STATSEQ: Gene Network Inference Meeting 2013
Paris, 28-29 March

Robert J. Flassig[1], Sandra Heise[1],

Kai Sundmacher[1,2], Steffen Klamt[1]

[1]Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

[2]Chair for Process Systems Engineering, Otto-von-Guericke University Magdeburg, Germany

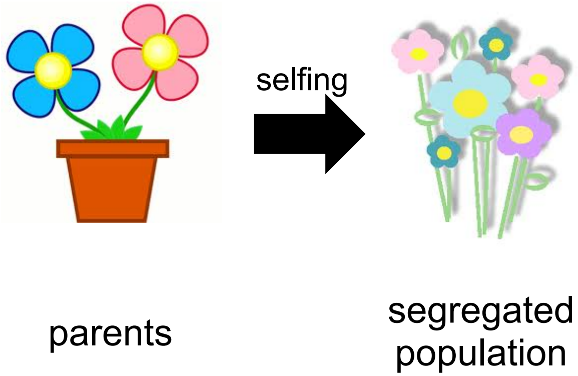
An Effective Framework for Reconstructing Gene Regulatory Networks From Genetical Genomics Data

OTTO-VON-GUERICKE UNIVERSITÄT MAGDEBURG

MAX-PLANCK-INSTITUT
DYNAMIK KOMPLEXER
TECHNISCHER SYSTEME
MAGDEBURG

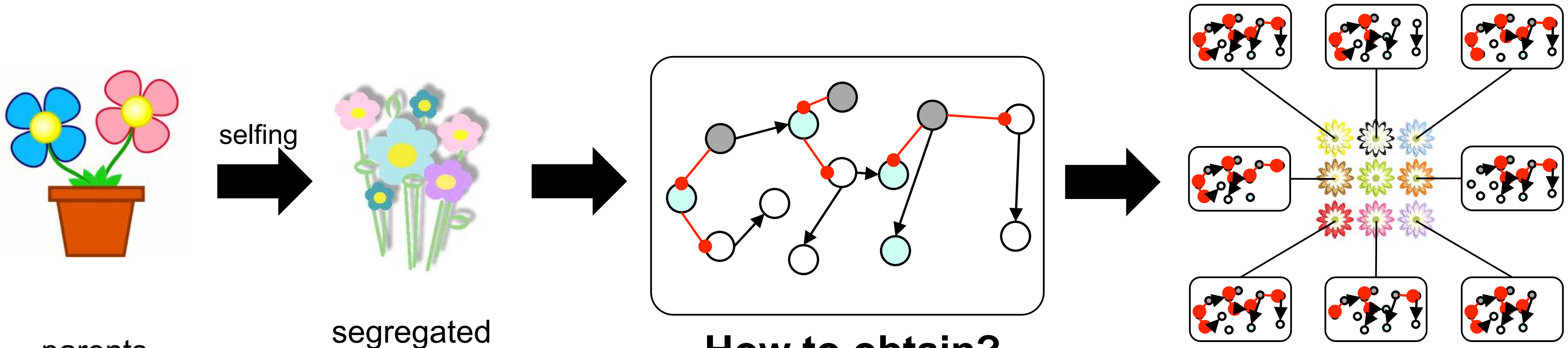
Motivation

- **Wanted**
  - **understanding of gene interactions** via gene regulatory networks
  - optimization of phenotype via **optimization of gene interactions**



Motivation

- **Wanted**
  - **understanding of gene interactions** via gene regulatory networks
  - optimization of phenotype via **optimization of gene interactions**
- **Need**
  - infer interaction structure (GRN) from data
    - **systems genetics approach: use segregated population**
    - ...



parents

segregated population

How to obtain?

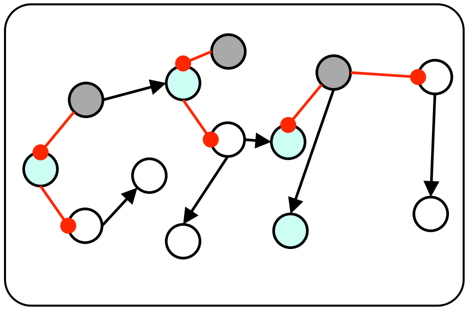
Motivation

- Many (complex) methods
- What about simple correlation?

parents

selfing

segregated population



How to obtain?

Challenge

$$n_{\text{sample}}/n_{\text{genes}} \ll 1$$

Outline

- Motivation
- Method
- Application
- Conclusion

Overview Method

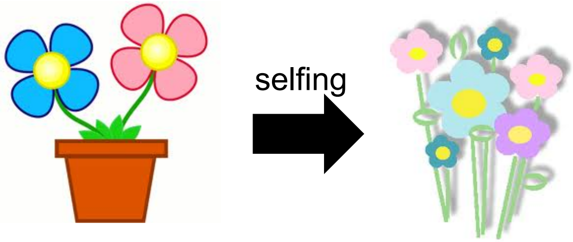
1. Data preprocessing

2. Reconstruct interactions between genes
   - raw perturbation graph G1 (no edge weights nor signs)
   - raw perturbation graph G2 (edge weights & signs)
   - identify eQTLs
   - select ONE candidate gene

