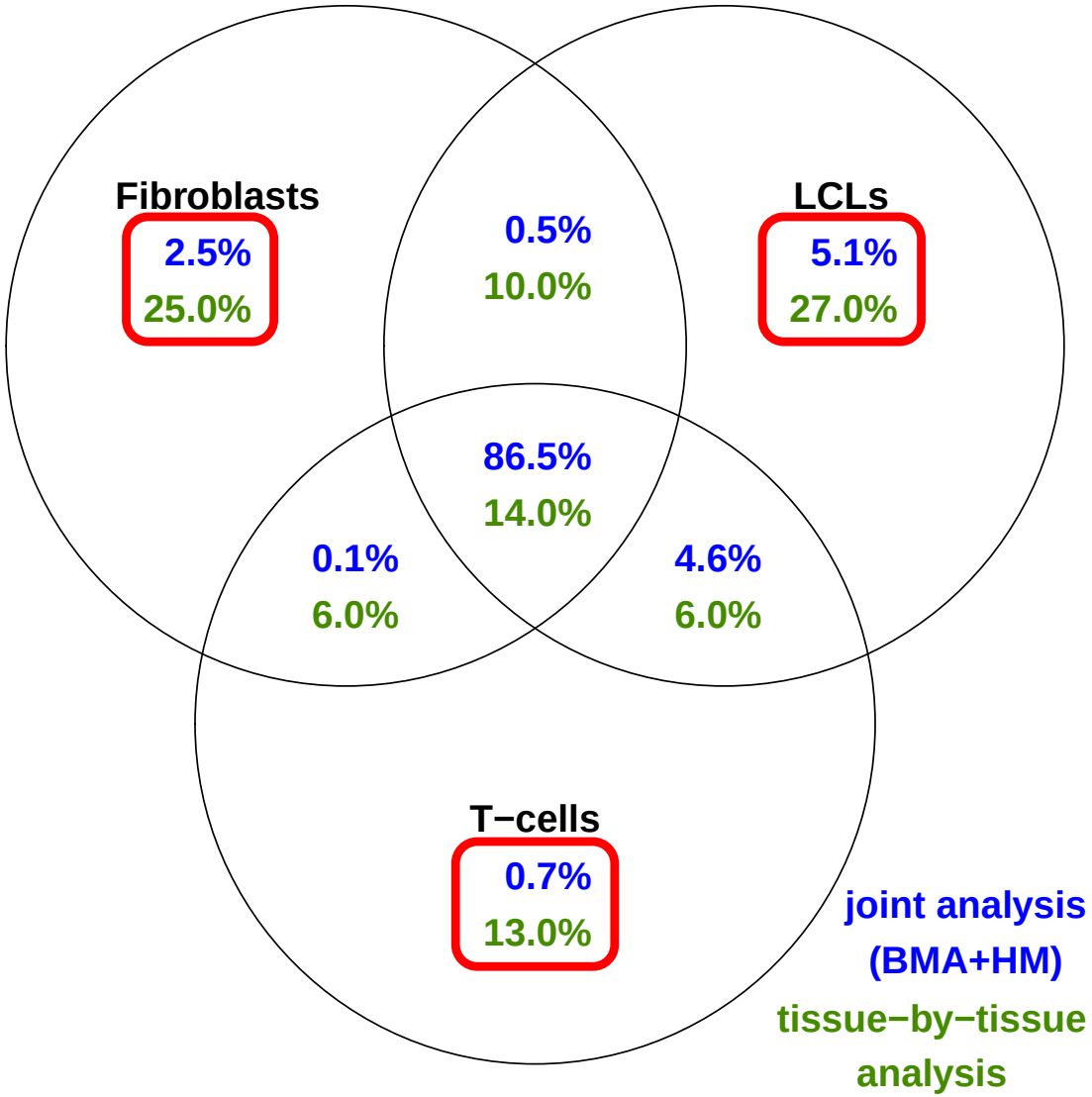
Gain in power from the joint analysis



