

DREAM5 Systems Genetics Challenge 3A

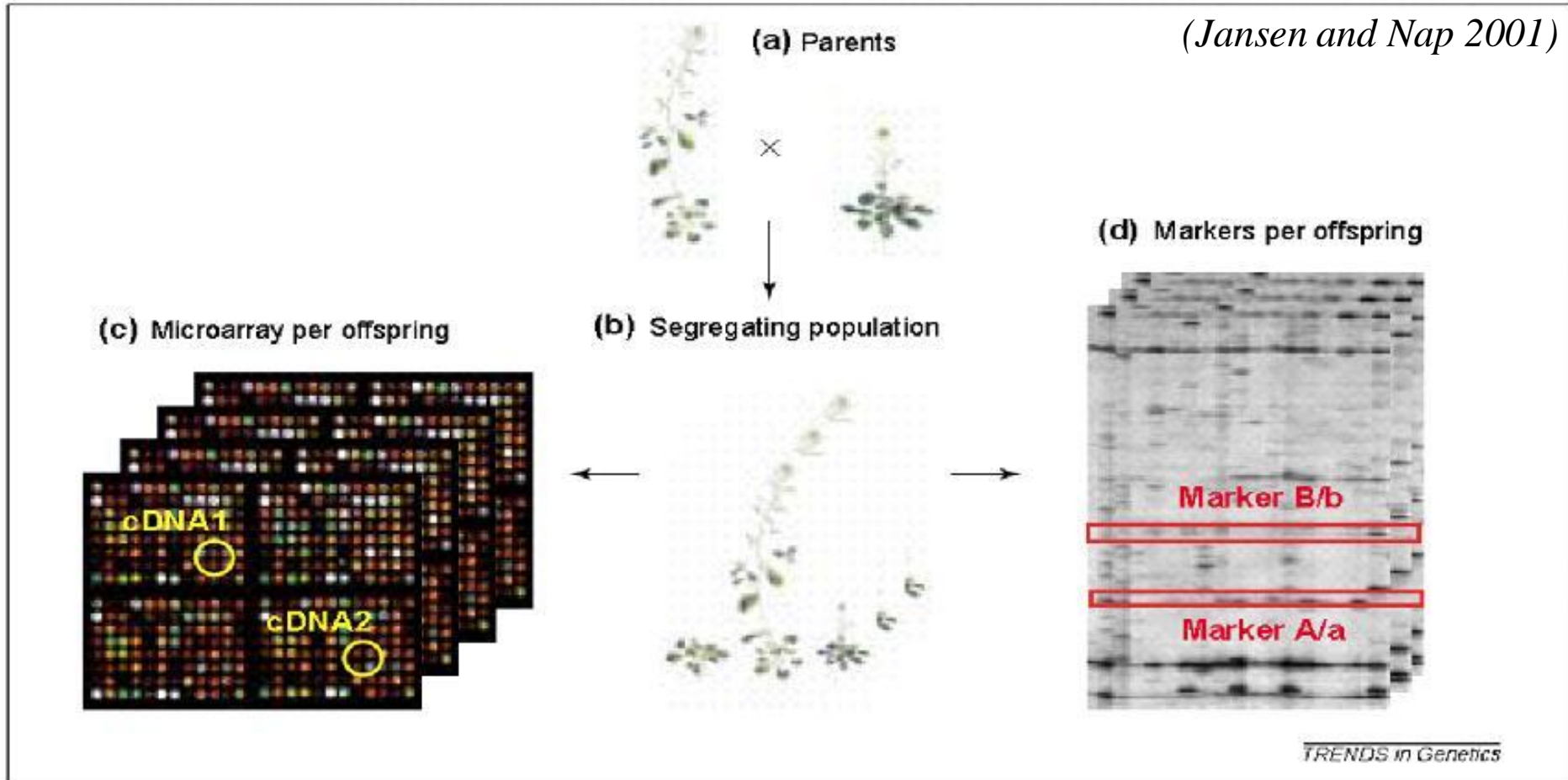
Gene regulatory network reconstruction using Bayesian Networks, the Dantzig selector and the Lasso: a meta-analysis

David Allouche, Christine Cierco-Ayrolles, Simon de Givry,
Brigitte Mangin, Nidal Ramadan-Alban, Thomas Schiex,
Jimmy Vandael, Matthieu Vignes

