Validation of gene networks constructed based on the 50K SNP chip

Joanna Szyda Tomasz Suchocki

Wroclaw University of Natural and Environmental Sciences, Institute of Genetics, Biostatistics Lab
Wroclaw , Poland

Motivation

Identify (all) genes underlying a complex trait

„Biology emerges from pathways,
not from single genes”

Eric Lander

ICQG2012

Motivation

Identify (all) genes underlying a complex trait

STATISTICAL MODELS (GWAS, genomic selection)

- genes with large effects → easy
- genes with medium effect → possible in large samples
- genes with small effect → impossible

BIOINFORMATIC TOOLS

- gene networks
- gene set enrichment
- …

Motivation

Identify (all) genes underlying a complex trait

STATISTICAL MODELS (GWAS, **genomic selection**)

- genes with large effects
- genes with medium effect ples
- genes with small effect

BIOINFORMATIC TOOLS

- **gene networks**
- gene set enrichment
- …

Motivation

Identify (all) genes underlying a complex trait

STATISTICAL MODELS (GWAS, **genomic selection**)

- genes with large effects
- genes with medium effect
- genes with small effect

BIOINFORMATIC TOOLS

- **gene networks**
- gene set enrichment
- …

PROBLEMS

- **SNP information lost**
- **LD information lost**
- **No network validation**

1. Data
2. **Gene selection**
3. Functional features
4. **Network validation**
5. Results
6. Conclusions

Data

4 375 HF animals

SNP information

Gene information

61% EBV | **39% no EBV**

- **animals bulls, cows, heifers** → MASinBULL project
- **SNP genotypes** → 50K chip
- **SNP position** → Illumina + manually corr.
- **SNP pairwise LD** → PLINK
- **Gene position** → Ensemble rel.68 07.2012
- **EBV for milk yield** → evaluation 04.2012