3. Prune false positive interactions via transitive reduction
   - remove redundant candidate genes/interactions

1. Data Preprocessing

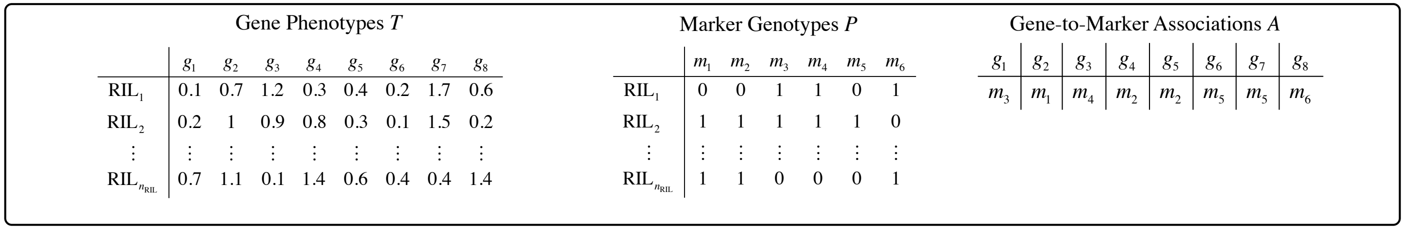
- data set from study population:



➢ **phenotype** of genes *T*

➢ **genotype** of marker *P*

➢ **gene-to-marker** association *A*



Gene Phenotypes *T*

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
RIL_1	0.1	0.7	1.2	0.3	0.4	0.2	1.7	0.6
RIL_2	0.2	1	0.9	0.8	0.3	0.1	1.5	0.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\text{RIL}_{n_\text{RIL}}$	0.7	1.1	0.1	1.4	0.6	0.4	0.4	1.4

Marker Genotypes *P*

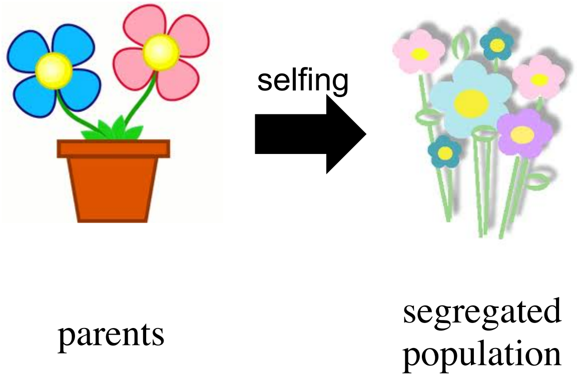
	m_1	m_2	m_3	m_4	m_5	m_6
RIL_1	0	0	1	1	0	1
RIL_2	1	1	1	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\text{RIL}_{n_\text{RIL}}$	1	1	0	0	0	1

Gene-to-Marker Associations *A*

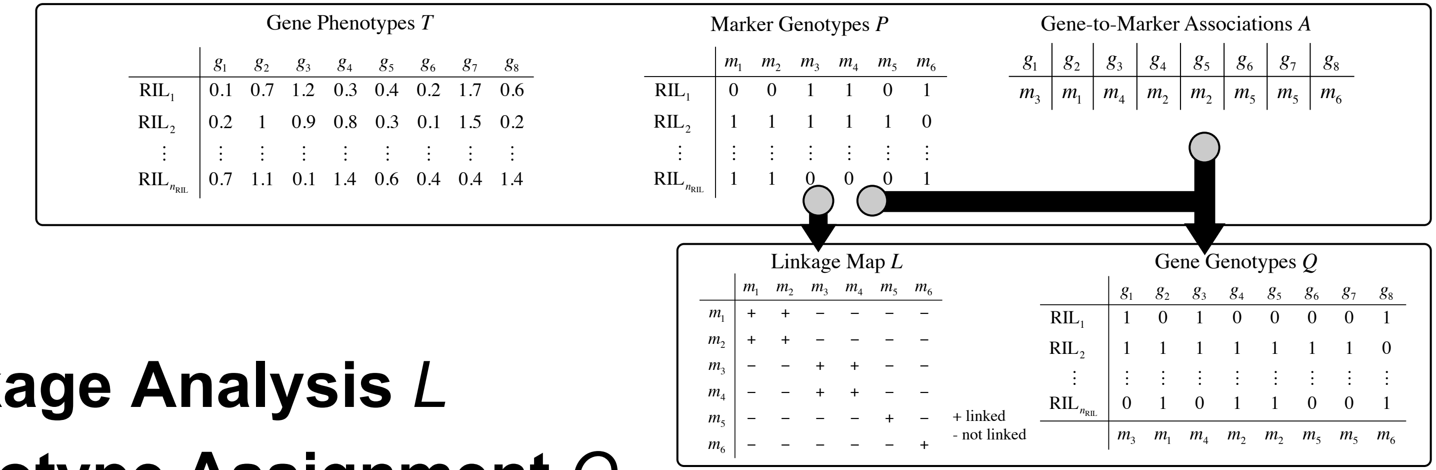
g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
m_3	m_1	m_4	m_2	m_2	m_5	m_5	m_6

1. Data Preprocessing

- data set from study population:



- ➢ **phenotype** of genes *T*
- ➢ **genotype** of marker *P*
- ➢ **gene-to-marker** association *A*



Gene Phenotypes *T*

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
RIL_1	0.1	0.7	1.2	0.3	0.4	0.2	1.7	0.6
RIL_2	0.2	1	0.9	0.8	0.3	0.1	1.5	0.2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\text{RIL}_{n_\text{RIL}}$	0.7	1.1	0.1	1.4	0.6	0.4	0.4	1.4

Marker Genotypes *P*

	m_1	m_2	m_3	m_4	m_5	m_6
RIL_1	0	0	1	1	0	1
RIL_2	1	1	1	1	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\text{RIL}_{n_\text{RIL}}$	1	1	0	0	0	1

Gene-to-Marker Associations *A*

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
m_3	m_1	m_4	m_2	m_2	m_5	m_5	m_6