SaAB Team, INRA – MIA, Toulouse, France



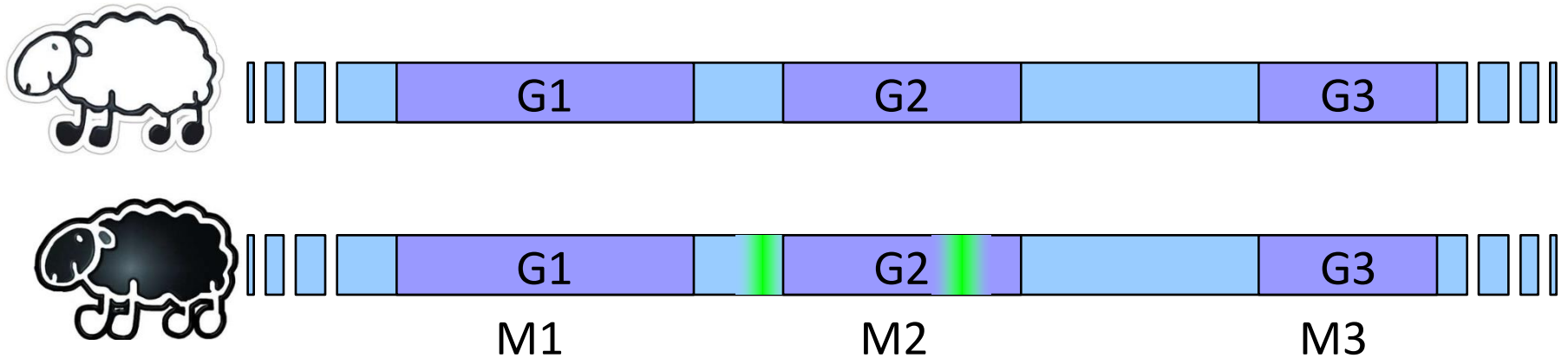
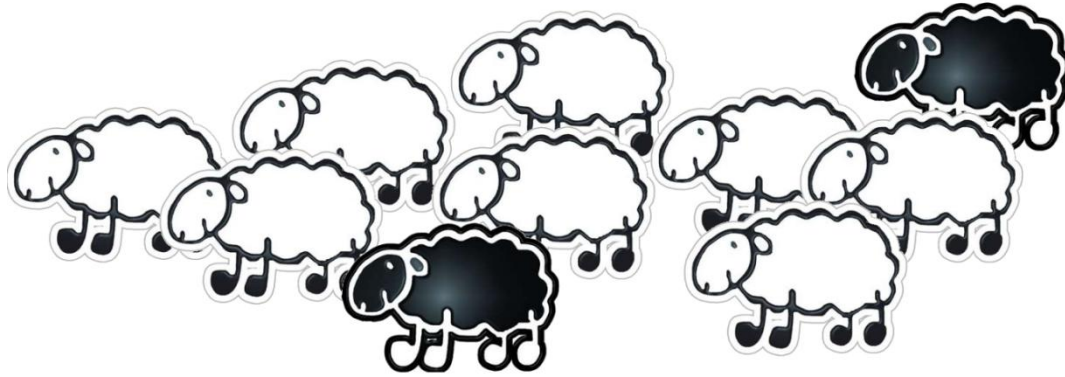
Genetical Genomics



Data: 1000 Expression levels, 1000 Marker genotypes (SNP)

RIL population size: A1: 100 individuals, A2: 300 ind., A3: 999 ind.

Polymorphism



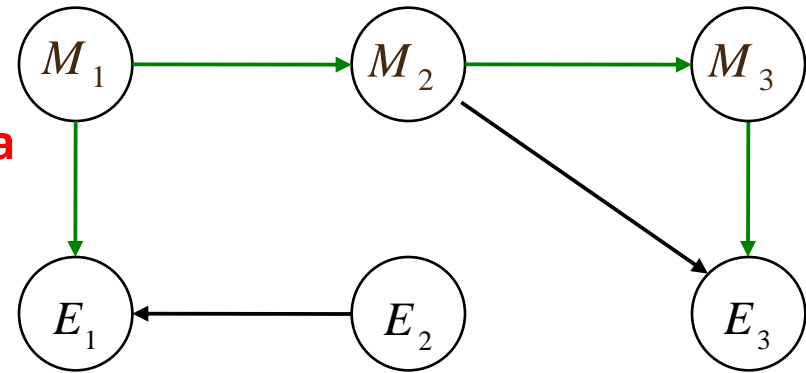
- DNA mutations in genes (1 marker / gene)
 - In promoter region (impact on gene activity)
« *cis-effect* »
 - In coding region (modify protein structure)
« *trans-effect* »