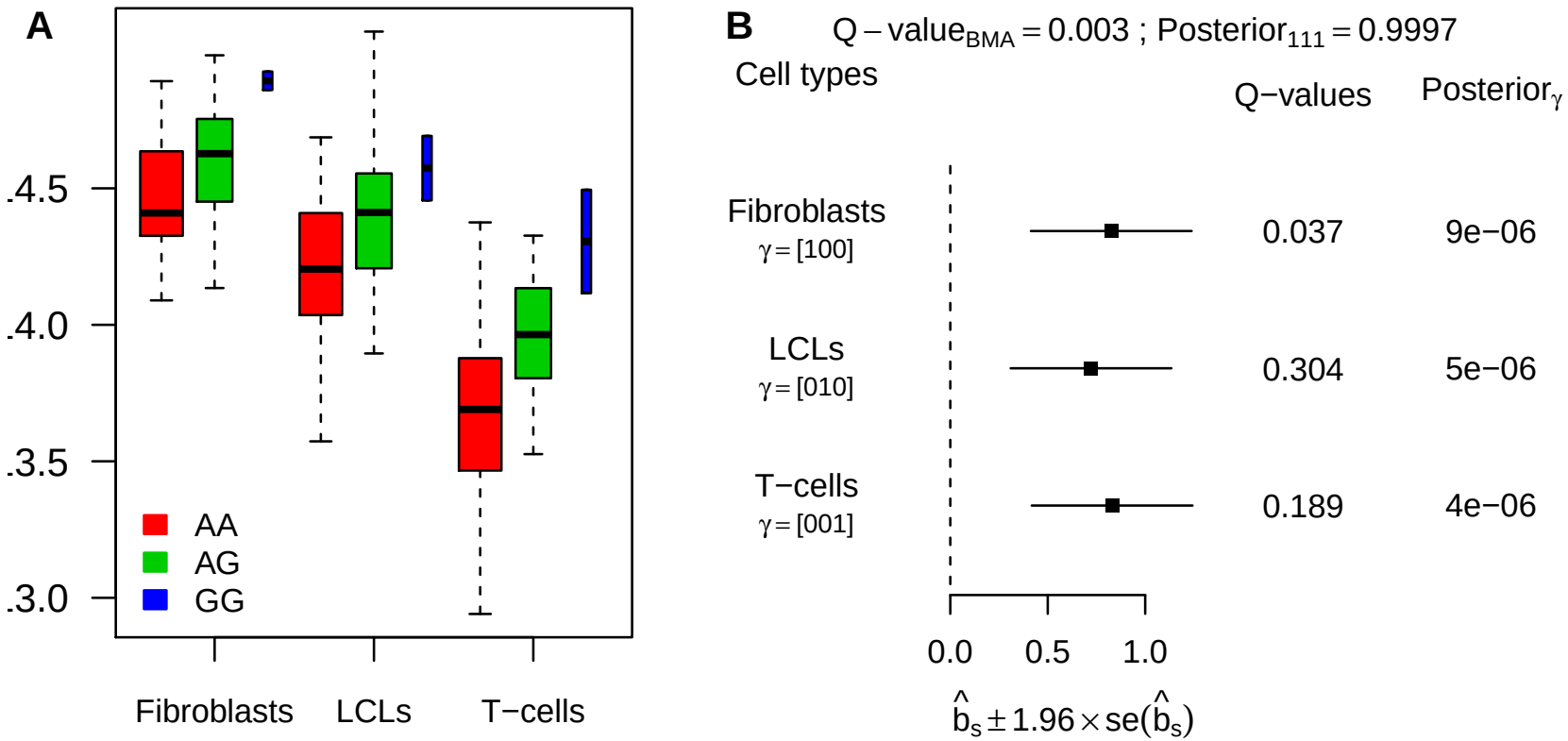
Reliable inference of the proportion of tissue specificity



