Integrated transcriptional profiling and linkage analysis for mapping disease genes and regulatory gene networks analysis

Enrico Petretto

Research Fellow in Genomic Medicine

Imperial College Faculty of Medicine
enrico.petretto@imperial.ac.uk

MRC Clinical Sciences Centre
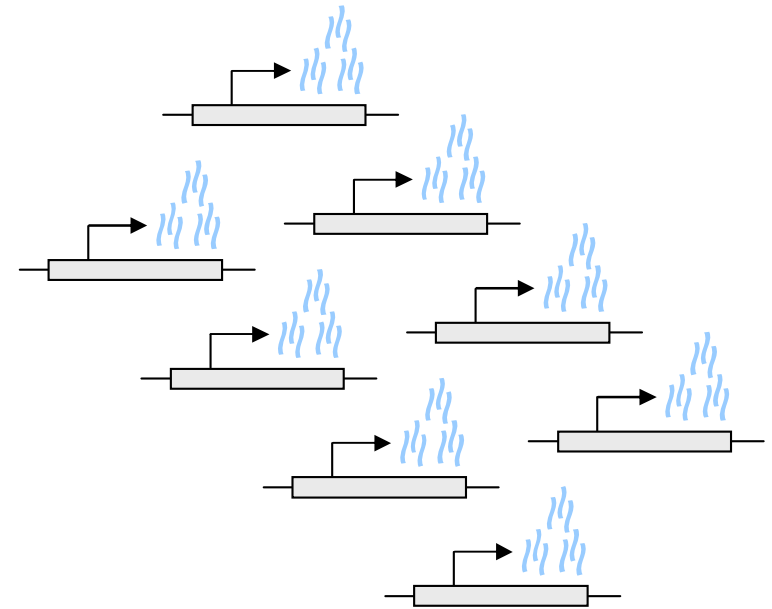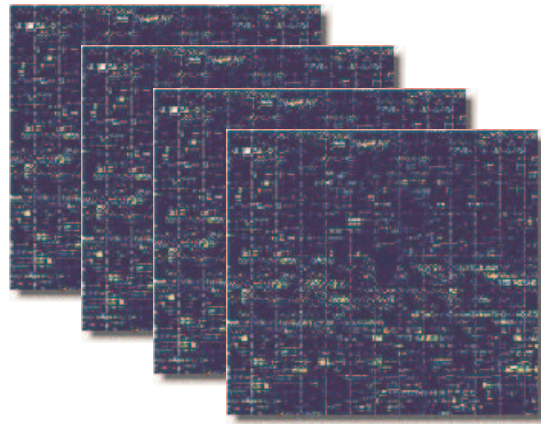
RESEARCH COUNCILS UK
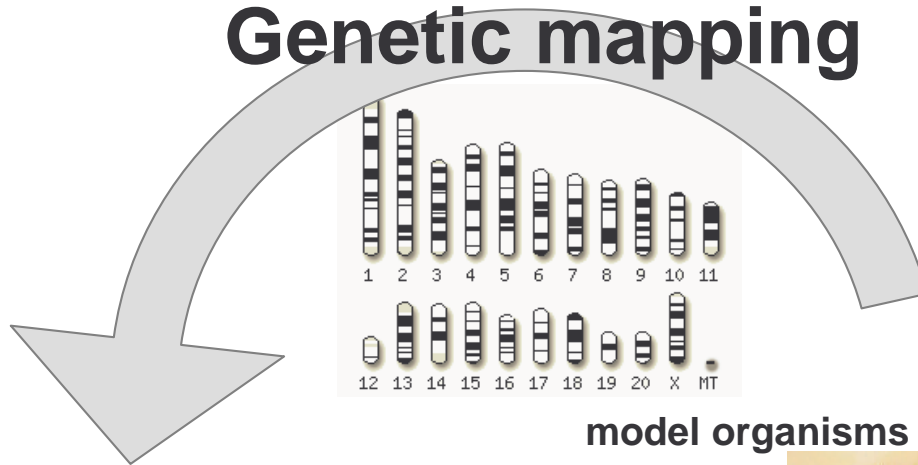
Imperial College London

Outline

- **Introduction: the biological framework**

  – Expression QTL mapping using animal models

  – eQTL analysis in multiple tissues

- **Integrating genome-wide eQTL data to identify gene association networks**

  – Data mining of eQTLs

  – Graphical Gaussian models (GGMs)

  – Example of identification of disregulated pathway
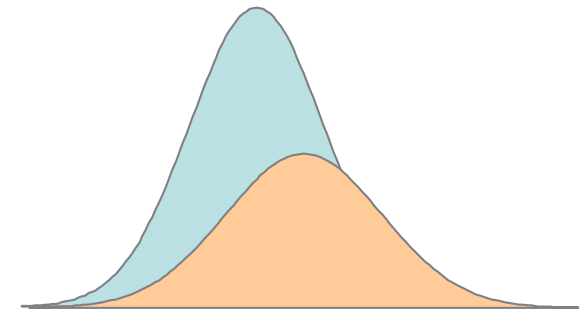
  – Master transcriptional regulator

Imperial College
London

Genetical Genomics

Genetic mapping

Expression QTLs
genetic determinants
of gene expression

model organisms

quantitative variation of mRNA levels
in a segregating population

Imperial College
London

The rat is among the leading model species for research in physiology, pharmacology, toxicology and for the study of genetically complex human diseases

Spontaneously Hypertensive Rat (SHR): A model of the metabolic syndrome

- Spontaneous hypertension

- Decreased insulin action

- Hyperinsulinaemia

- Central obesity

- Defective fatty acid metabolism

- Hypertriglyceridaemia

SHR/Kyo
動物番号 21

Specialized tools for genetic mapping:
Rat Recombinant Inbred (RI) strains

Spontaneously Hypertensive Rat
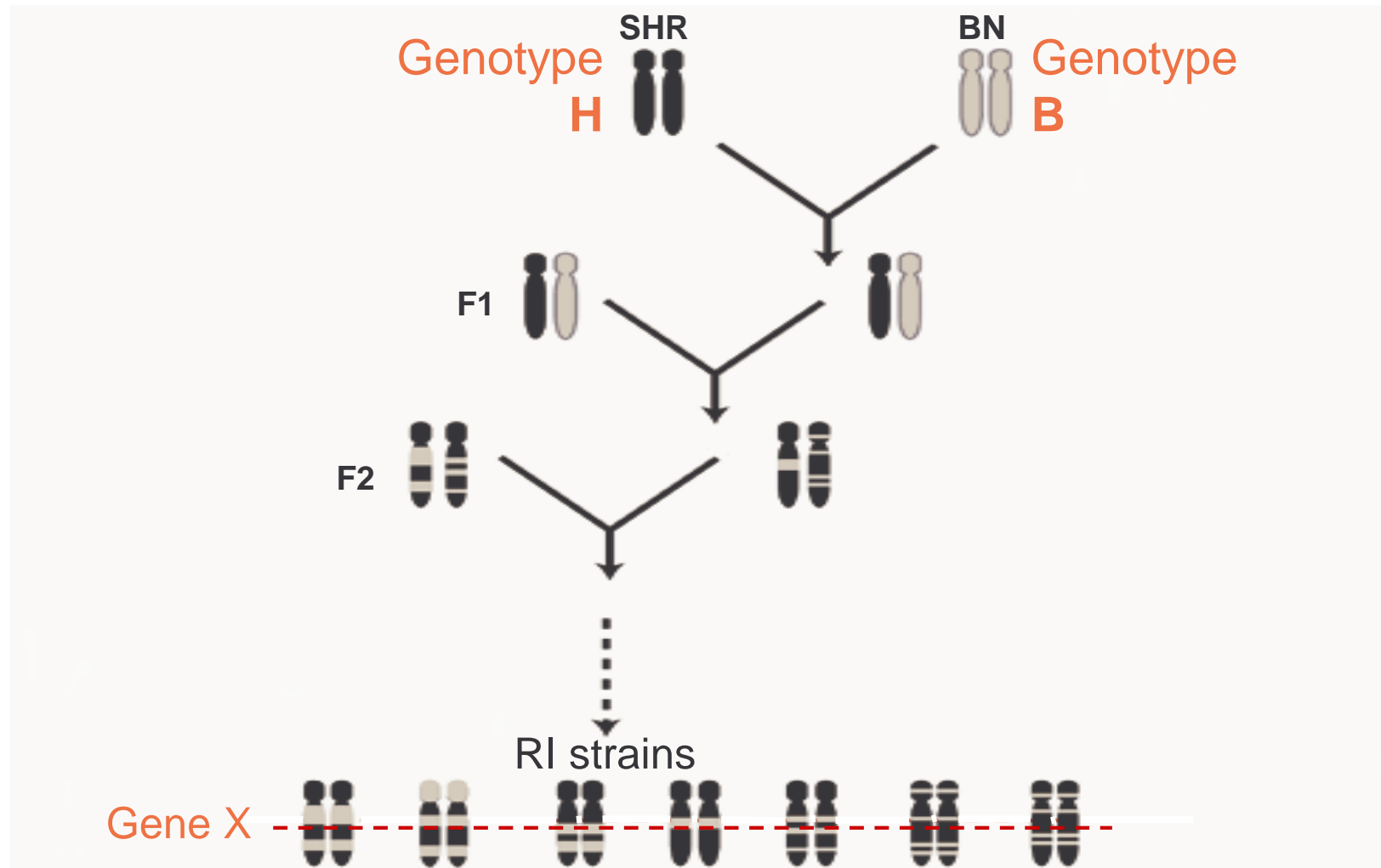
Normotensive Rat (BN)

Mate two inbred strains

F1 offspring are identical

F2 offspring are different
(due to recombination)