→ **Prior test for linear regression to detect cis (hence non-cis)-acting regulation**

Probabilistic graphical models

Discrete graphical model

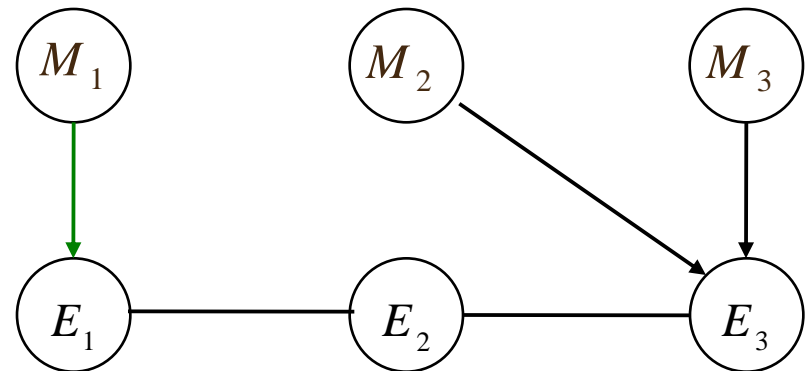
→ **1-Bayesian network on discrete data**
(Friedman 2000), (Vandel et al. 2010),...



Linear model

Graphical Gaussian Models

→ **Local regressions:**
2-Lasso (Tibshirani 1996),
ElasticNet (Zou and Hastie 2004)
3-Dantzig (Candès and Tao 2007)



+ **Meta-analysis**

Score-based Bayesian Network learning

➤ Bayesian Network on discrete random variables

➤ **Directed Acyclic Graph** G

(in)dependencies between variables

➤ **Conditional probability distribution** $P_G(X_i / Pa_i)$

$$P_G(X) = \prod P_G(X_i / Pa_i)$$

➤ Find the graph $G_{best} = \operatorname{argmax}_G P(G / D)$ with dataset D

$$P(G / D) \propto P(D / G)P(G)$$

➤ $P(D / G)$ **marginal likelihood of the graph**
Bayesian Dirichlet score (BDeu)

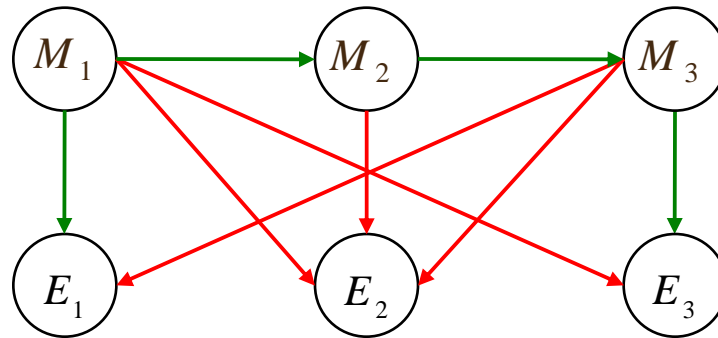
with equivalent sample size $\alpha = 1(A1), 2(A2), 5(A3)$

➤ $P(G)$ **prior probability of the graph**
with **uniform prior on the number of parents**

(Chen 2008)

$\gamma = 0.5$

Restricted DAG search space



$$M_i: \{0,1\}$$

$$E_i: \{1,2,3\}$$

$$-0.5 \leq \log_2 \left(\frac{E_i^l}{\hat{E}_i} \right) \leq 0.5$$

→ Genetic linkage between markers (Carthagene mapping software (*Givry et al. 2006*))

→ *Cis-effect*: mutation in promoter region of gene i (example: M_1 and M_3)

- **Enforce arc** $M_i \rightarrow E_i$
- **Forbid arcs** $M_i \rightarrow E_j \quad \forall j \neq i$

→ *Trans-effect*: mutation in coding region of gene i (example: M_2)

- **Forbid arc** $M_i \rightarrow E_i$

Sparse candidate greedy search

(Friedman et al. 1999)

- Sparse list of candidate parents per E_i
 - Test one parent (gene-expression or marker) versus no parent
$$P_{BIC}(E_i | X_j) > P_{BIC}(E_i) \quad \forall X_j \in \{E_j, M_j\}$$
 - Select at most one best marker inside a sliding window (50 cM) along the chromosomes.
- Maximum number of parents ≤ 7 (observed was 4)
- Start with an *empty* DAG, greedy algorithm: insert/reverse/delete edges
- Edge weight: *influence score* (Yu et al. 2002)

2&3-Regression model

Gene-by-gene linear regressions. For gene i :

$$E_i = \mathbf{E} \cdot \beta_i + \mathbf{M} \cdot \theta_i + \varepsilon_i,$$


- **E**: gene expression levels ($n \times p$ matrix)
- **M**: genotypes ($n \times p$ matrix)
- β_i : effects of expression levels on y_i (p -vector, $\beta_{ii}=0$)
- θ_i : effects of markers on y_i (p -vector)
- ε_i : Gaussian residual error term.