Linkage Map *L*

	m_1	m_2	m_3	m_4	m_5	m_6
m_1	+	+	−	−	−	−
m_2	+	+	−	−	−	−
m_3	−	−	+	+	−	−
m_4	−	−	+	+	−	−
m_5	−	−	−	−	+	−
m_6	−	−	−	−	−	+

+ linked
- not linked

Gene Genotypes *Q*

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
RIL_1	1	0	1	0	0	0	0	1
RIL_2	1	1	1	1	1	1	1	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\text{RIL}_{n_\text{RIL}}$	0	1	0	1	1	0	0	1
	m_3	m_1	m_4	m_2	m_2	m_5	m_5	m_6

- **Linkage Analysis** *L*
- **Genotype Assignment** *Q*

1. Data Preprocessing

Linkage Analysis

- identify genetic linkage of markers via **genotype-genotype correlation** $r^{P_iP_j}$ and threshold d_{min}
- if $r^{P_iP_j} \geq d_{min}$ then $m_j \in \mu_i$, with μ_i set of markers linked to marker m_i

Marker Genotypes *P*

	m_1	m_2	m_3	m_4	m_5	m_6
RIL_1	0	0	1	1	0	1
RIL_2	1	1	1	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\text{RIL}_{n_\text{RIL}}$	1	1	0	0	0	1

Linkage Map *L*

	m_1	m_2	m_3	m_4	m_5	m_6
m_1	+	+	−	−	−	−
m_2	+	+	−	−	−	−
m_3	−	−	+	+	−	−
m_4	−	−	+	+	−	−
m_5	−	−	−	−	+	−
m_6	−	−	−	−	−	+

1. Data Preprocessing

Genotype Assignment

- n_{m} genotyped markers, n_{g} phenotyped genes for n_{RIL} RILs
- gene-to-marker association

1. Data Preprocessing

Genotype Assignment

- n_{m} genotyped markers, n_{g} phenotyped genes for n_{RIL} RILs
- gene-to-marker association
- assign genotype to genes from associated marker genotypes

Marker Genotypes *P*

	m_1	m_2	m_3	m_4	m_5	m_6
RIL_1	0	0	1	1	0	1
RIL_2	1	1	1	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\mathrm{RIL}_{n_{\mathrm{RIL}}}$	1	1	0	0	0	1

Gene-to-Marker Associations *A*

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
m_3	m_1	m_4	m_2	m_2	m_5	m_5	m_6

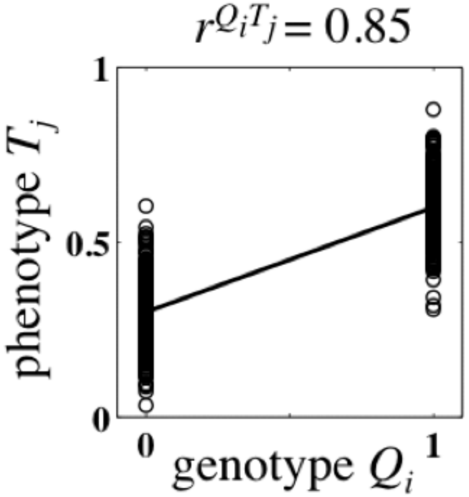
Gene Genotypes *Q*

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
RIL_1	1	0	1	0	0	0	0	1
RIL_2	1	1	1	1	1	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\mathrm{RIL}_{n_{\mathrm{RIL}}}$	0	1	0	1	1	0	0	1
	m_3	m_1	m_4	m_2	m_2	m_5	m_5	m_6

2. Reconstruction of Perturbation Graphs

Raw Perturbation Graph G1

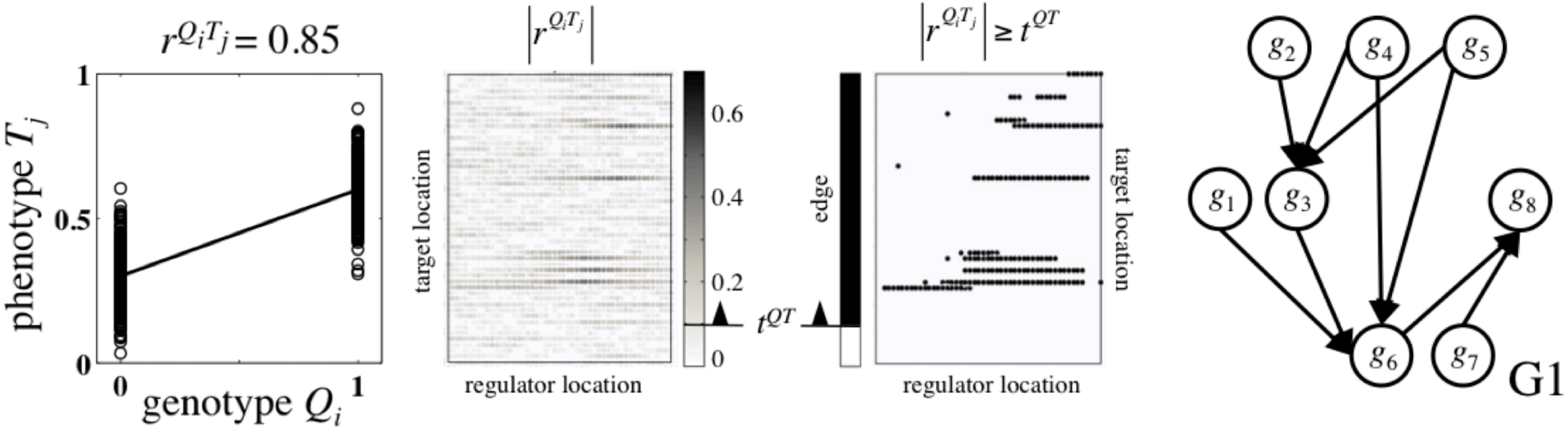
- directed **edge detection** based on **genotype-phenotype correlation** between genes and threshold t^{QT}: $\left| r^{Q_i T_j} \right| \geq t^{QT}$