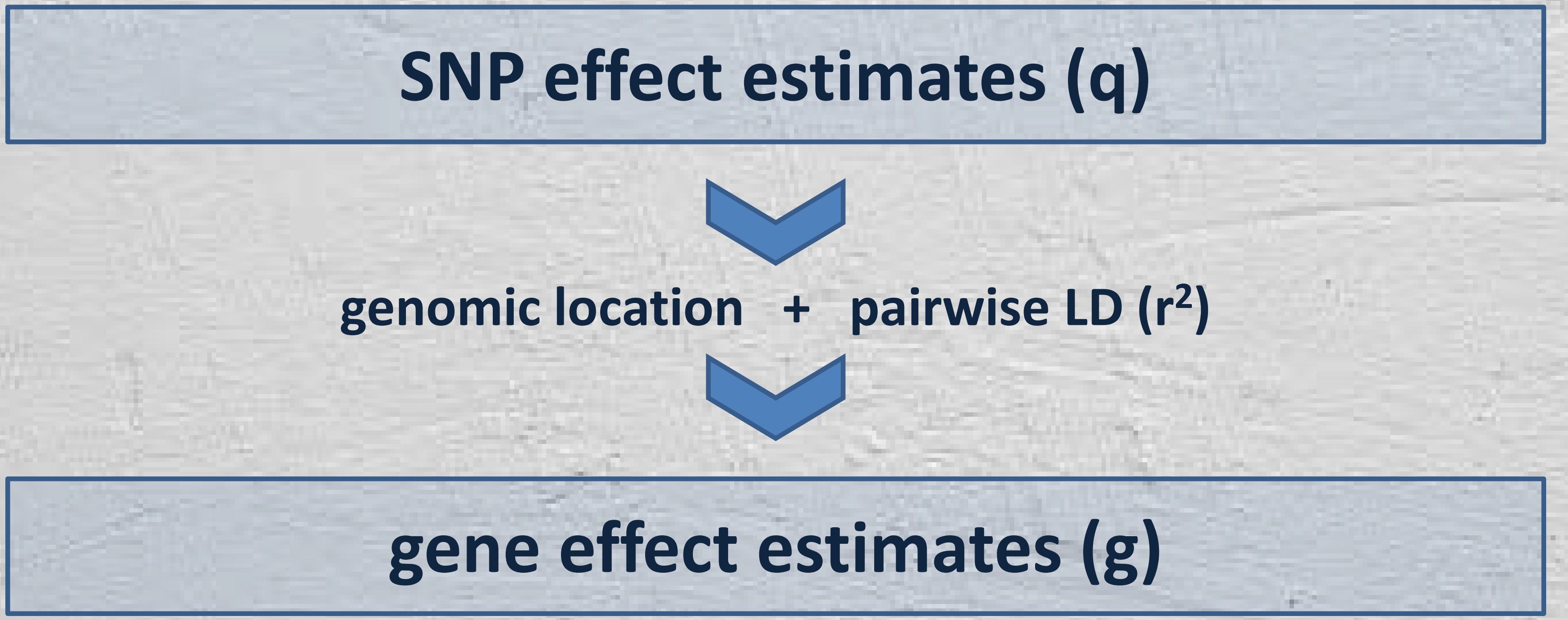
SNP effect estimation

$$y = \mu + Zq + e$$

- y deregressed EBV for milk yield
- µ general mean
- q additive SNP $\sim N\left(0, \mathbf{I}\sigma_q^2\right)$
- Z ∈{ -1, 0, 1 }