Reliable inference of the proportion of tissue specificity

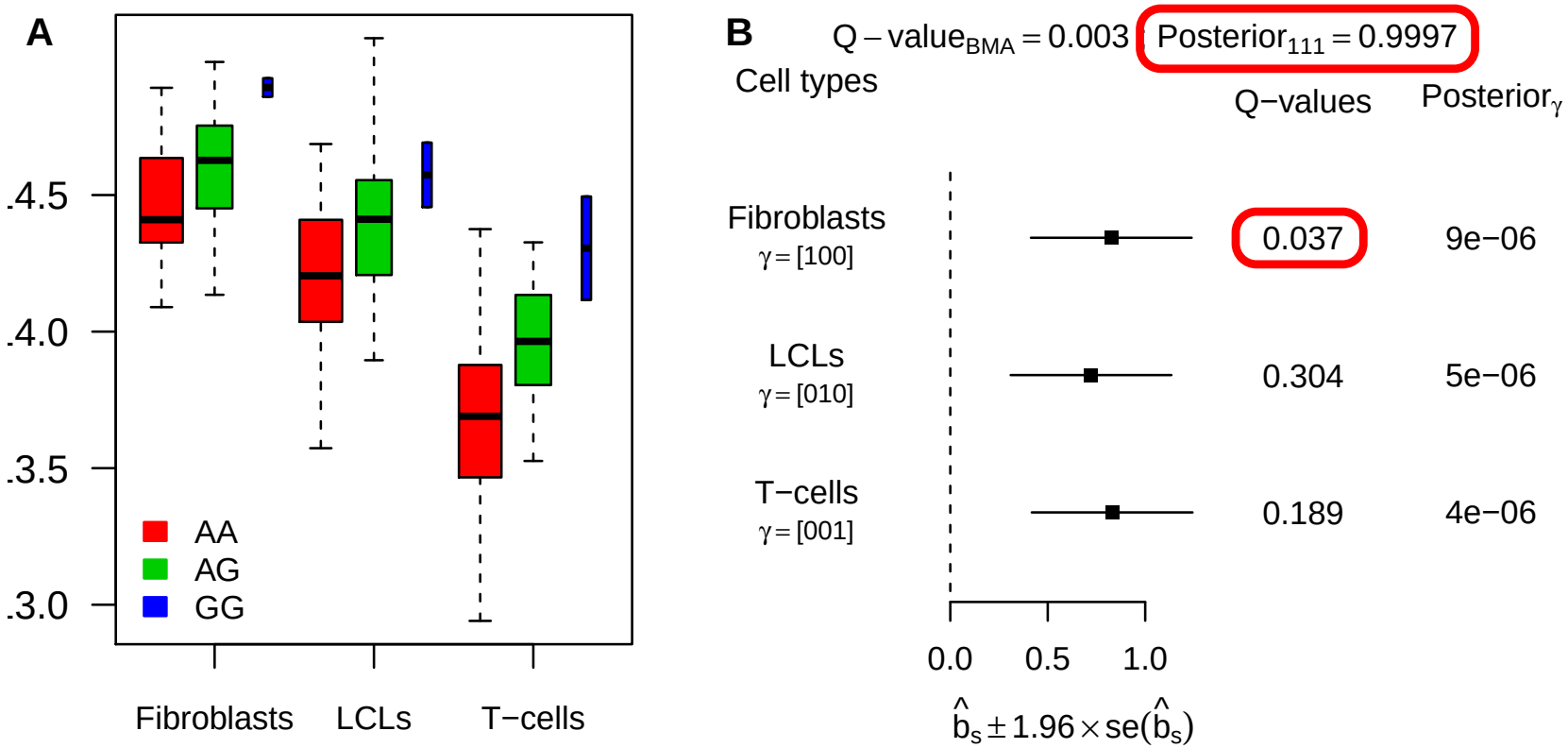


Wrong tissue-specific call by the tissue-by-tissue analysis



Example of gene ENSG00000106153 and SNP rs4948093 (MAF=0.23).
See also Ding *et al.* (2010, AJHG).

Wrong tissue-specific call by the tissue-by-tissue analysis



Example of gene ENSG00000106153 and SNP rs4948093 (MAF=0.23).
See also Ding *et al.* (2010, AJHG).

Conclusions and perspectives

Our framework:

- maps eQTLs jointly across tissues and explicitly models heterogeneity;
- has more power and gives more reliable estimates of tissue specificity than a tissue-by-tissue analysis;

Conclusions and perspectives

Our framework:

- maps eQTLs jointly across tissues and explicitly models heterogeneity;
- has more power and gives more reliable estimates of tissue specificity than a tissue-by-tissue analysis;

- a non-exhaustive version of our framework (`BMAlite`) can handle data sets with "many" tissues (eg. more than 15-20);
- our hierarchical model can also incorporate some genomic annotations.