2. Reconstruction of Perturbation Graphs

Raw Perturbation Graph G1

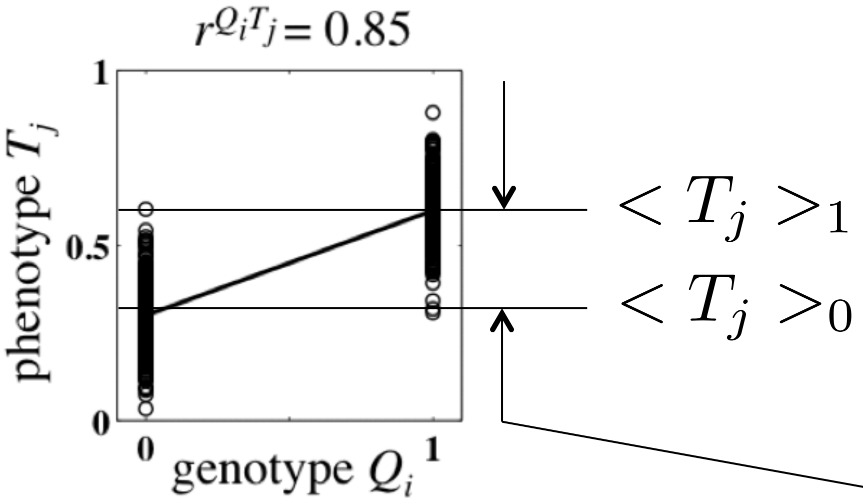
- directed **edge detection** based on **genotype-phenotype correlation** between genes and threshold t^{QT}: $\left| r^{Q_i T_j} \right| \geq t^{QT}$



2. Reconstruction of Perturbation Graphs

Raw Perturbation Graph G1

- directed **edge detection** based on **genotype-phenotype correlation** between genes and threshold t^{QT}: $\left|r^{Q_iT_j}\right| \geq t^{QT}$



- $r^{Q_iT_j}$ is a z-score of deviations $$r^{Q_iT_j} = \frac{< T_j >_1 - < T_j >_0}{s_{T_j}/2}$$

2. Reconstruction of Perturbation Graphs

Raw Perturbation Graph G2

- **sign detection** from **pheno-phenotype correlation** $r^{T_iT_j}$

Gene Phenotypes *T*

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
RIL_1	0.1	0.7	1.2	0.3	0.4	0.2	1.7	0.6
RIL_2	0.2	1	0.9	0.8	0.3	0.1	1.5	0.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\mathrm{RIL}_{n_{\mathrm{RIL}}}$	0.7	1.1	0.1	1.4	0.6	0.4	0.4	1.4

2. Reconstruction of Perturbation Graphs

Raw Perturbation Graph G2

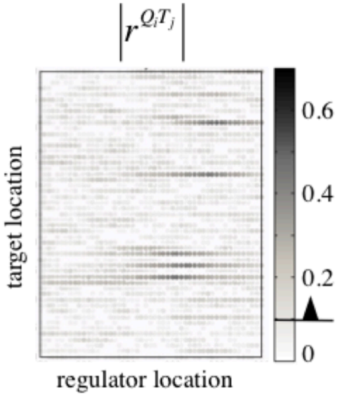
- **sign detection** from **pheno-phenotype correlation** $r^{T_iT_j}$

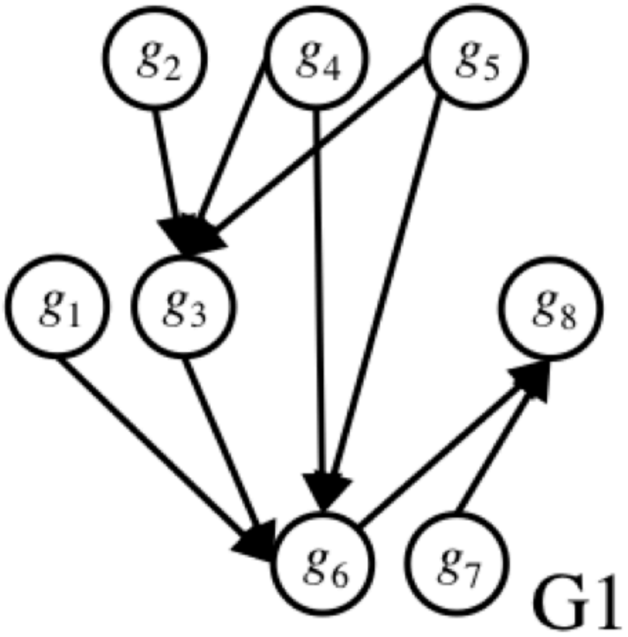
Gene Phenotypes *T*

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
RIL_1	0.1	0.7	1.2	0.3	0.4	0.2	1.7	0.6
RIL_2	0.2	1	0.9	0.8	0.3	0.1	1.5	0.2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathrm{RIL}_{n_{\mathrm{RIL}}}$	0.7	1.1	0.1	1.4	0.6	0.4	0.4	1.4

- **assign edge weights**

$$w_{ij} = \left(\left|r^{Q_iT_j}\right| + \left|r^{T_iT_j}\right|\right)\Big/2$$