Brother sister mating over >20
generations to achieve
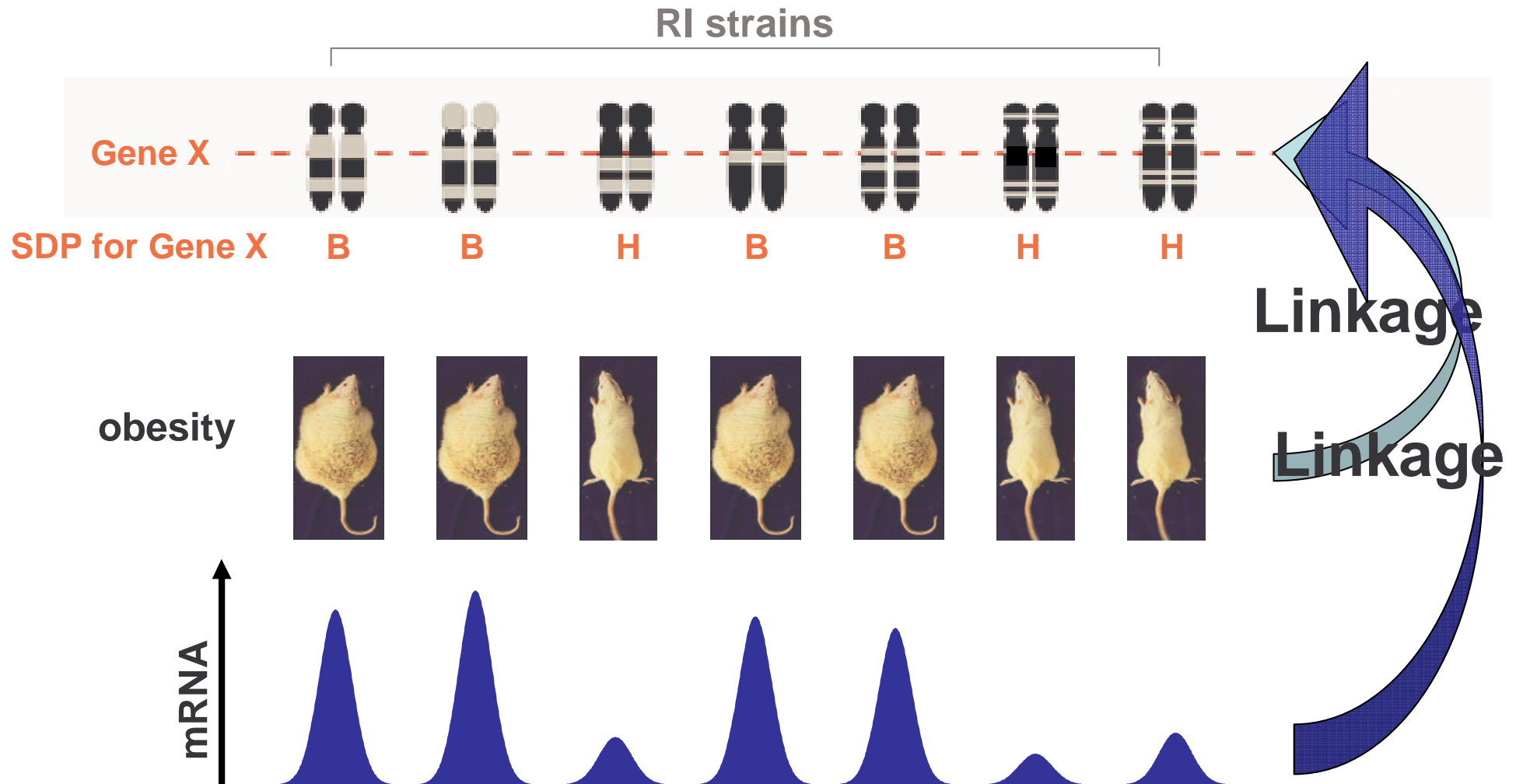homozygosity at all genetic loci

Pravenec et al. *J Hypertension*, 1989

Cumulative, renewable resource for phenotypes and genetic mapping

Strain Distribution Pattern for Gene X

Mapping of QTLs
compare strain distribution pattern of markers and traits

RI strains

Gene X

SDP for Gene X   B   B   H   B   B   H   H

Linkage

Linkage

obesity

mRNA

Gene expression analysis in the Rat

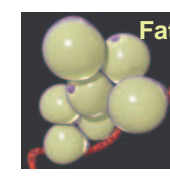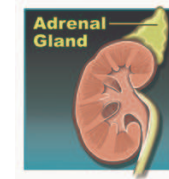30 RI strains + 2 parental strains
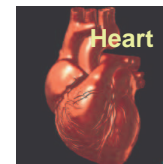
↓

4 animals per strain (no pooling)

↓

Expression profiling

↓

Affymetrix RAE230A

Affymetrix RAE230_2

640 microarray data sets
~ 16,000 probe sets per array (fat, kidney, adrenal)
~ 30,000 probe sets per array (heart, skeletal muscle)

eQTL Linkage Analysis

- For each probe set on the microarray, expression profiles were regressed against all 1,011 genetic markers

Multiple testing issues

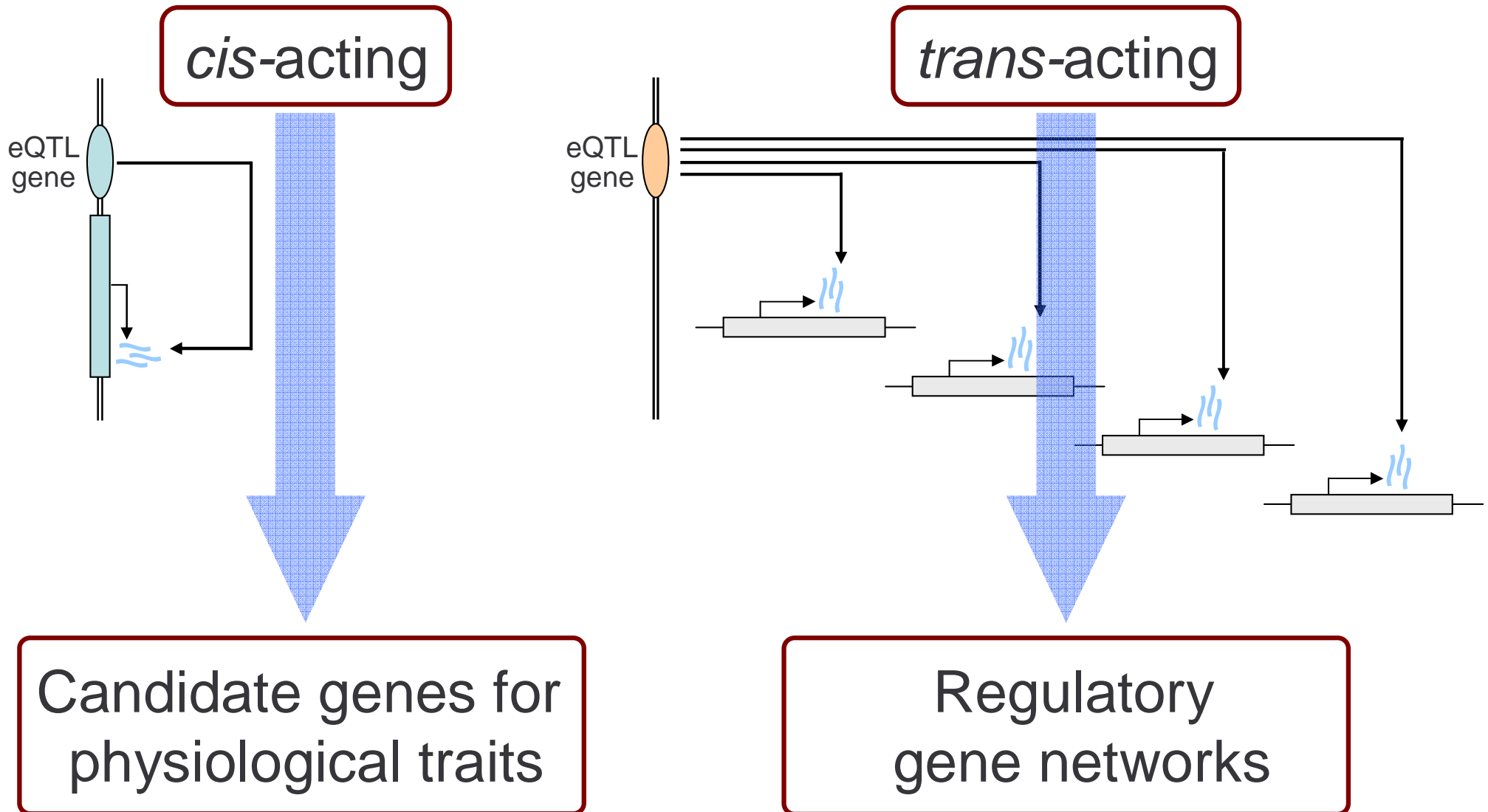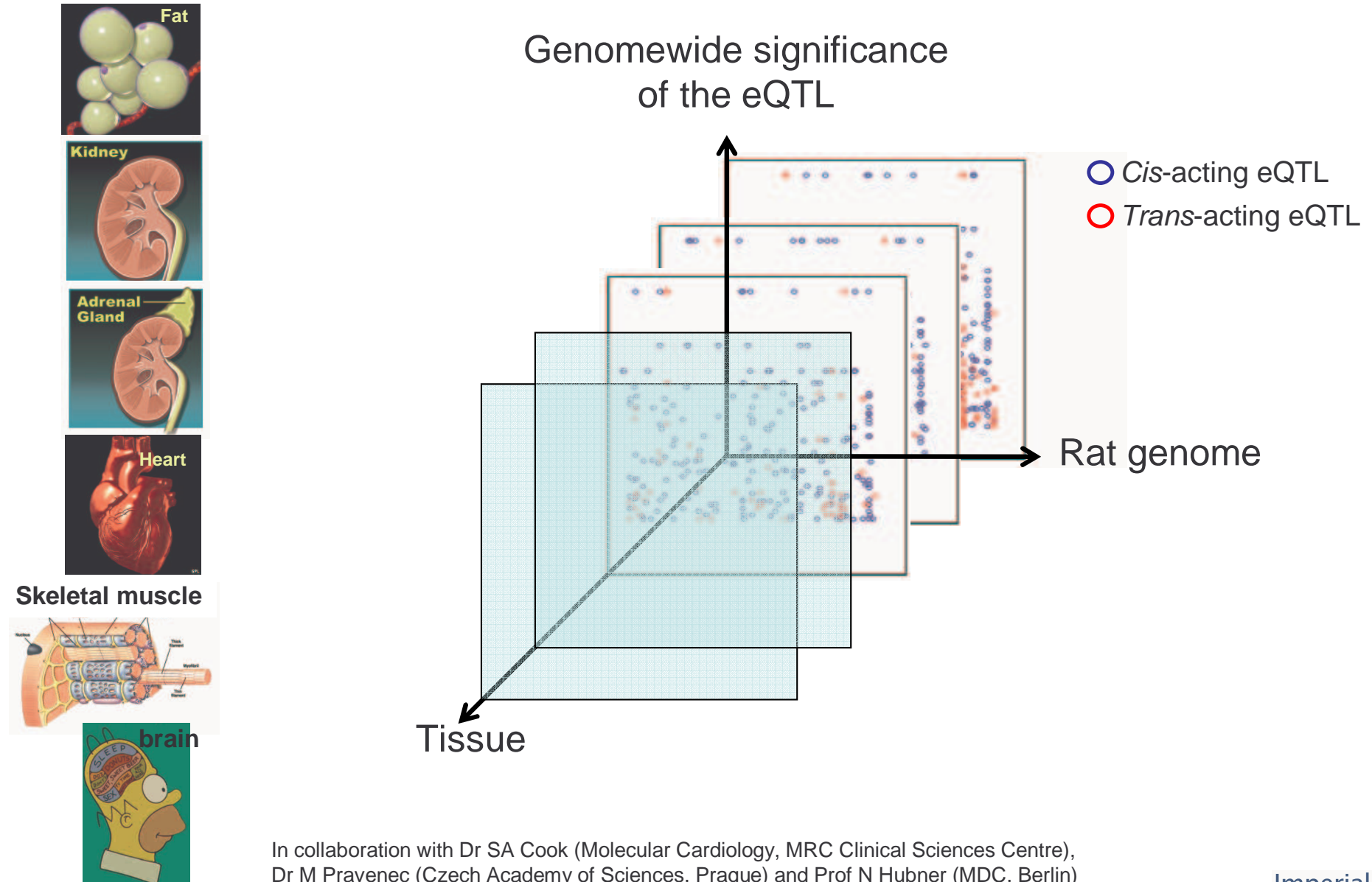| 1,011 genetic markers	15,923 probe sets
Evaluate the linkage statistics for each genetic marker and use permutation testing to provide genome-wide corrected P-values	Expected proportion of false positives among the probe sets called significant in the linkage analysis (False Discovery Rate*)