- e residual $\sim N\left(0, \mathbf{D}\sigma_e^2\right)$

genomic evaluation
04.2012

Gene effect estimation



SNP effect estimates (q)

genomic location + pairwise LD (r^2)

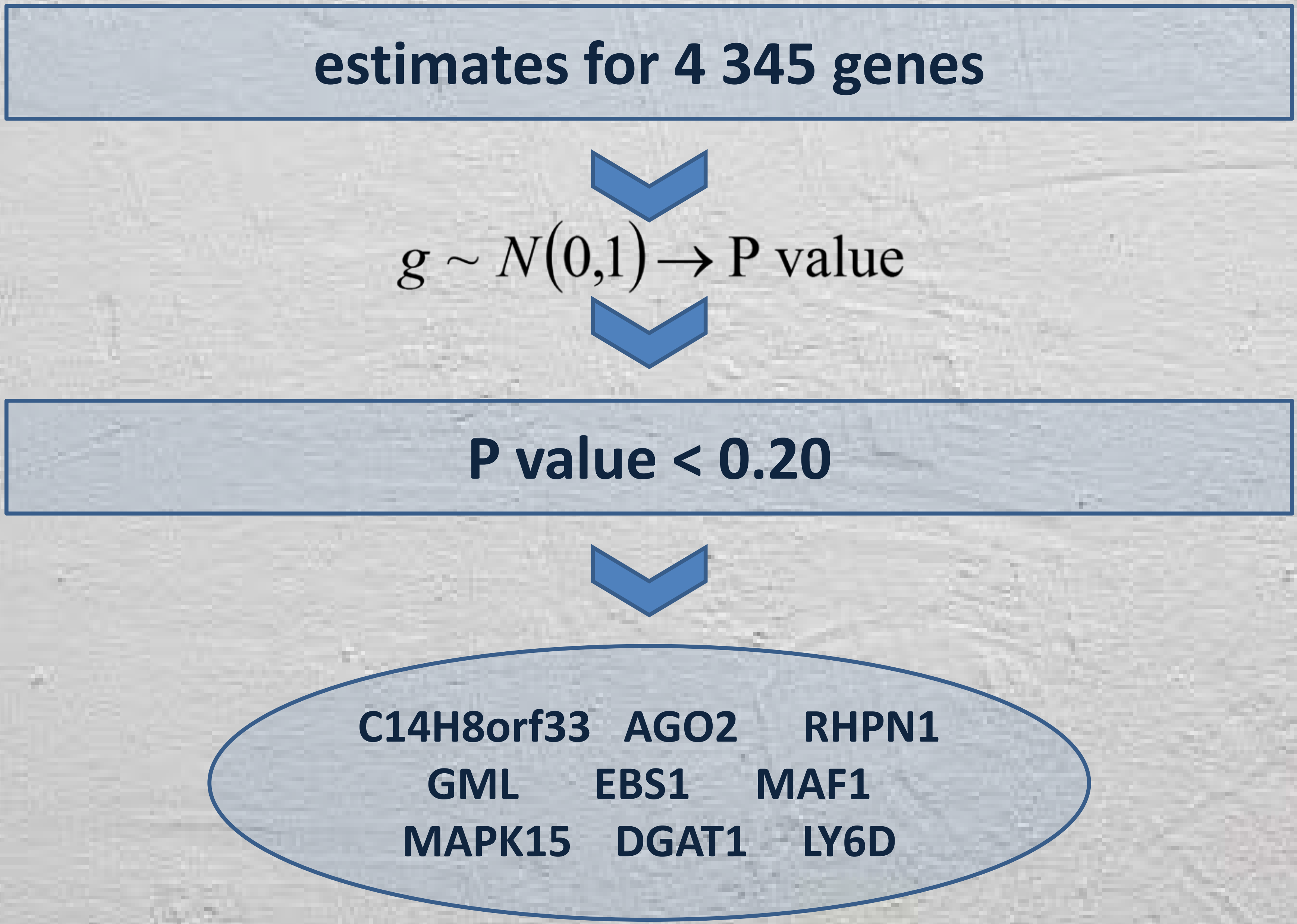
gene effect estimates (g)