2. Reconstruction of Perturbation Graphs

Raw Perturbation Graph G2

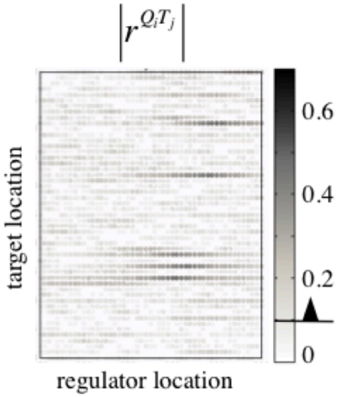
- **sign detection** from **pheno-phenotype correlation** $r^{T_iT_j}$

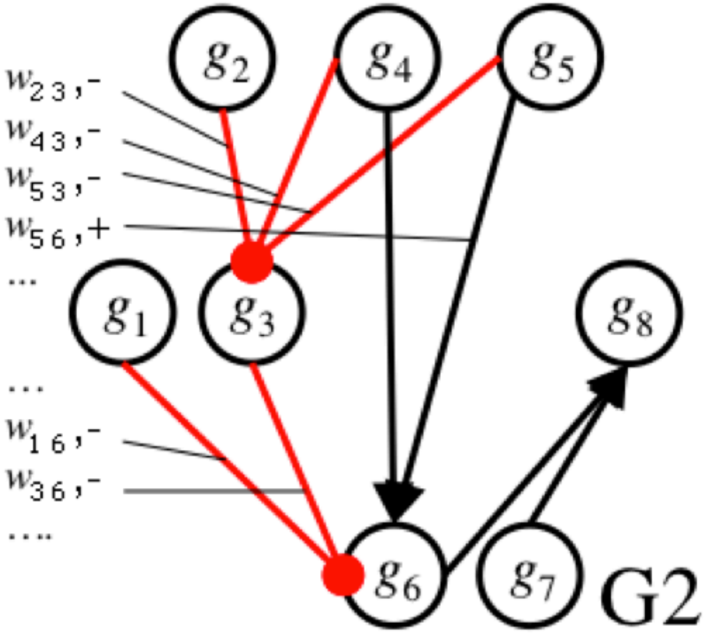
Gene Phenotypes *T*

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
RIL_1	0.1	0.7	1.2	0.3	0.4	0.2	1.7	0.6
RIL_2	0.2	1	0.9	0.8	0.3	0.1	1.5	0.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\mathrm{RIL}_{n_{\mathrm{RIL}}}$	0.7	1.1	0.1	1.4	0.6	0.4	0.4	1.4

- **assign edge weights**

$$w_{ij} = \left(\left|r^{Q_iT_j}\right| + \left|r^{T_iT_j}\right|\right)\Big/2$$





2. Reconstruction of Perturbation Graphs

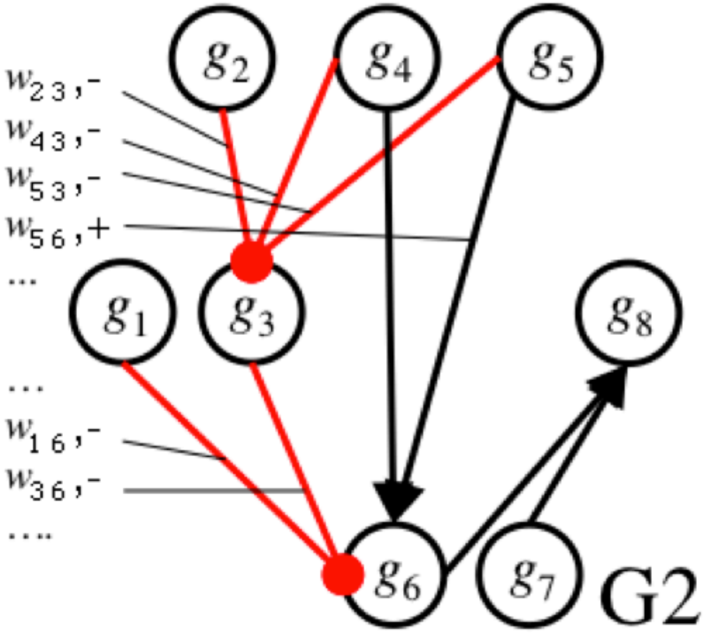
Raw Perturbation Graph G2

= weighted signed digraph

eQTL Graph (G3)
→ digraph with eQTLs

- **candidate regulator selection,** identify one regulator-target edge from each eQTL