* Storey 2000

cis- and *trans*-acting eQTLs

eQTL datasets in the rat model system

Fat

Kidney

Adrenal Gland

Heart

Skeletal muscle

brain

Genomewide significance
of the eQTL

○ *Cis*-acting eQTL
○ *Trans*-acting eQTL

Rat genome

Tissue

Imperial College
London

Genetic architecture of genetic variation in gene expression

+ *cis*-eQTL

+ *trans*-eQTL

trans-eQTLs:
small genetic effect

cis-eQTLs:
big genetic effect
highly heritable

Petretto *et al.* 2006 *PLoS Genet*
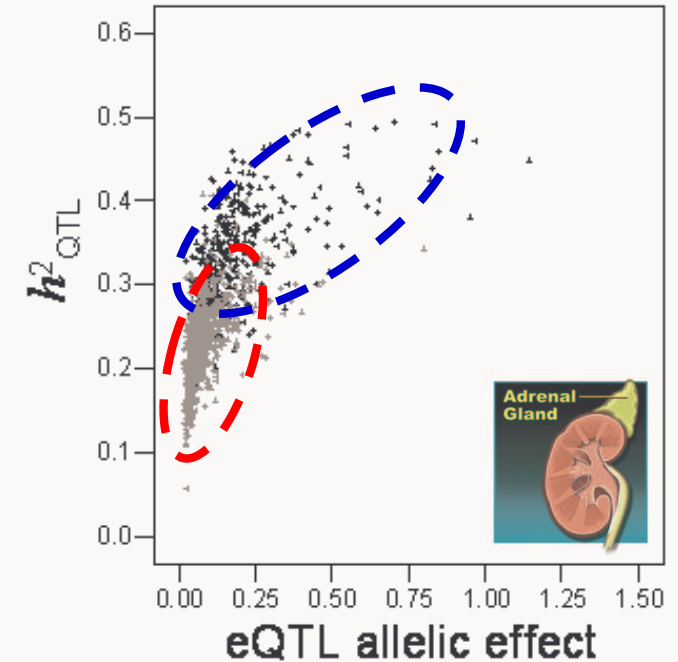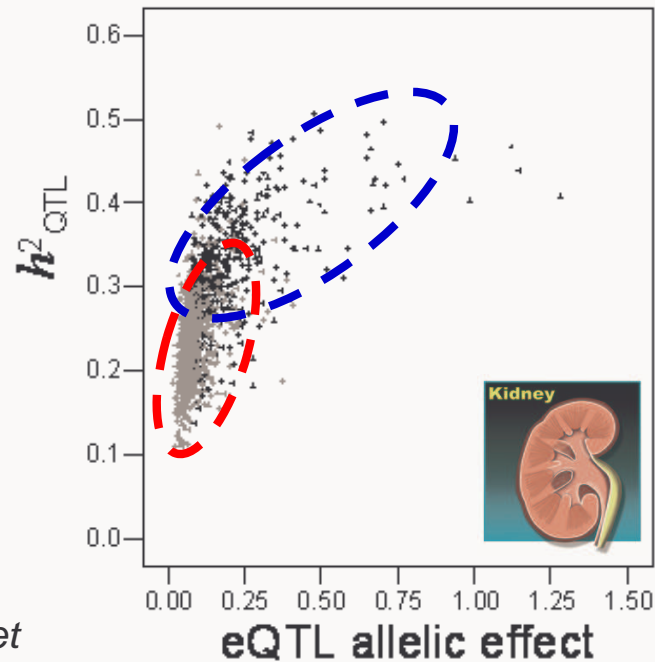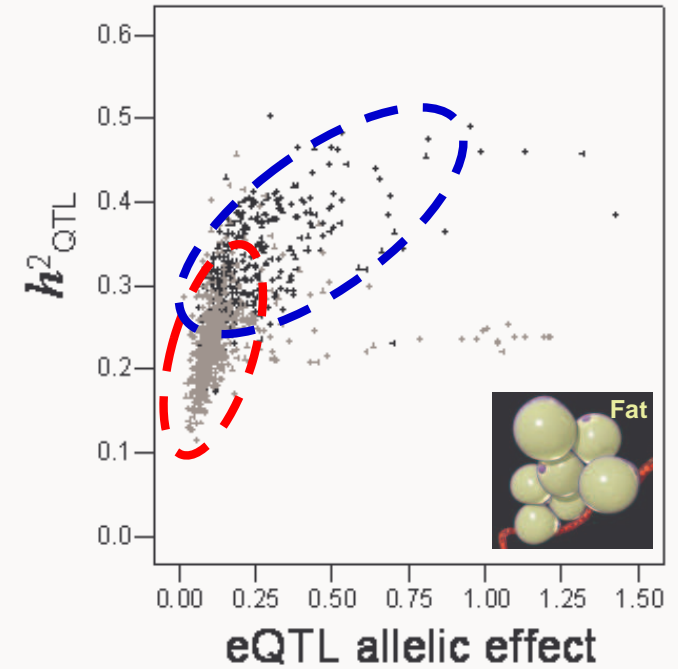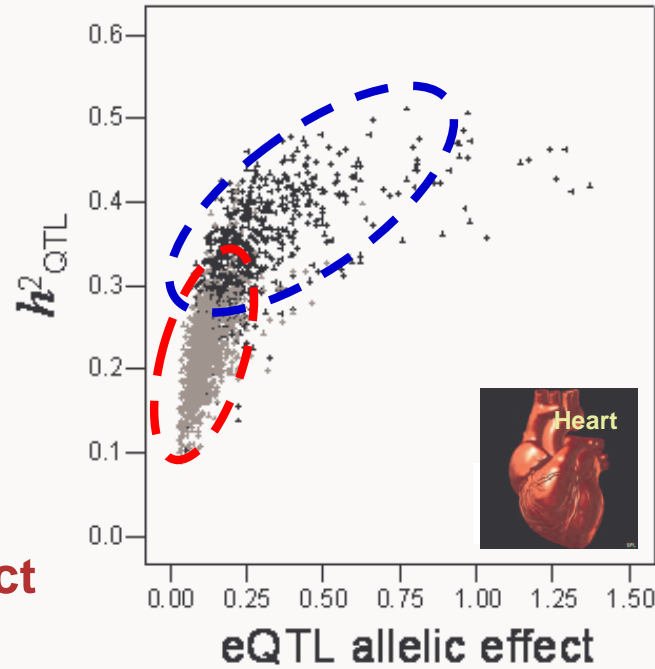
FDR for *cis-* and *trans-*eQTLs

homogeneous tissues — ⬇ FDR

heterogeneous tissues — ⬆ FDR

Petretto *et al.* 2006 *PLoS Genet*

Imperial College
London
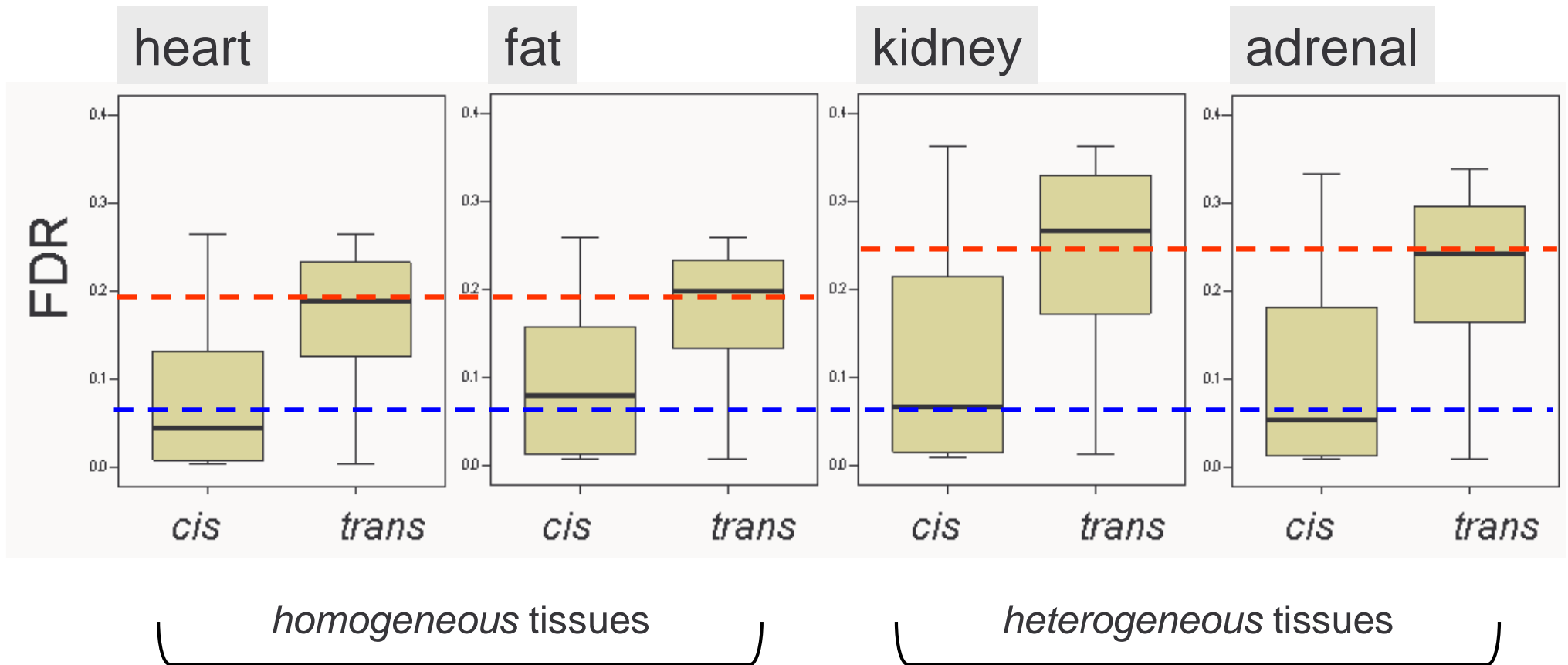
trans-eQTLs hot-spots