The network structure is *encoded in non-zero entries of matrices β and θ* that need estimation.

2-Lasso regression

(Tibshirani, 1996)

Gene i :
$$\min \left\| \mathbf{E} - \mathbf{E} \cdot \boldsymbol{\beta} - \mathbf{M} \cdot \boldsymbol{\theta} \right\|_{\ell_2} + \lambda \left\| (\boldsymbol{\beta}, \boldsymbol{\theta}) \right\|_{\ell_1}$$

 Estimates $\boldsymbol{\beta}^\lambda, \boldsymbol{\theta}^\lambda$ for given λ (repeated for 20 different values $\lambda_{\max}/20$ to λ_{\max})

- Solved with LAR (Efron et al. 2003) algorithm. No model selection (BIC, cross validation, Meinshausen and Bühlman 2006...) , rather a consensus.
- ~~Post proc: cis effect enforces θ_{ij} to 0 for $j \neq i$ in range $[i-F, i+F]$.~~
- Edges that have no causal basis are symetrized. Causality is inferred from θ .
- Reliability of $i \rightarrow j$ is the ratio of occurence on λ grid. Halved for undirected edges.


3-The Dantzig selector

(Candès and Tao, 2007)

Gene i

$$\min \left\| (\beta_i, \theta_i) \right\|_{\ell_1}$$

s.t. $\left\| [\mathbf{E}_{\setminus i}, \mathbf{M}]^* r_i \right\|_{\ell_\infty} \leq \delta$ where r_i is the residual vector
(bounded residual/variables correlations)

 Estimates $\beta_{ij}^\delta, \theta_{ij}^\delta$ for bound δ

- Reduces to linear programming
- Solved for 20 evenly spaced values of $\delta \in [0, \delta_{\max} [$ where δ_{\max} : minimum δ that leads to an empty network.
- Postprocessing as in LASSO.

1+2+3 = Meta analysis


$$\mathfrak{M} = \{Lasso, Dantzig, BayesNet\}$$

$$r_{ij}^{meta} = 1 - \exp\left(\sum_{m \in \mathfrak{M}} \log(1 - r_{ij}^m)\right)$$

r_{ij}^m : reliability of edge $i \rightarrow j$ for method m

~ Fisher's inverse χ^2 method

(Hedge and Olkin 1985)

 **Calibration of the reliabilities between methods:**
No change for Dantzig and BayesNet
Reliabilities for Lasso set between 0 and $\frac{1}{2}$




Implementation details and CPU times

- **BayesNet:** Greedy Search using Banjo (*Hartemink 2005*)
A1: ~ 20' A2: ~ 70' A3: ~ 180'
- **Lasso:** R scripts based on glmnet package
A1: ~ 10' A2: ~ 20' A3: ~ 60'
- **Dantzig:** glpk linear programming solver
A1: ~ 300' A2: ~ 1300' A3: ~ 6600'
- **Meta:** few R code lines
runs in a few seconds

Acknowledgements:

Lasso and Dantzig ran on GenoToul and GenoQuest bioinformatic platforms.