$$g = \frac{\sum \hat{q}_i}{\sigma_g}$$

$$\sigma_g^2 = \sum \sigma_{qi}^2 + 2\sum\sum \sigma_{qij}$$

$$\sigma_g^2 = n\sigma_q^2 + 2\sum_i \sum_{j>i} r_{ij}^2 \sigma_q^2$$

Gene selection



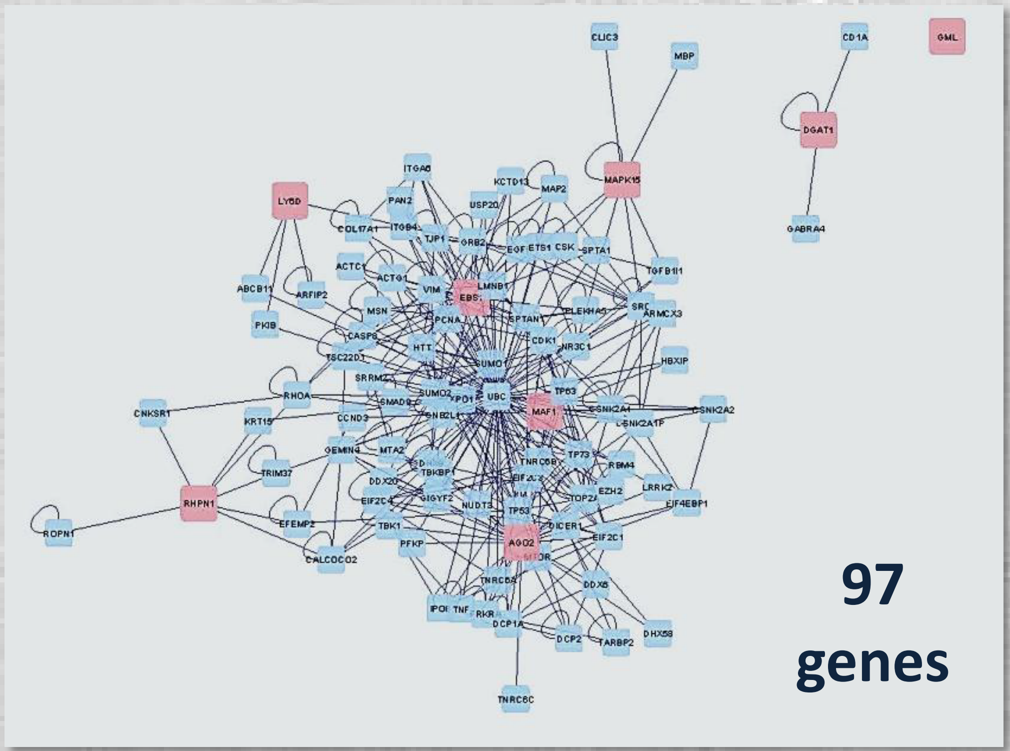
estimates for 4 345 genes

$$g \sim N(0,1) \rightarrow \text{P value}$$

P value < 0.20

C14H8orf33 AGO2 RHPN1
GML EBS1 MAF1
MAPK15 DGAT1 LY6D

Gene network



97 genes

Bisogenet, Martin et al. 2010 BMC Bioinformatics

retrieve functional information

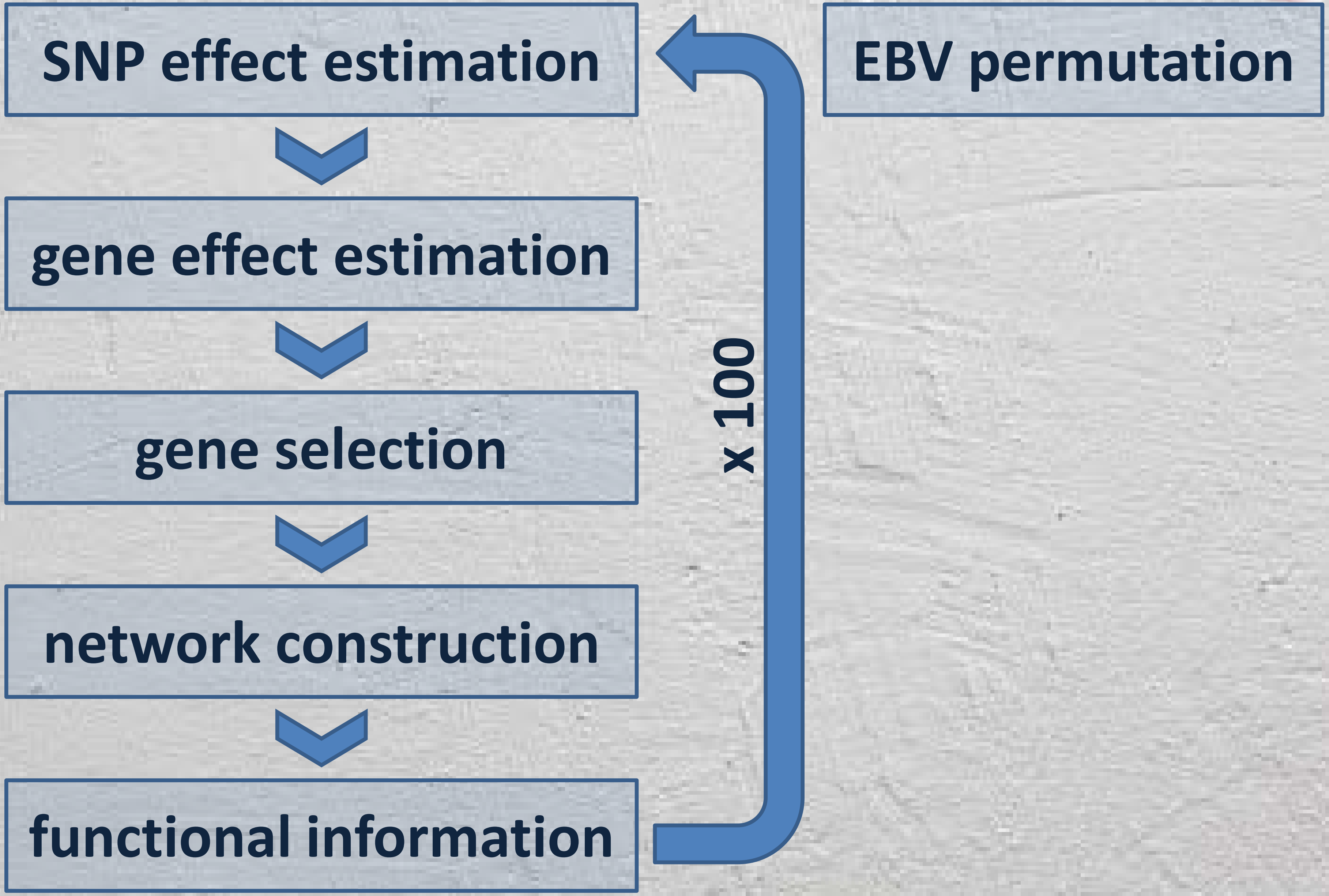
326 KEGG pathways

Kobas, Xie et al. 2011 Nucleid Acids Research