Trans-eQTLs
- ● heart (orange)
- ● fat (blue)
- ● adrenal (green)
- ● kidney (red)

$P_{GW} < 0.05$

Rat chromosome 8

tissue-specific clusters

Master transcriptional regulator ?

not tissue-specific cluster
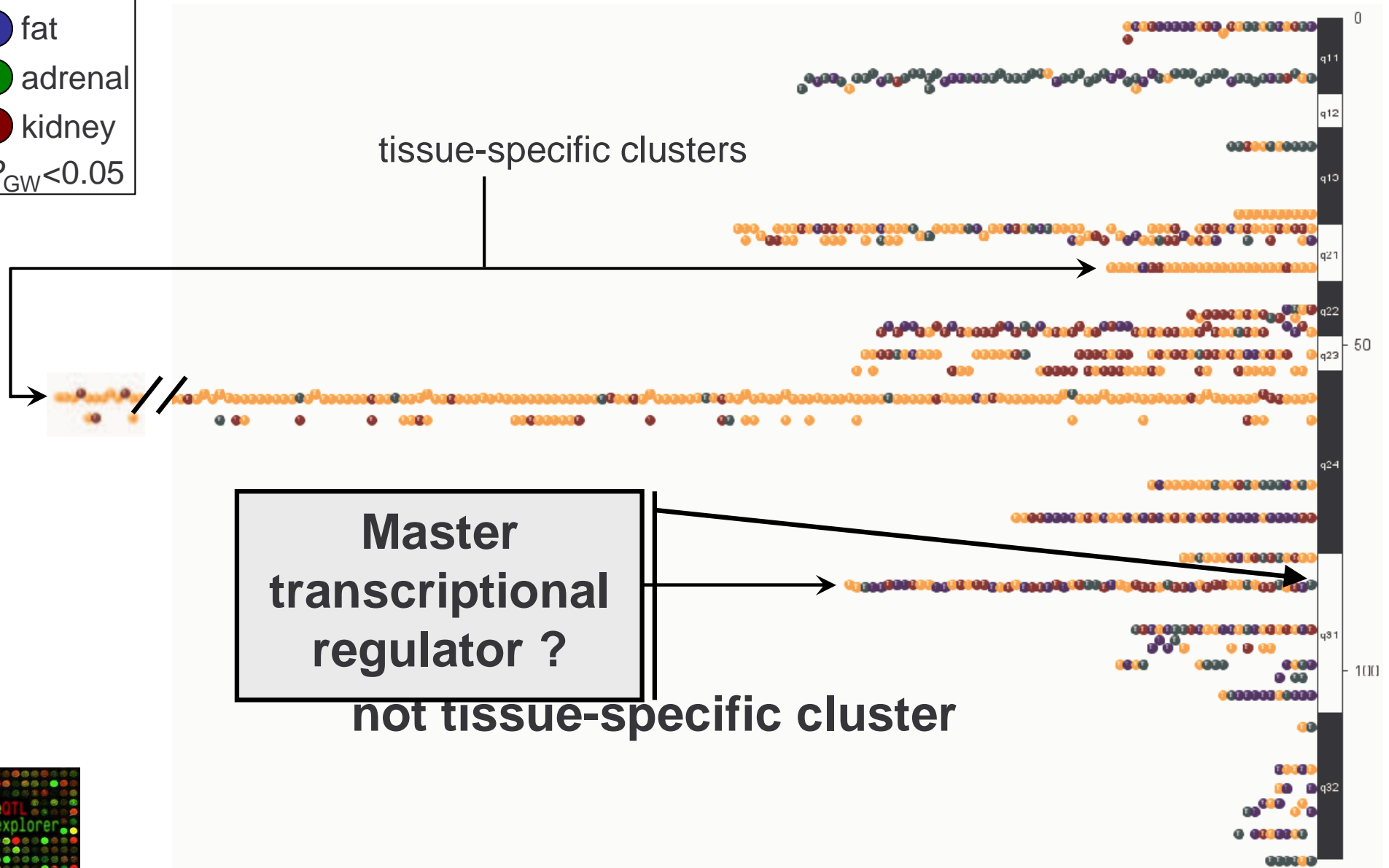
Strategy to identify master transcriptional regulators

Gene expression

Genetic markers

eQTLs

Data mining

trans

cis

GGMs

TF binding data

Functional Analysis (GSEA, etc.)

Association networks

Multi-tissues

Model for master transcriptional regulator

genetic variant

cis-linked gene

Transcription Factor (TF) activity profile

Expression of *trans*-linked genes
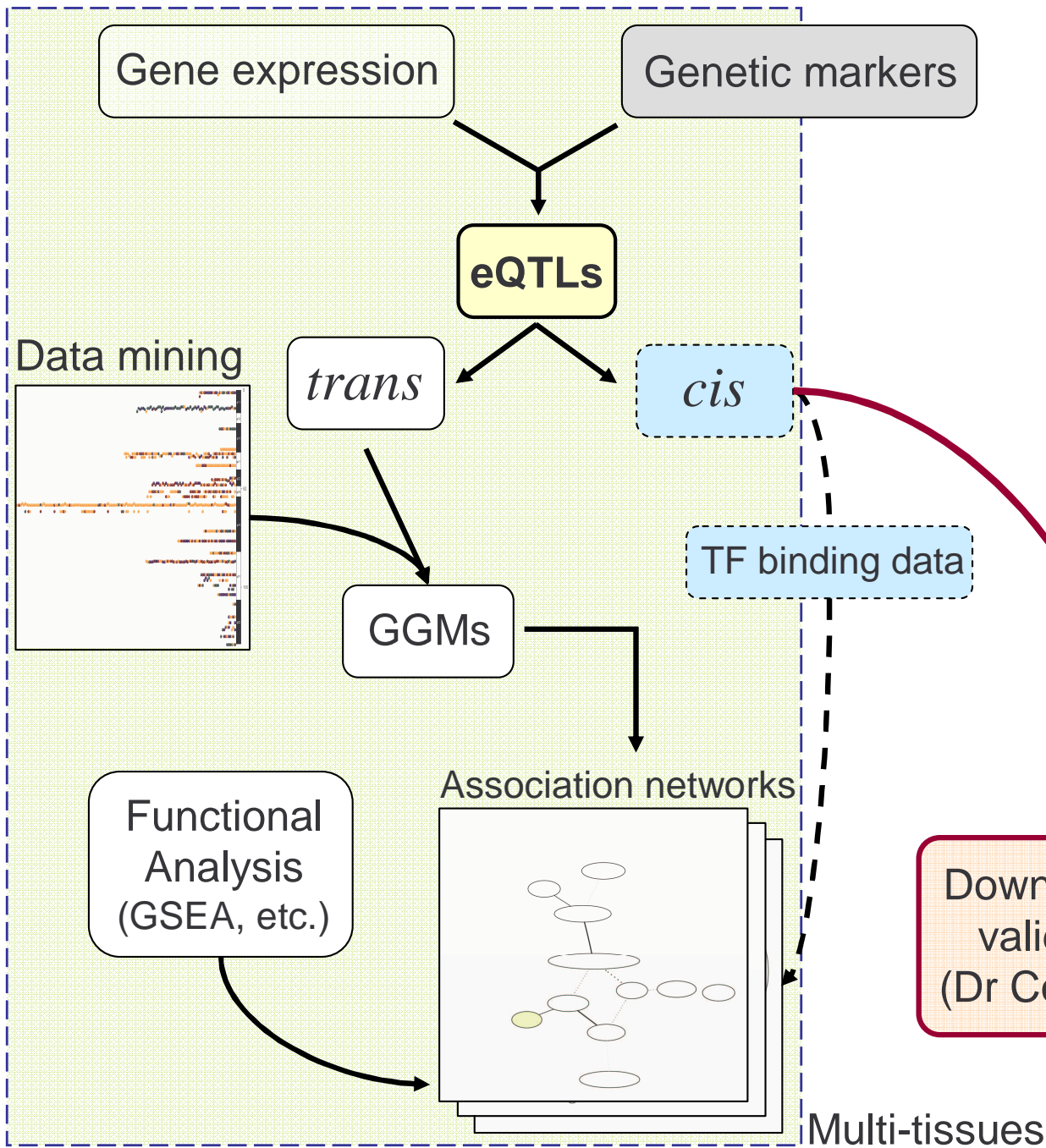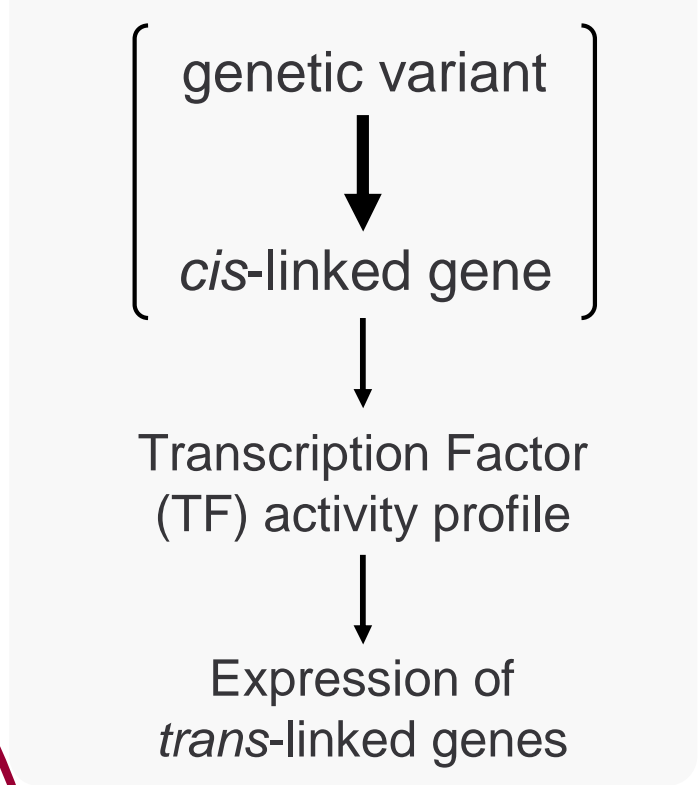
Downstream functional validation in the lab (Dr Cook / Prof Aitman)

Imperial College London

GGMs

- Partial correlation matrix $\quad \Pi = (\pi_{ij})$
- Inverse of variance covariance matrix P

$$\Omega = (\omega_{ij}) = P^{-1}$$

$$\pi_{ij} = -\omega_{ij} / (\omega_{ii}\,\omega_{jj})^{\frac{1}{2}}$$

- small n, large p
- Regularized covariance matrix estimator by shrinkage (Ledoit-Wolf approach)
  - Guarantees positive definiteness