Final Perturbation Graph (G4)
→ weighted signed digraph

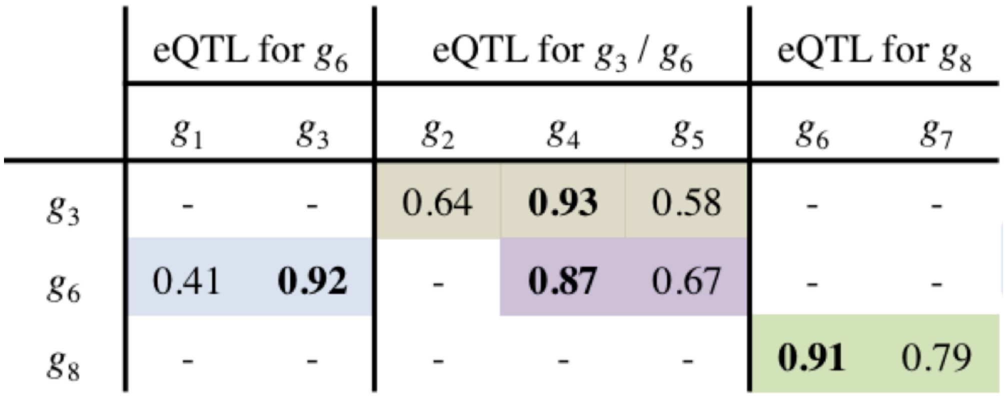


2. Reconstruction of Perturbation Graphs

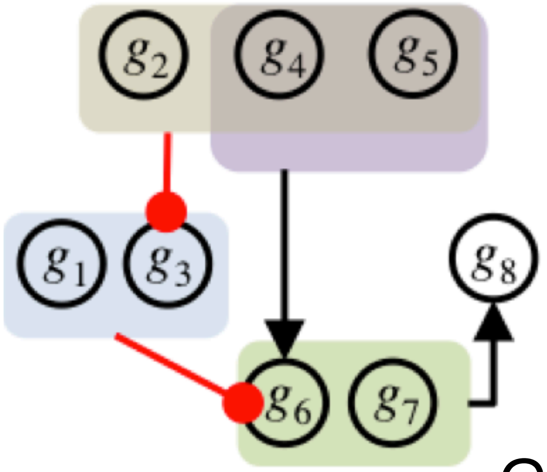
eQTL Perturbation Graph G3

- eQTLs are derived from candidate regulators and marker linkage map
- e.g., $\{g_2, g_4, g_5\}$ form an eQTL for target g_3, due to linkage of their associated markers m_1 and m_2.

Edge weights of regulator gene → target gene



	eQTL for g_6		eQTL for g_3 / g_6			eQTL for g_8	
	g_1	g_3	g_2	g_4	g_5	g_6	g_7
g_3	-	-	0.64	**0.93**	0.58	-	-
g_6	0.41	**0.92**	-	**0.87**	0.67	-	-
g_8	-	-	-	-	-	**0.91**	0.79



G3

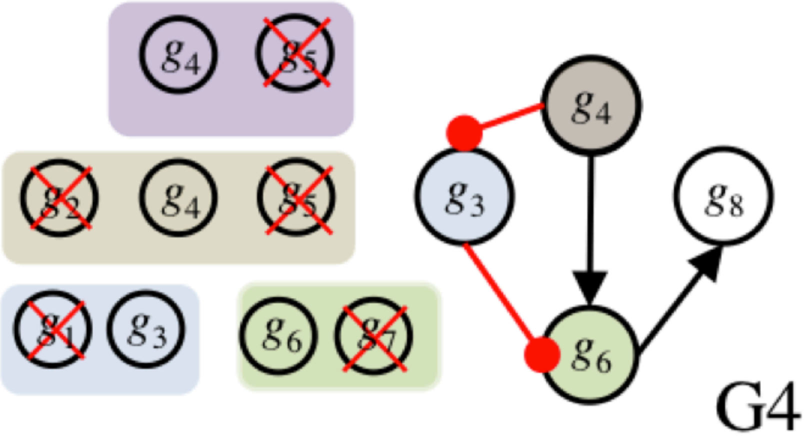
2. Reconstruction of Perturbation Graphs

eQTL Perturbation Graph G3

- eQTLs are derived from candidate regulators and marker linkage map
- e.g., $\{g_2, g_4, g_5\}$ form an eQTL for target g_3, due to linkage of their associated markers m_1 and m_2.

Edge weights of regulator gene → target gene

	eQTL for g_6		eQTL for g_3 / g_6			eQTL for g_8	
	g_1	g_3	g_2	g_4	g_5	g_6	g_7
g_3	-	-	0.64	**0.93**	0.58	-	-
g_6	0.41	**0.92**	-	**0.87**	0.67	-	-
g_8	-	-	-	-	-	**0.91**	0.79