2 289 GO terms

Bisogenet, Martin et al. 2010 BMC Bioinformatics

Permutation



EBV permutation

SNP effect estimation

gene effect estimation

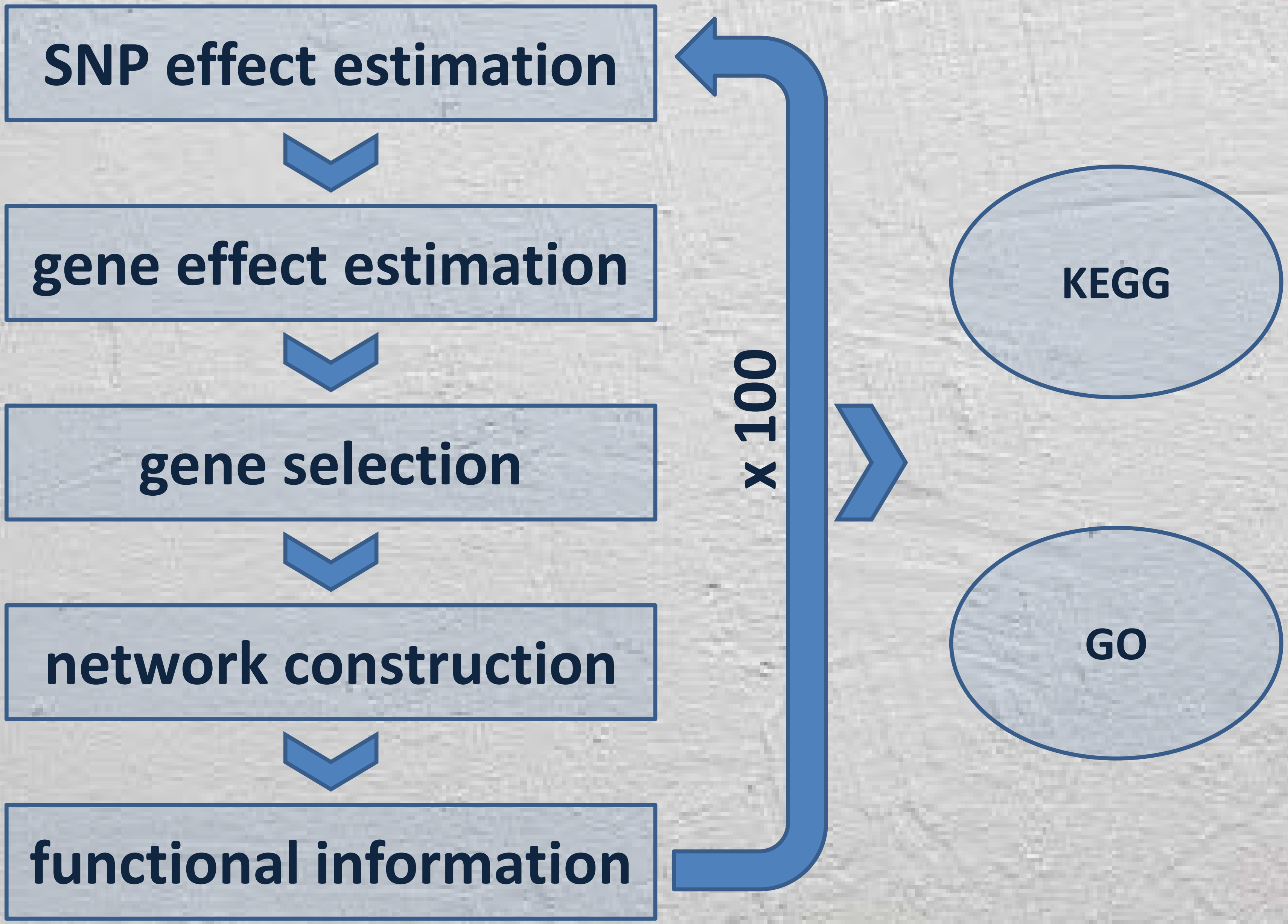
gene selection

network construction

functional information

x 100

Permutation



SNP effect estimation

gene effect estimation

gene selection

network construction

functional information

x 100

KEGG

GO

Validation

Odds Ratio for KEGG/GO

$H_0 : P(O) = P(P) \quad H_1 : P(O) \neq P(P)$

$$\ln(OR) = \ln\left(\frac{C_O / (N_O - C_O)}{C_P / (N_P - C_P)} \right)$$

- **original data**
- **permuted data** (pooled)