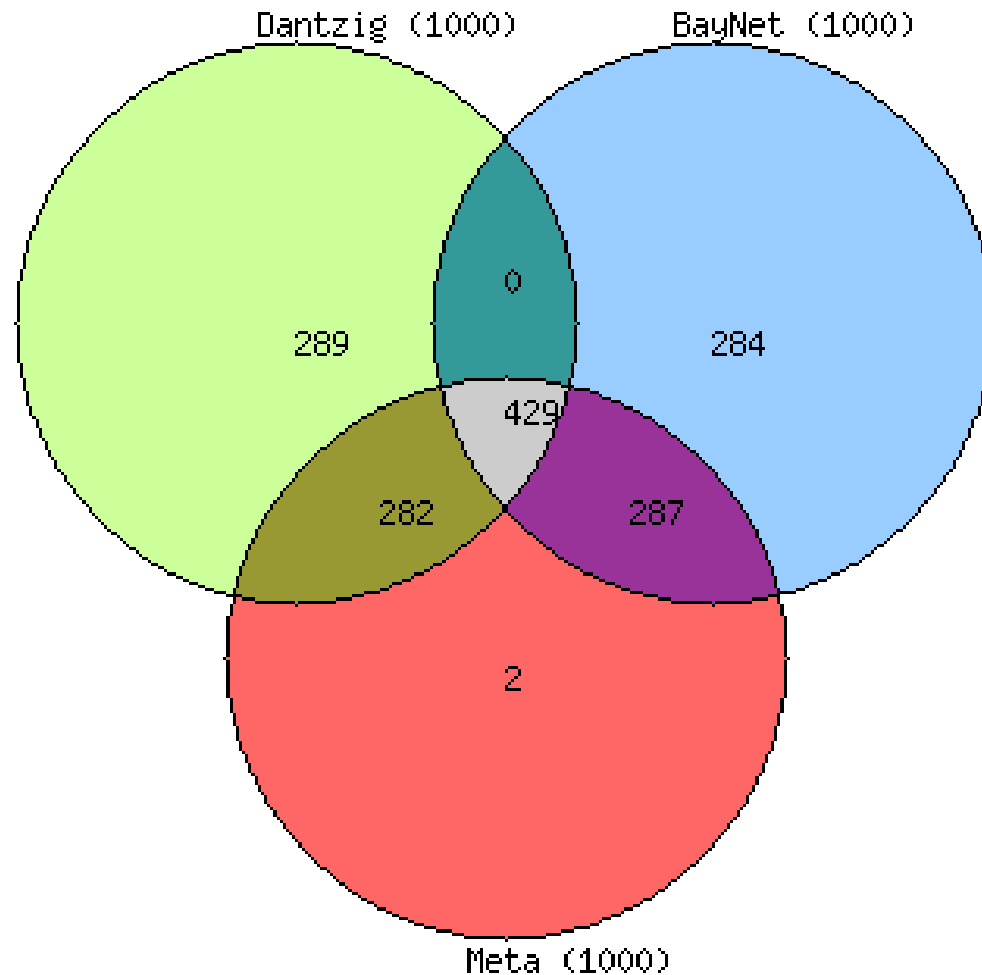
Results

	Sample size 100		Sample size 300		Sample size 999	
	rang	score	rang	score	rang	score
Meta	1	81.87	1	89.40	1	140.56
Dantzig	3	78.64	2	87.92	2	135.91
BayesNet	13	0.00	12	0.00	8	3.52
Lasso						

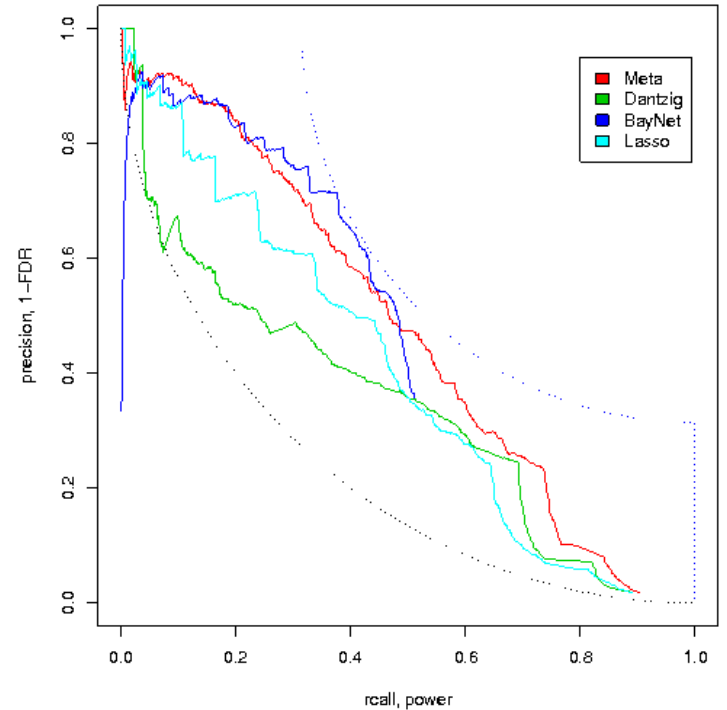
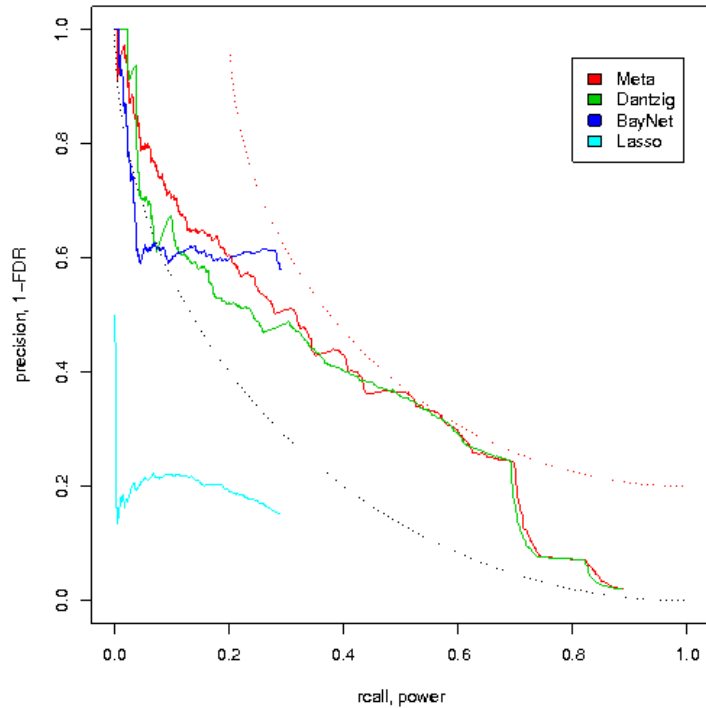