Schafer and Strimmer 2004, Rainer and Strimmer 2007

Partial correlation graphs

- Multiple testing on all partial correlations
  - Fitting a mixture distribution to the observed partial correlations (p)

$$f(p) = \eta_0 f_0(p;\kappa) + \eta_A f_A(p) \qquad \eta_0 + \eta_A = 1, \quad \eta_0 \gg \eta_A$$

$$f_0(p;\kappa) = (1-p^2)^{(\kappa-3)/2} \frac{\Gamma\left(\frac{\kappa}{2}\right)}{\pi^{1/2}\,\Gamma\left(\frac{\kappa-1}{2}\right)}$$

uniform [-1, 1]

$$\hat{\eta}_0\,,\ \hat{\kappa}$$

$$\text{Prob (non-zero edge|p)} = 1 - \frac{\hat{\eta}_0\,f_0(p;\hat{\kappa})}{f(p)}$$

Schafer and Strimmer 2004, Rainer and Strimmer 2007

GGMs

Infer partial ordering of the node

- Standardized partial variances (SPV_i)
- Proportion of the variance that remains unexplained after regressing against all other variables
- Log-ratios of standardized partial variances $B = (SPV_i / SPV_j)^{1/2}$

Log(B) |rest = 0

undirected

Log(B) |rest ≠ 0

directed

exogenous variable

bigger SPV

endogenous variable

smaller SPV
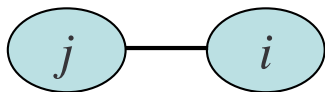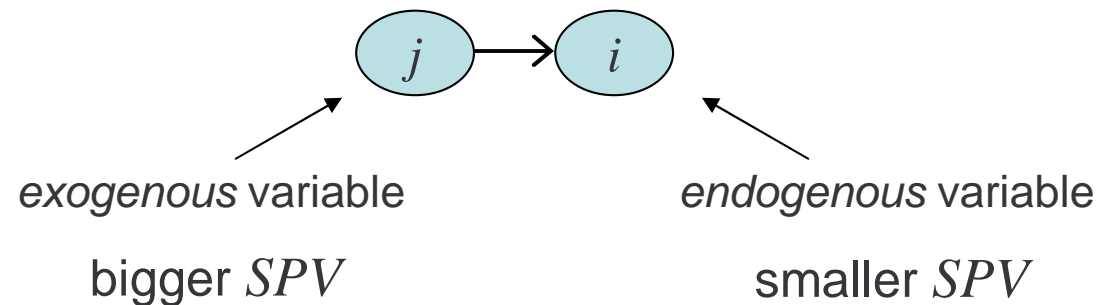
Inclusion of a directed edge into the network is conditional on a non-zero partial correlation coefficient

Schafer and Strimmer 2004, Rainer and Strimmer 2007

Imperial College
London

Hypothesis driven analysis

1. Gene expression levels under genetic control (i.e., 'structural' genetic perturbation)

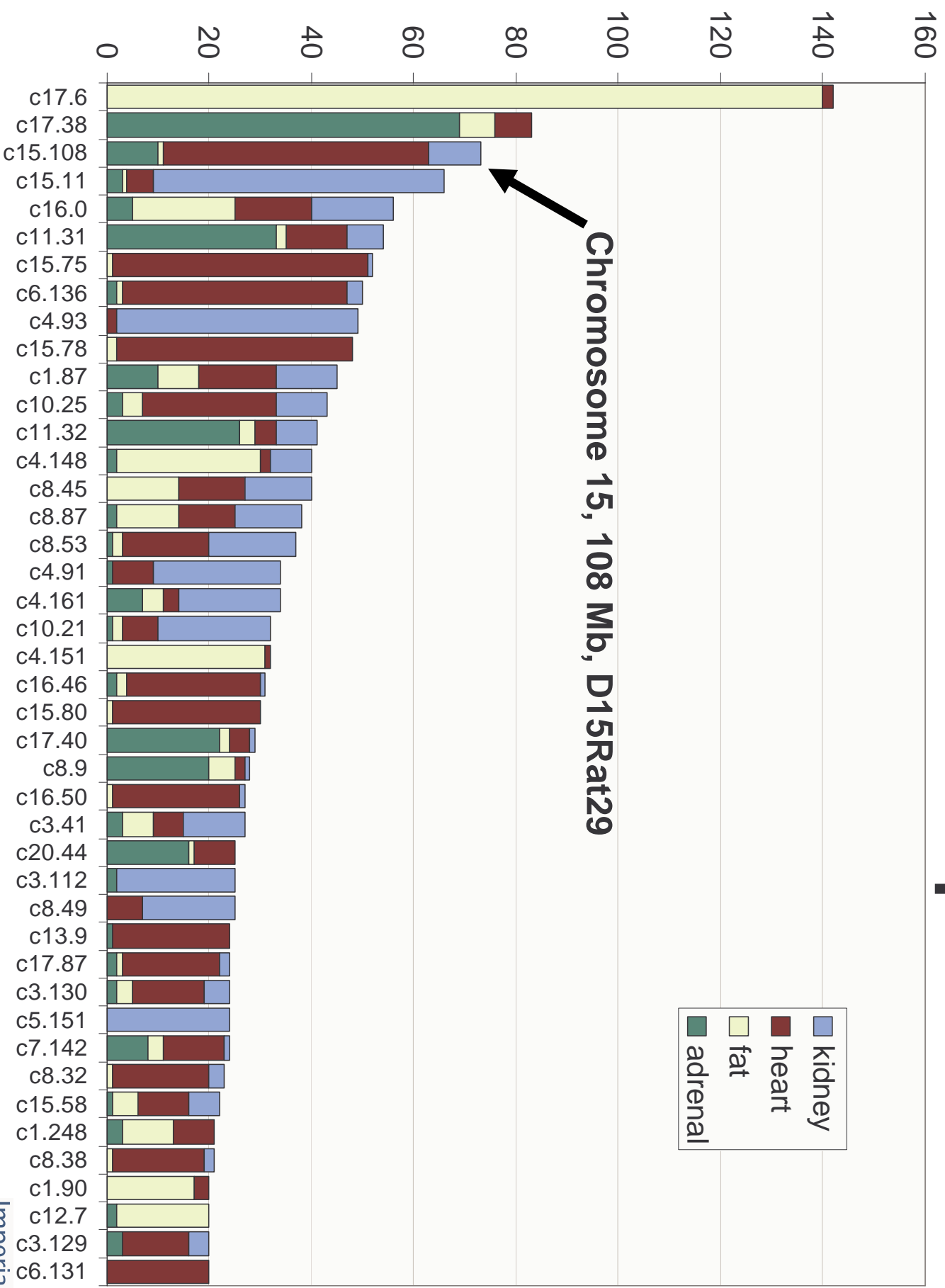2. Co-expression of *trans*-eQTLs point to common regulation by a single gene

Graphical Gaussian models

- Detect conditionally dependent *trans*-eQTL genes
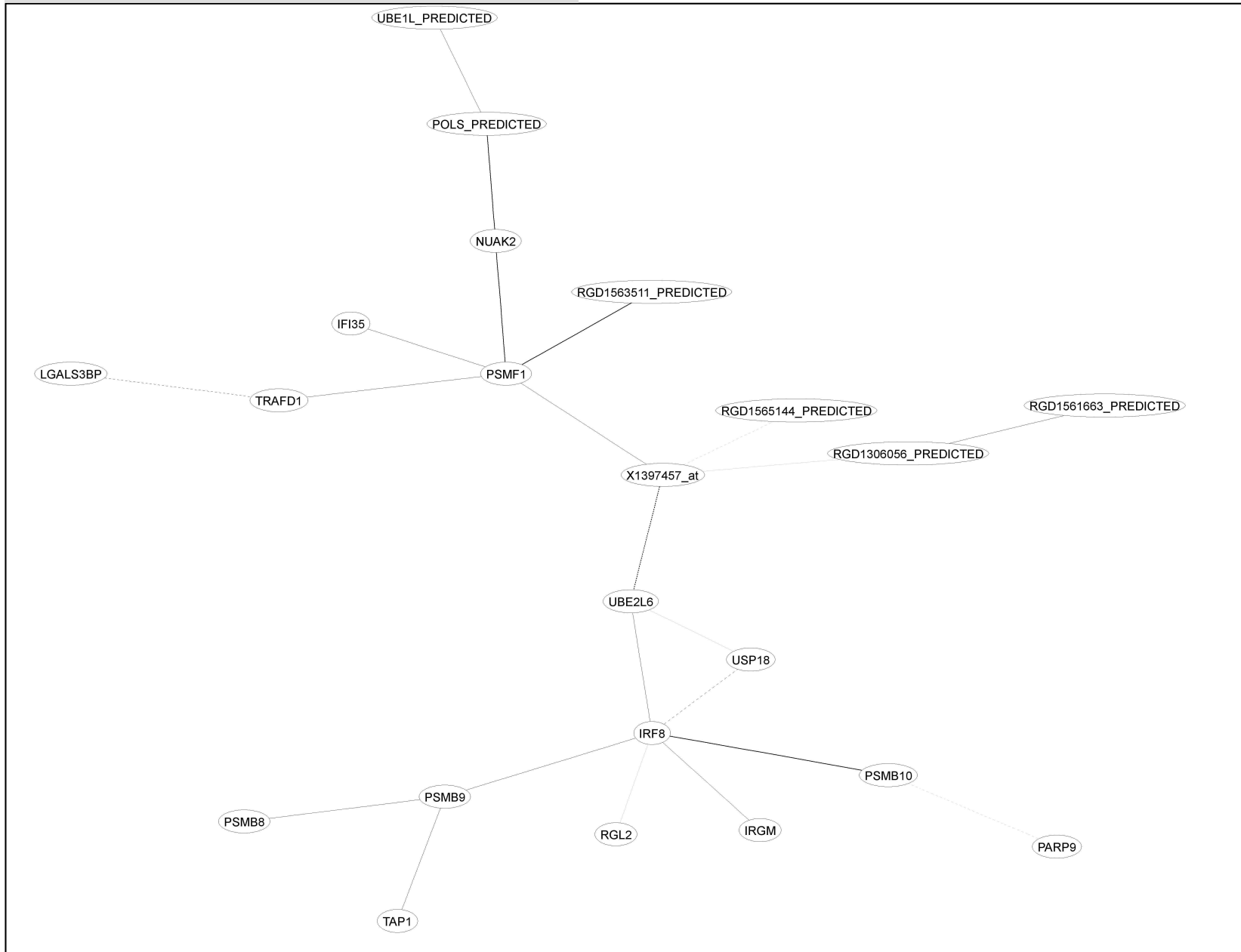
- Infer partial ordering of the nodes (directed edges)