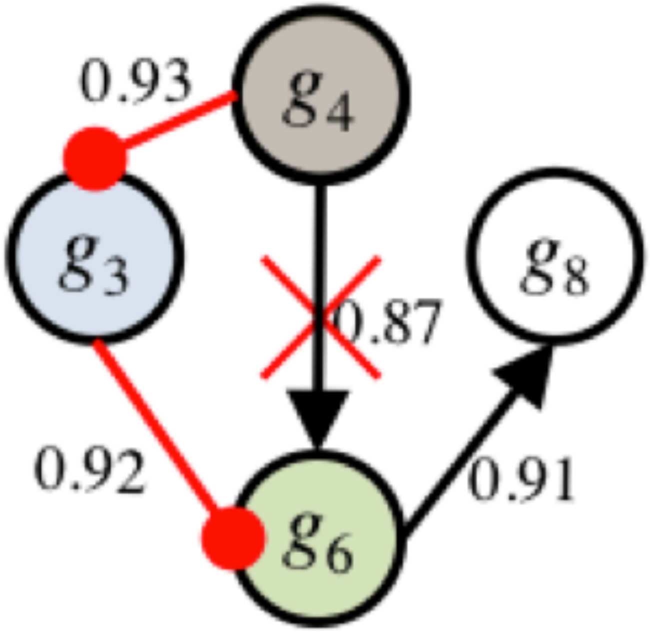


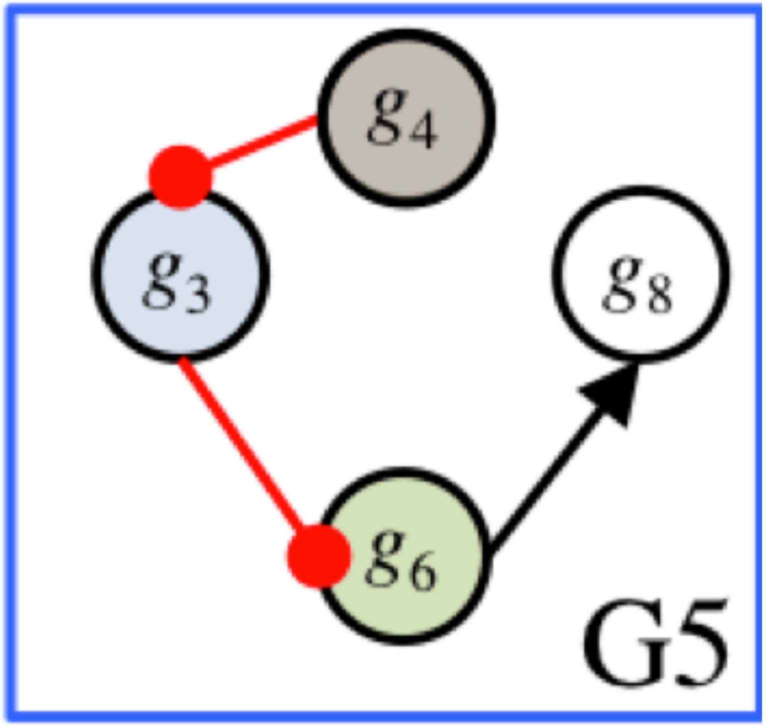
Final Perturbation Graph G4

- from each eQTL, pick the regulator gene with highest edge weight

3. Transitive Reduction

- Input: final perturbation graph G4
- **remove indirect path effects** via transitive reduction using TRANSWESD → final graph G5

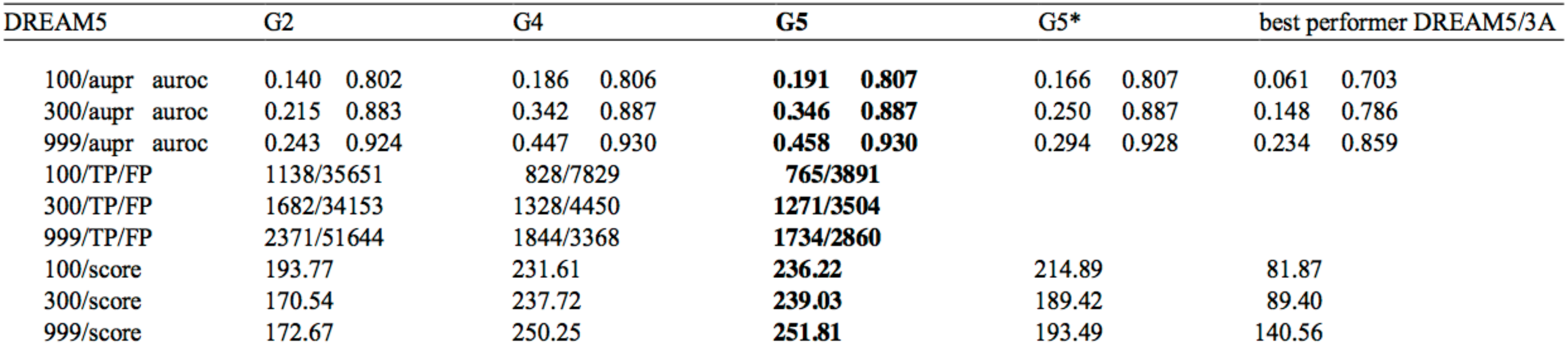




- optionally: **sorted edge list**

Application DREAM 5/3A

- 100, 300, 999 RILs
- 1000 genes on 20 chromosomes



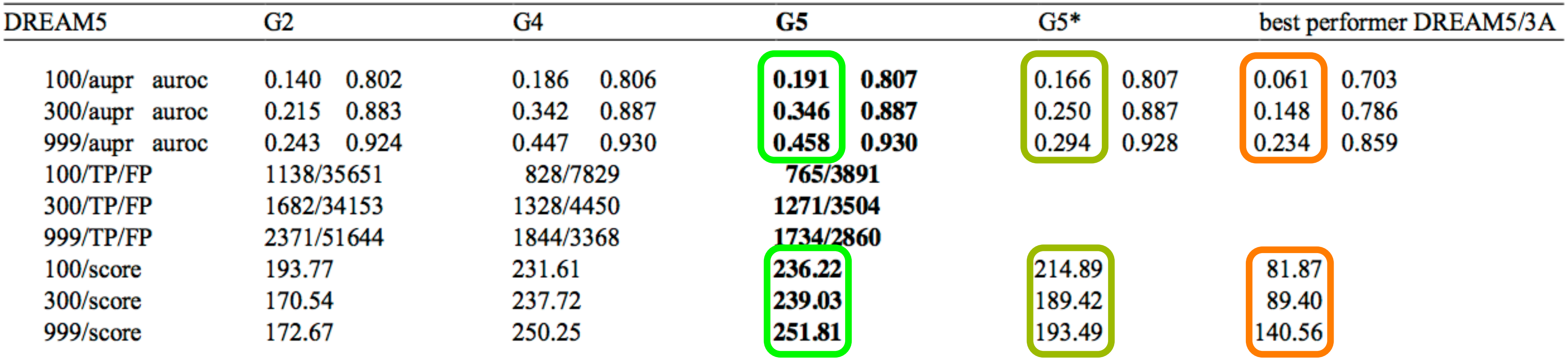
DREAM5	G2	G4	**G5**	G5*	best performer DREAM5/3A
100/aupr auroc	0.140 0.802	0.186 0.806	**0.191 0.807**	0.166 0.807	0.061 0.703
300/aupr auroc	0.215 0.883	0.342 0.887	**0.346 0.887**	0.250 0.887	0.148 0.786
999/aupr auroc	0.243 0.924	0.447 0.930	**0.458 0.930**	0.294 0.928	0.234 0.859
100/TP/FP	1138/35651	828/7829	**765/3891**		
300/TP/FP	1682/34153	1328/4450	**1271/3504**		
999/TP/FP	2371/51644	1844/3368	**1734/2860**		
100/score	193.77	231.61	**236.22**	214.89	81.87
300/score	170.54	237.72	**239.03**	189.42	89.40
999/score	172.67	250.25	**251.81**	193.49	140.56