Lasso had errors in edge direction,
reliability calibrated accordingly

Venn diagram for the first 1,000 edges



Precision vs Recall curves (left: old, right: new)



Conclusions & Prospects

- Genetical genomics data: potential for causal inference in gene regulatory networks.
- Accuracy increases with sample size. Seems to decrease a wee bit with average degree.
- Results in terms of absolute Precision/Recall (slightly) disappointing.
- Check results according to data/network features.
- Elastic Net procedure to clean out.
- Application on real genuine datasets (FRAGENOMICS ANR research project)

References

- * **Chickering.** *Efficient Approximations for the Marginal Likelihood of Incomplete Data Given a Bayesian Network.* AI 1997.
- * **Chen.** *Extended bayesian information criteria for model selection with large model spaces.* Biometrika 95(3), 2008.
- * **Candès and Tao.** *The Dantzig selector: statistical estimation when p is much larger than n .* Ann. Stat., 2007.
- * **Darwiche.** *Modeling and Reasoning with Bayesian Networks.* Cambridge University Press, 2009.
- * **Efron, Hastie, Johnstone, and Tibshirani.** *Least Angle Regression.* Ann. Stat., 2004.
- * **Friedman, Linial, Nachman, Pe'er.** *Using bayesian networks to analyse expression data.* J. Comp. Biol., 2000.
- * **Friedman, Nachman, Pe'er.** *Learning Bayesian network structure from massive datasets: The “sparse candidate” algorithm.* In Proc. of UAI, 1999.
- * **Givry, Bouchez, Chabrier, Milan, Schiex.** *Carthagene: multipopulation integrated genetic and radiated hybrid mapping.* Bioinformatics, 2005.
- * **Hartemink.** *Reverse engineering gene regulatory networks.* Nature Biotechnology, 2005.
- * **Jansen, Nap.** *Genetical genomics: the added value from segregation.* Trends in Genetics, 2001.
- * **Meinshausen and Bühlmann.** *High dimensional graphs and variable selection with the lasso.* Ann. Stat., 2006.
- * **Liu, Fuente, Hoeschele.** *Gene network inference via structural equation modeling in genetical genomics experiments.* Genetics 178, 2008.
- * **Tibshirani.** *Regression shrinkage and selection via the lasso.* J. Royal. Statist. Soc B., 1996.
- * **Vandel, Mangin, Vignes, Givry.** *Extended Bayesian scores for reconstructing gene regulatory networks.* ECCS workshop on graphical models for reasoning on biological systems, 2010.
- * **Yu, Smith, Wang, Hartemink, Jarvis.** *Using bayesian network inference algorithms to recover molecular genetic regulatory networks.* In Int. Conf. on Sys. Biol., 2002.
- * **Zhu, Wiener, Zhang, Fridman, Minch, Lum, Sachs, Schadt.** *Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations.* PLOS Comp. Bio. 3(4), 2007.
- * **Zou and Hastie.** *Regularization and variable selection via the elastic net.* J. Royal. Statist. Soc B., 2005.