Imperial College London

trans-eQTLs hot spots

Chromosome 15, 108 Mb, D15Rat29

Locus (chromosome.Mb)

Legend: kidney, heart, fat, adrenal

Loci (top to bottom): c17.6, c17.38, c15.108, c15.11, c16.0, c11.31, c15.75, c6.136, c4.93, c15.78, c1.87, c10.25, c11.32, c4.148, c8.45, c8.87, c8.53, c4.91, c4.161, c10.21, c4.151, c16.46, c15.80, c17.40, c8.9, c16.50, c3.41, c20.44, c3.112, c8.49, c13.9, c17.87, c3.130, c5.151, c7.142, c8.32, c15.58, c1.248, c8.38, c1.90, c12.7, c3.129, c6.131

Imperial College London

Heart tissue, *trans*-eQTLs hot-spot (chromosome 15)

posterior probability for non-zero edge 0.8

Heart tissue, *trans*-eQTLs hot-spot (chromosome 15)

Relaxing the threshold…

LY6C

LOC362934

STAT1_c

Involved in the transport of antigens from the cytoplasm to the endoplasmic reticulum for association with MHC class I molecules

STAT1_a

PARP9

Signal transducer / activator of transcription
IFN gamma activated, drive expression of the target genes, inducing a cellular antiviral state

RT1_S3_b

degradation of cytoplasmic antigens for MHC class I antigen presentation pathways

STAT1_b

TAP1

RT1_S3_a

PSMB9   PSMB10

PSMB8

IRF8   RGL2

TYKI_PREDICTED   USP18   RGD1561663_PREDICTED

MHC class I antigen antigen processing and presentation

RGD1306056_PREDICTED

UBE2L6   RGD1310093_PREDICTED

NUAK2   POLS_PREDICTED

IGTP   X1397457_at   UBE1L_PREDICTED

TRAFD1

Ddx58   PSMF1   RGD1565144_PREDICTED
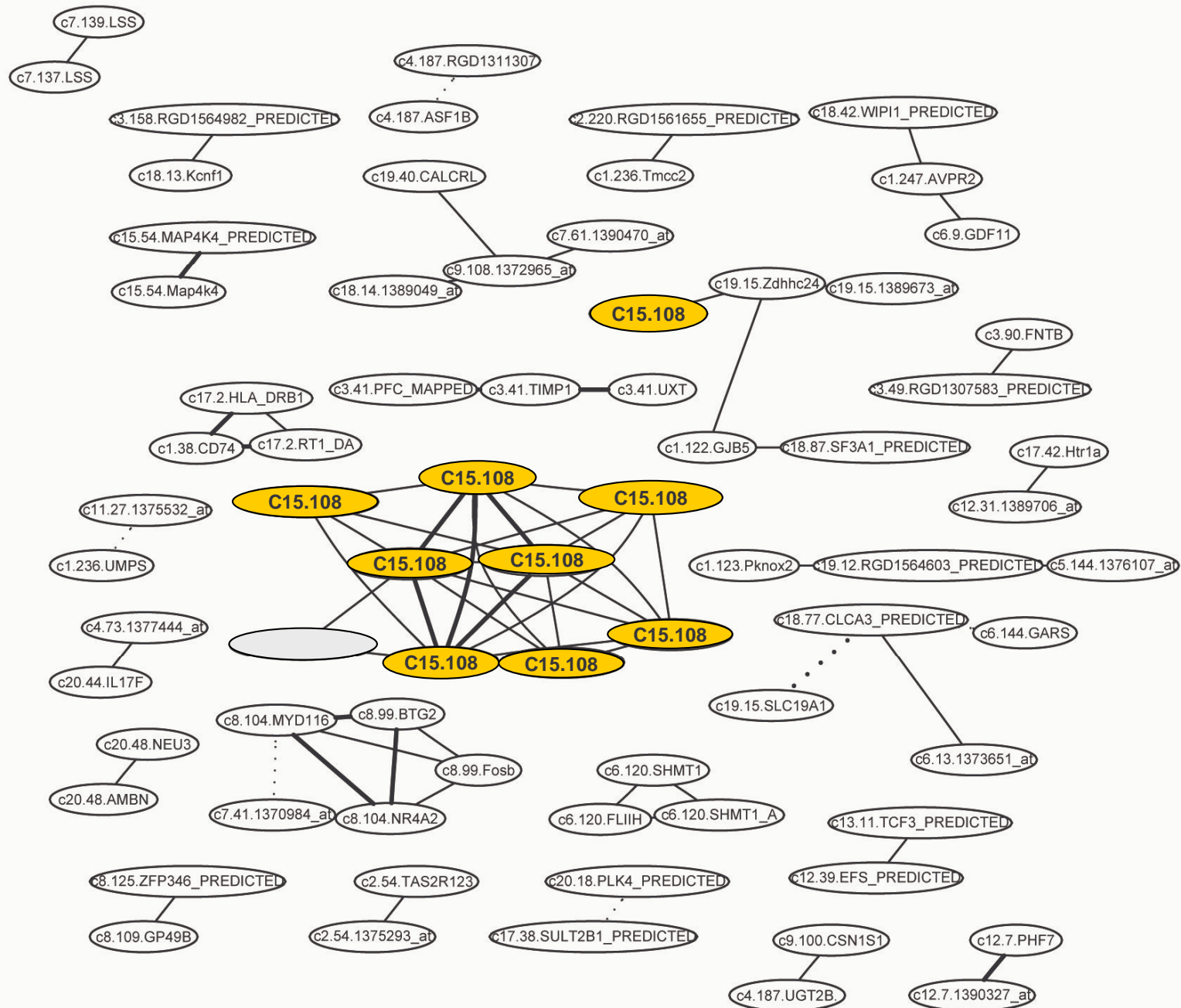
LGALS3BP