$\sim N\left(0, \sigma^2_{\ln(OR)}\right) \quad \rightarrow \quad \sim N(0,1) \quad \rightarrow$ **Bonferroni**

Results: GO

regulation of translation	→ $P<0.00001$
down regulation of translation involved in gene silencing by miRNA	→ $P<0.00001$
RNA-mediated gene silencing	→ $P<0.00001$
cytoplasmic mRNA processing body	→ $P<0.00001$
RNA-induced silencing complex	→ $P=0.00060$
double-stranded RNA binding	→ $P=0.00333$
down reg. of translational initiation	→ $P=0.01487$
pre-miRNA processing	→ $P=0.03088$
hemidesmosome assembly	→ $P=0.03630$

Results: KEGG

Carbohydrate Metabolism → glycolysis, lactogenesis → energy, lactose

- **Galactose metabolism** (30 genes) → P=0.01357
- **Pentose phosphate** (26) → P=0.03223
- **Fructose and mannose metabolism** (36) → P=0.03223
- **Measles** → P=0.04278
- **Dilated cardiomyopathy** → P=0.05933
- **p53 signaling pathway** → P=0.09567
- **Hypertrophic cardiomyopathy** → P=0.09567

Results: KEGG

- **Galactose metabolism** (30 genes) → P=0.01357
- **Pentose phosphate** (26) → P=0.03223

Immunogenesis → bacterial infection susceptibility

- **Measles** → P=0.04278
- **Dilated cardiomyopathy** → P=0.05933
- **p53 signaling pathway** → P=0.09567
- **Hypertrophic cardiomyopathy** → P=0.09567

Results: KEGG

- **Galactose metabolism** (30 genes) → P=0.01357
- **Pentose phosphate** (26) → P=0.03223
- [illegible]