- G2 unpruned PG
- G4 final PG (pruned eQTLs)
- G5 final graph after TRANSWESD
- G5* averaged over t^{QT} on [0.05...0.6]

[Flassig *et al.*, *Bioinformatics*, 2013]

Application DREAM 5/3A

- 100, 300, 999 RILs
- 1000 genes on 20 chromosomes



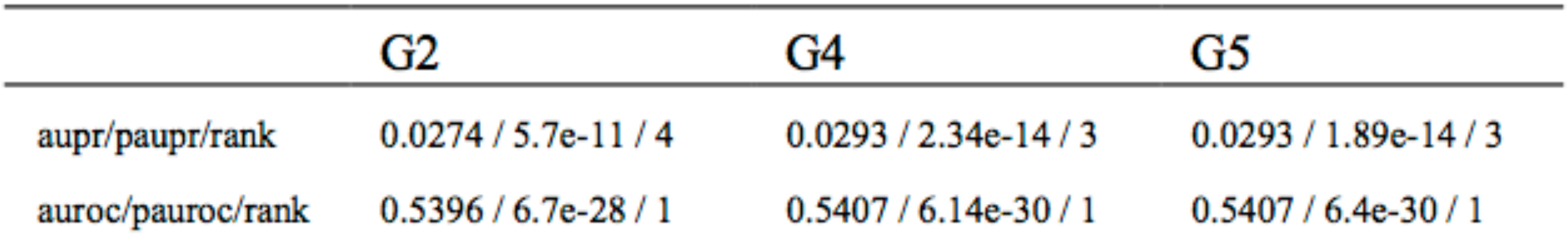
DREAM5		G2		G4		G5		G5*		best performer DREAM5/3A	
100/aupr	auroc	0.140	0.802	0.186	0.806	**0.191**	**0.807**	0.166	0.807	0.061	0.703
300/aupr	auroc	0.215	0.883	0.342	0.887	**0.346**	**0.887**	0.250	0.887	0.148	0.786
999/aupr	auroc	0.243	0.924	0.447	0.930	**0.458**	**0.930**	0.294	0.928	0.234	0.859
100/TP/FP		1138/35651		828/7829		**765/3891**					
300/TP/FP		1682/34153		1328/4450		**1271/3504**					
999/TP/FP		2371/51644		1844/3368		**1734/2860**					
100/score		193.77		231.61		**236.22**		214.89		81.87	
300/score		170.54		237.72		**239.03**		189.42		89.40	
999/score		172.67		250.25		**251.81**		193.49		140.56	

- much more effective in terms of AUPR and AUROC at all sample sizes
- especially at small sample sizes good performance wrt. best performer DREAM5/3A

[Flassig *et al.*, *Bioinformatics*, 2013]

Application *S. cerevisiae*

- data from 112 segregants obtained from a yeast cross (Brem and Kruglyak, PNAS, 2005)
- only 1573 of all 2956 markers were associated to at least one of the 5736 expression-profiled genes
- much less data than in DREAM5/3A
- compare to DREAM5/4.4 submissions



	G2	G4	G5
aupr/paupr/rank	0.0274 / 5.7e-11 / 4	0.0293 / 2.34e-14 / 3	0.0293 / 1.89e-14 / 3
auroc/pauroc/rank	0.5396 / 6.7e-28 / 1	0.5407 / 6.14e-30 / 1	0.5407 / 6.4e-30 / 1

- good performance for 112 samples vs. 536 microarrays with well defined perturbations

[Flassig *et al.*, *Bioinformatics*, 2013]

Summary

- framework for reconstructing GRN from systems genetics data
- **simple correlation** analysis for PG (just sums, no optimization, no matrix operation, no regularization,...) → **large scale GNR**

Summary

- framework for reconstructing GRN from systems genetics data
- **simple correlation** analysis for PG (just sums, no optimization, no matrix operation, no regularization,...) → **large scale GNR**
- performs especially well on **small sample sizes**

Summary

- framework for reconstructing GRN from systems genetics data
- **simple correlation** analysis for PG (just sums, no optimization, no matrix operation, no regularization,...) → **large scale GNR**
- performs especially well on **small sample sizes**
- modular, generate final PG in a different way, then TRANSWESD

Summary

- framework for reconstructing GRN from systems genetics data
- **simple correlation** analysis for PG (just sums, no optimization, no matrix operation, no regularization,...) → **large scale GNR**
- performs especially well on **small sample sizes**
- modular, generate final PG in a different way, then TRANSWESD
- robust wrt. tuning parametrs d_{min} and t^{QT}

Summary

- framework for reconstructing GRN from systems genetics data
- **simple correlation** analysis for PG (just sums, no optimization, no matrix operation, no regularization,...) → **large scale GNR**
- performs especially well on **small sample sizes**
- modular, generate final PG in a different way, then TRANSWESD
- robust wrt. tuning parametrs d_{min} and t^{QT}

References

- Flassig RJ, Heise S, Sundmacher K, Klamt S (2013) An effective framework for reconstructing gene regulatory networks from genetical genomics data. *Bioinformatics* 29 (2): 246-254
- Klamt S, Flassig R, Sundmacher K (2010) TRANSWESD: inferring cellular networks with transitive reduction. *Bioinformatics* 26, 2160-2168
- Brem RB and Kruglyak L. (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. PNAS **102**(5), 1572-1577