RGD1563511_PREDICTED   IFI35   Isgf3g_b
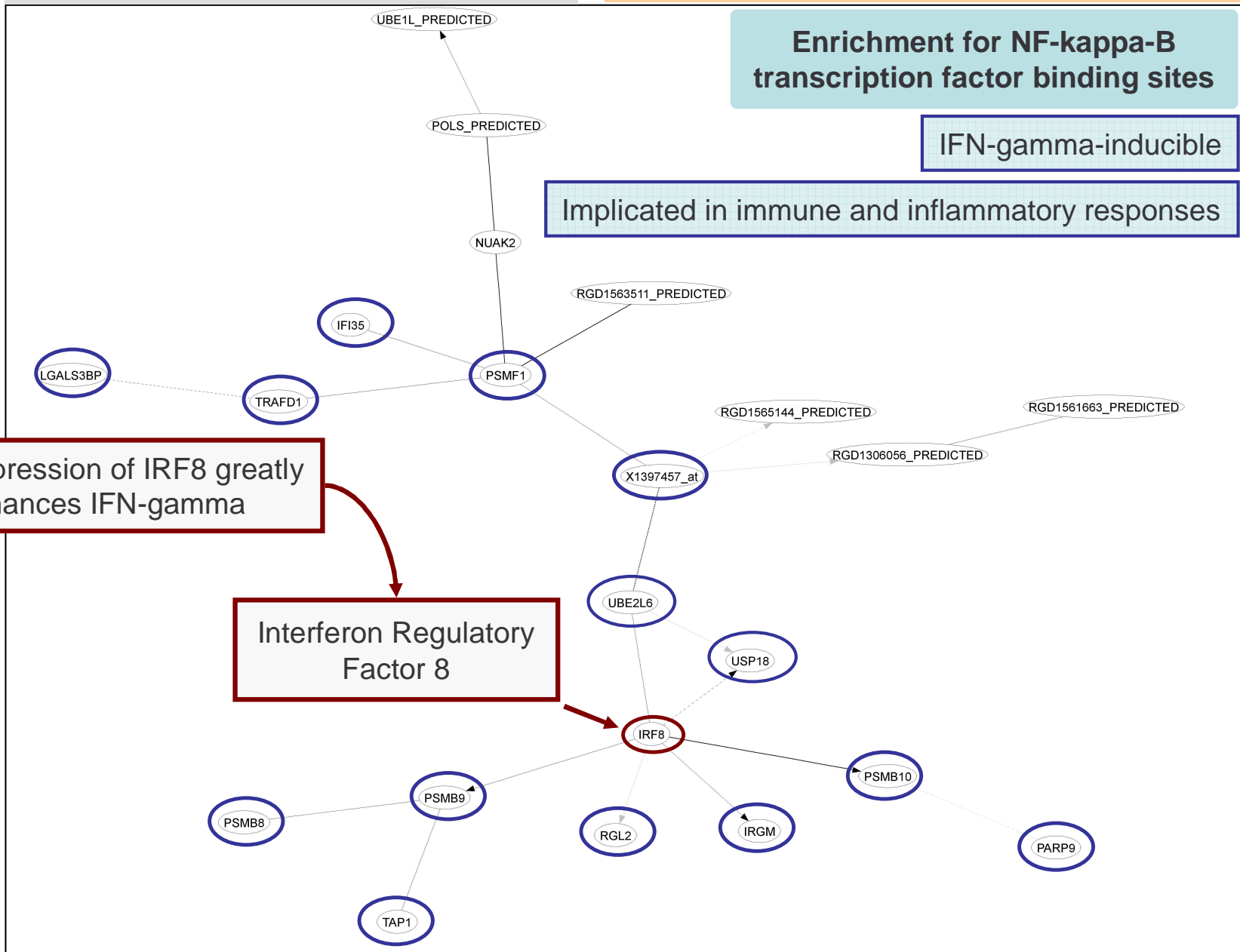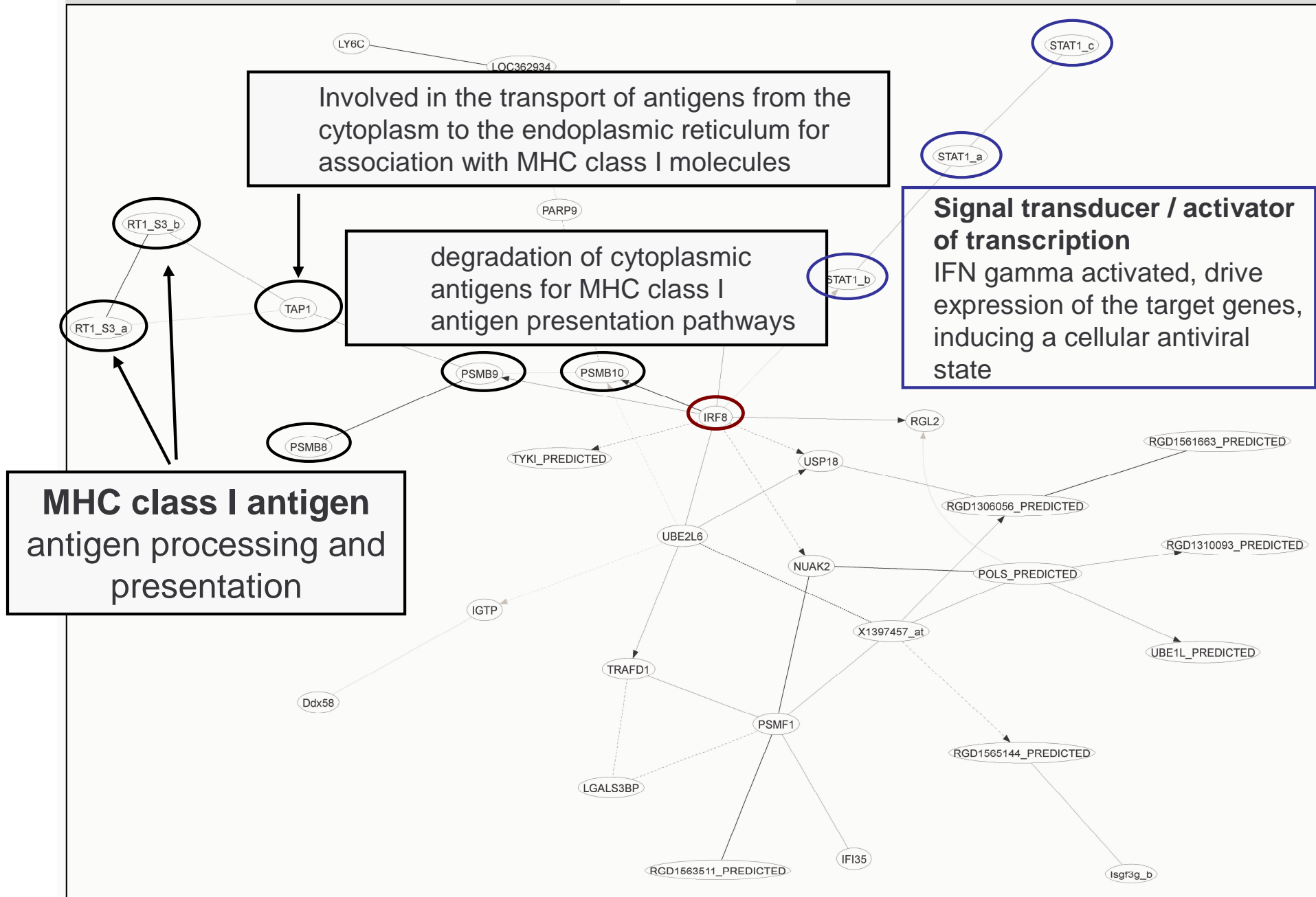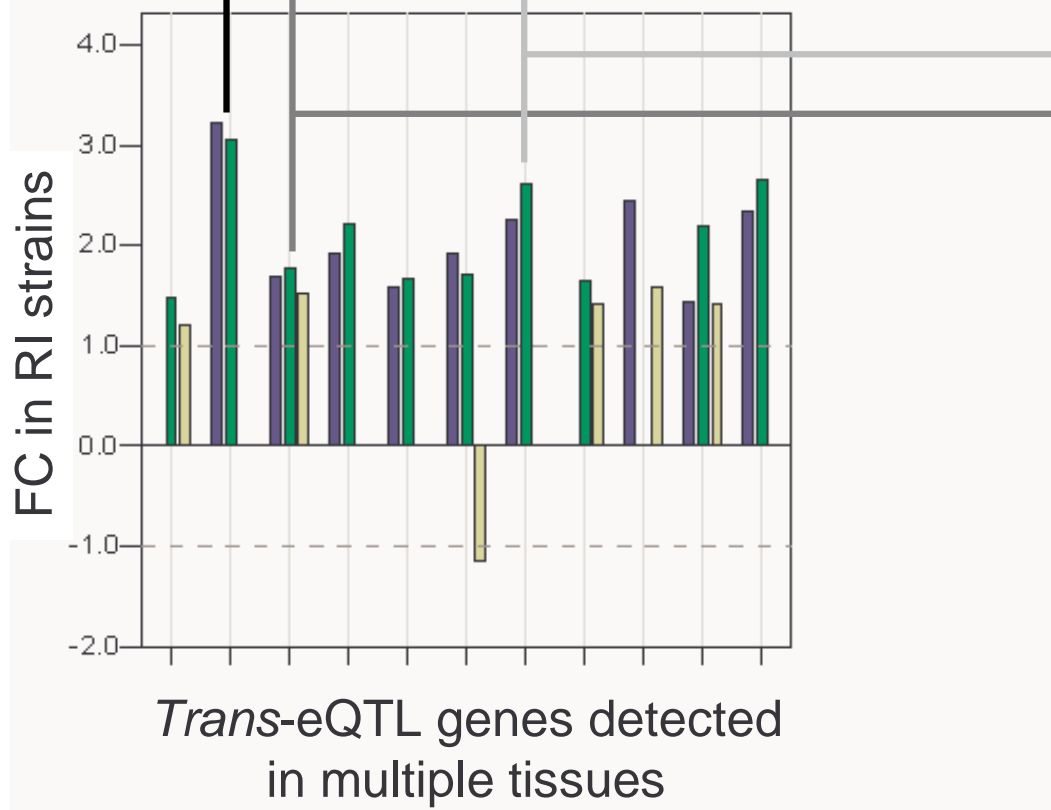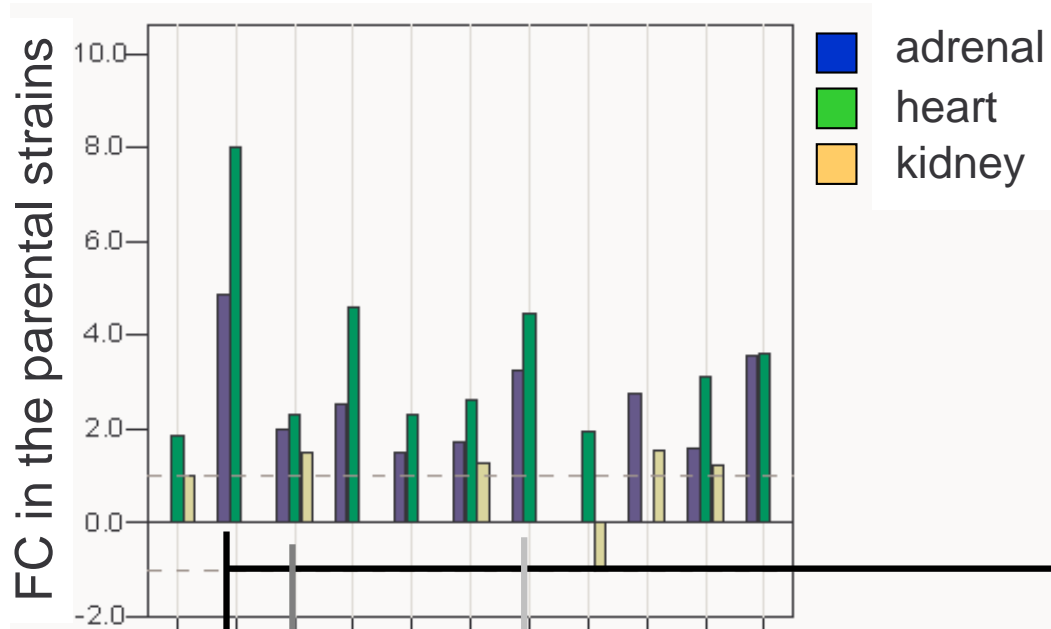
Is this association graph tissue specific?

kidney, all trans-eQTLs, posterior probability 0.95
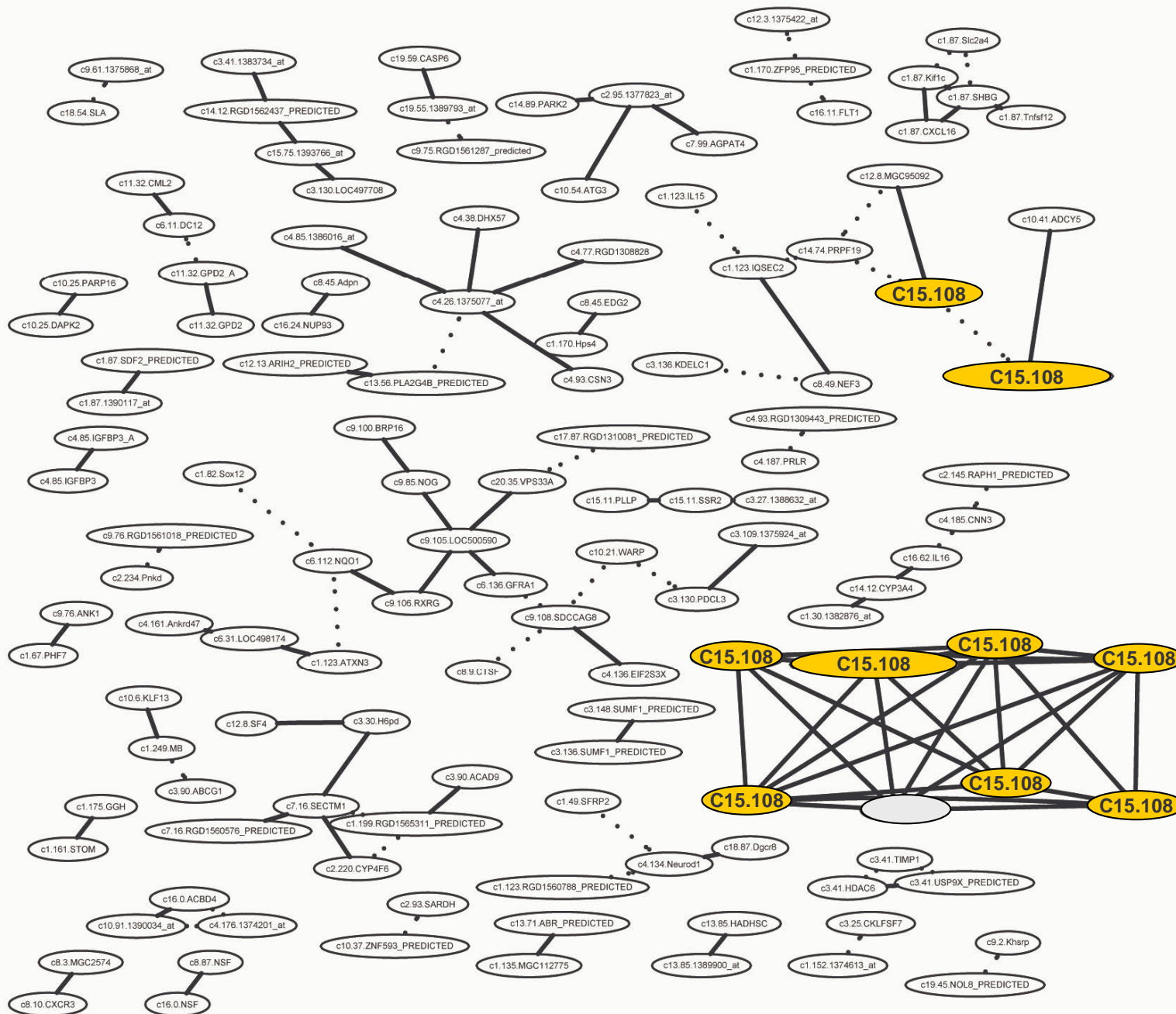
Adrenal, all trans-eQTLs, posterior probability 0.95