Significant in gene set enrichment analysis

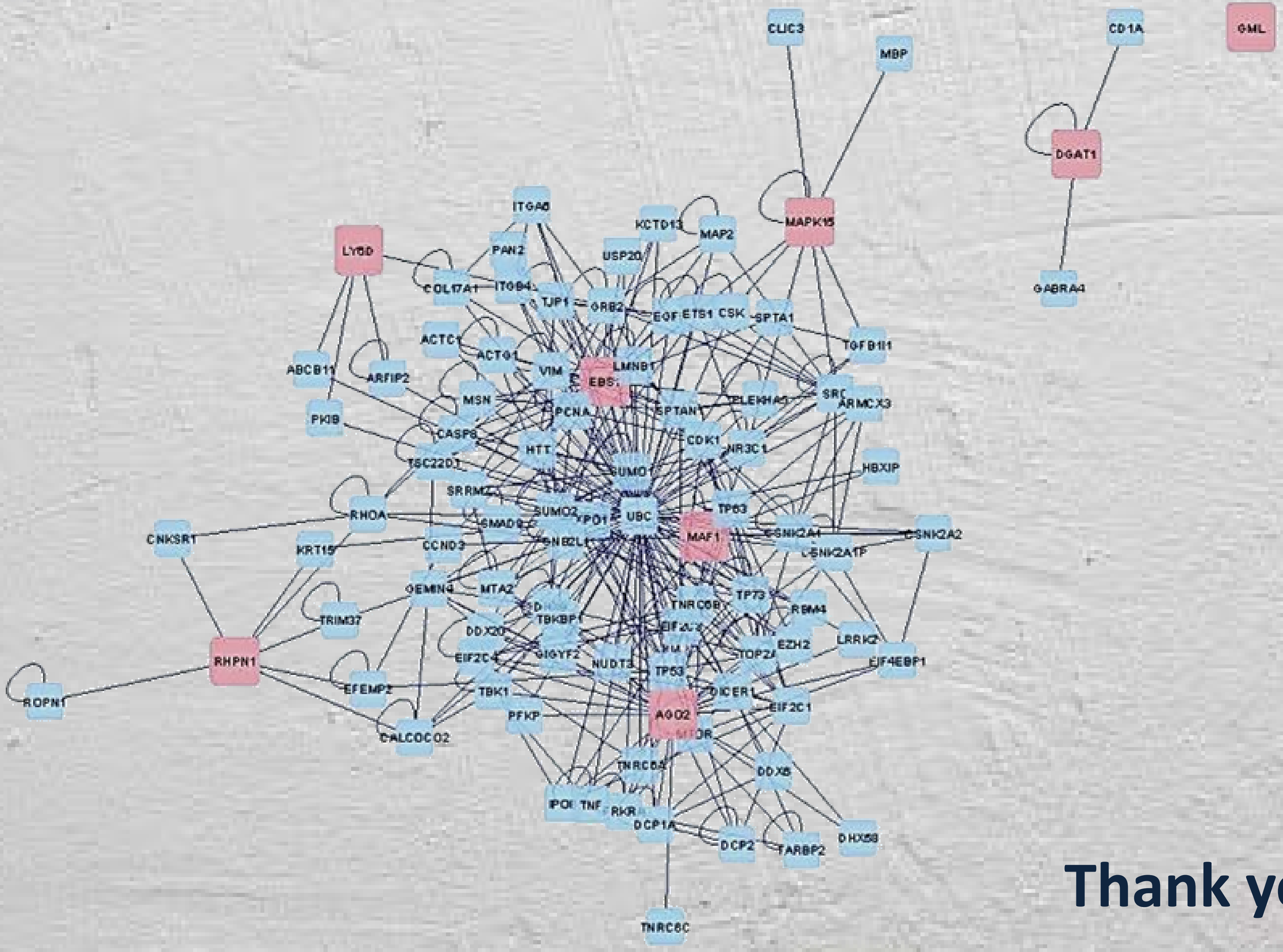
- **Measles** → P=0.04278
- **Dilated cardiomyopathy** → P=0.05933
- **p53 signaling pathway** → P=0.09567
- **Hypertrophic cardiomyopathy** → P=0.09567

Conclusions

1. **Pathway validation via EBV permutation**
   - time consuming
   - reflects sample size, informativeness and structure
2. **Identification of genes with small effects through pathways**
   - „logical" pathways identified
   - 50K poor resolution (not all genes represented)

Thanks

Katarzyna Wojdak-Maksymiec


CLIC3
CD1A
GML
MBP
DGAT1
MAPK15
GABRA4
ITGA6
KCTD13
MAP2
LY6D
PAN2
USP20
COL17A1
ITGB4
TJP1
GRB2
EGF
ETS1
CSK
SPTA1
ACTC1
ACTG1
TGFB1I1
ABCB11
ARFIP2
VIM
EBS
LMNB1
MSN
PLEKHA5
SRC
ARMCX3
PKIB
PCNA
SPTAN1
CASP8
CDK1
NR3C1
HTT
TSC22D1
HBXIP
SUMO1
SRRM2
RHOA
SUMO2
XPO1
UBC
TP63
SMAD9
MAF1
CSNK2A1
CSNK2A2
CNKSR1
KRT15
CCND3
SNB2L1
CSNK2A1P
GEMIN4
MTA2
TNRC6B
TP73
TRIM37
TBKBP1
RBM4
DDX20
EIF2C2
EZH2
LRRK2
EIF2C4
GIGYF2
NUDT3
TOP2A
EIF4EBP1
RHPN1
TP53
EFEMP2
TBK1
DICER1
ROPN1
EIF2C1
PFKP
AGO2
MTOR
CALCOCO2
TNRC6A
DDX6
IPO
TNF
RKR
DCP1A
DCP2
TARBP2
DHX58
TNRC6C

Thank you