Acknowledgments

Co-authors:

- William Wen
- Matthew Stephens
- Jonathan Pritchard

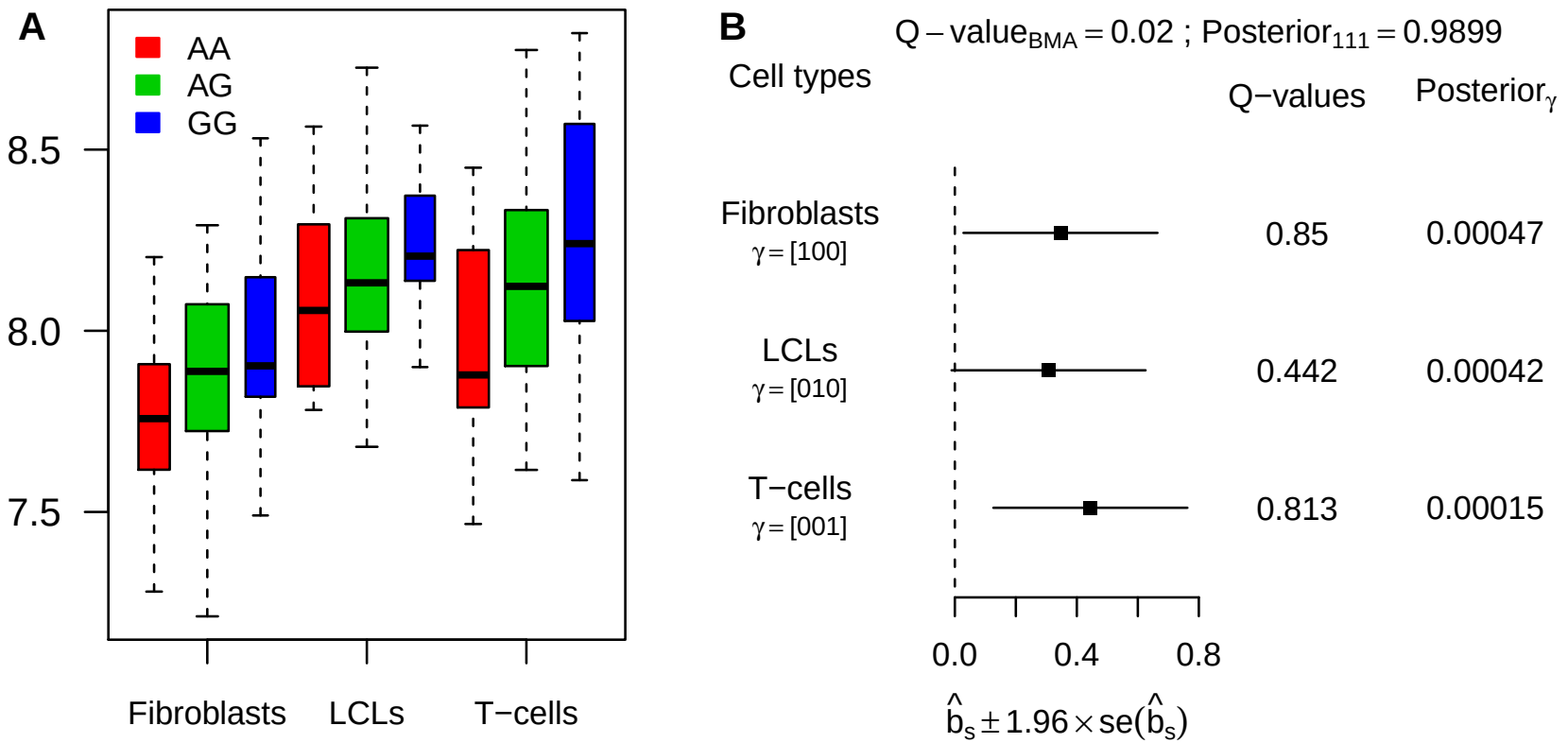
Funding of T. Flutre:

- INRA
- NIH (GTEx project)

References:

- Wen & Stephens (2011, arXiv), Wen (2012, arXiv)
- Han & Eskin (2011, AJHG)
- Ding *et al.* (2010, AJHG)
- Lebrec *et al.* (2010, SAGMB)
- Veyrieras *et al.* (2008, PLoS Genetics)

Weak, yet consistent eQTL called only by BMA



Example of gene ENSG00000090924 and SNP rs755690.