Microarray data: dysregulated genes

Imperial College London

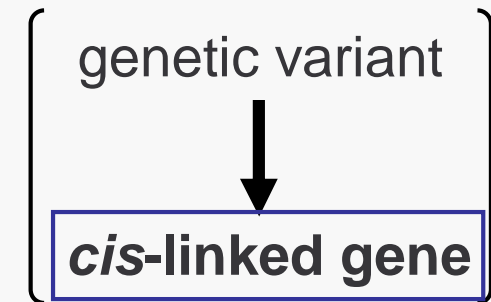**cis-acting eQTLs within the cluster region**

Transcripts representing *Dock9* gene
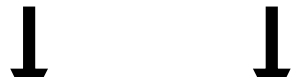
Model for master transcriptional regulator

genetic variant

↓

cis-linked gene

↓

Transcription Factor (TF) activity profile

↓

Expression of *trans*-linked genes

Gene Set Enrichment Analysis

Correlation between *Dock9* and __all__ *trans*-eQTLs (heart)

LCP2
IRF8
TAP1
PSMB9
PSMB8
PSMB10
IRGM
IFIT3
STAT1
USP18
IFI35
IRF7
LGALS3BP

Transcript 1370905_at
Enrichment Score -0.73
Normalized Enrichment Score -0.93
p-value 0.004
FDR q-value 3%

IRF8
PSMB8
PSMB10
TAP1
PSMB9
IRGM
STAT1
USP18
IFIT3
IFI35
IRF7
LGALS3BP

Transcript 1385378_at
Enrichment Score -0.69
Normalized Enrichment Score -1.85
Nominal p-value 0.015
FDR q-value 7%

Functional gene-sets correlated with *Dock9*

Genes whose expression is altered greater than twofold in **mouse** livers experiencing graft-versus-host disease (GVHD) as a result of allogenic bone marrow transplantation…
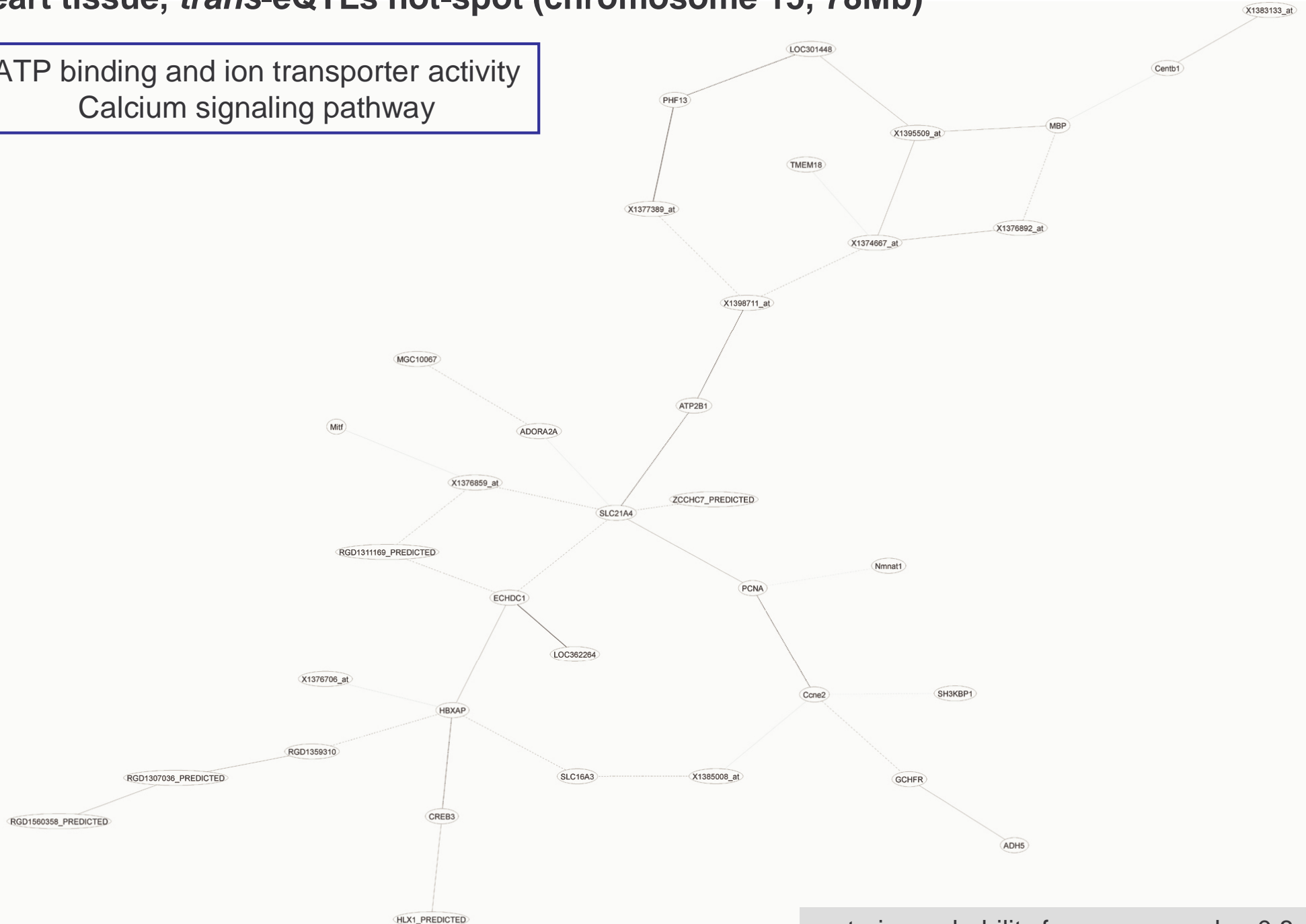
Imperial College London

Other examples

Heart tissue, *trans*-eQTLs hot-spot (chromosome 15, 78Mb)

ATP binding and ion transporter activity
Calcium signaling pathway

posterior probability for non-zero edge 0.8
posterior probability for directed edge 0.8

Fat *tissue specific*, *trans*-eQTLs hot-spot (chromosome 17)

posterior probability for non-zero edge 0.8
posterior probability for directed edge 0.8
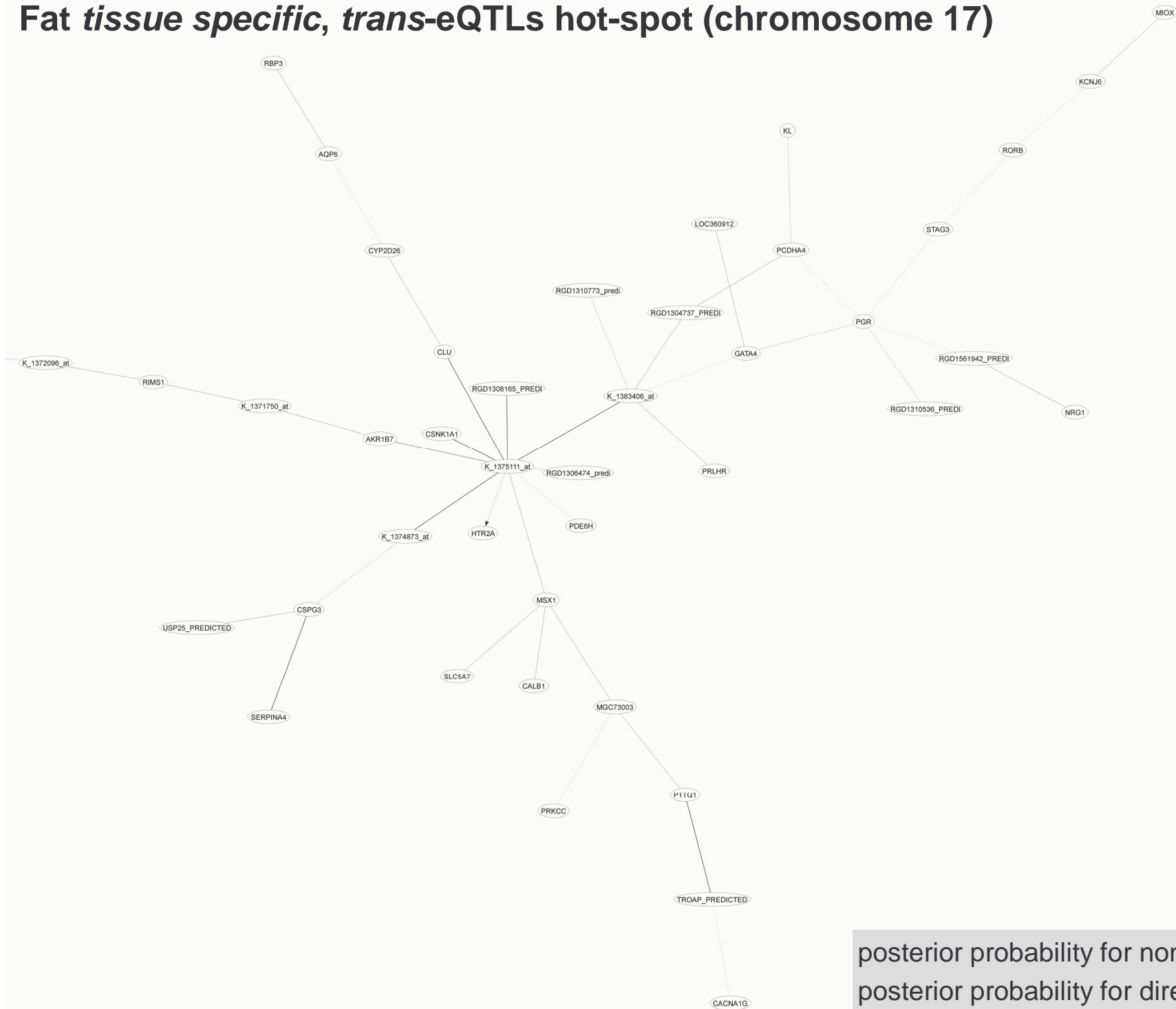
Summary

- Genome-wide eQTL data provide new insights into gene regulatory networks

- GGMs applied to *trans*-eQTL hotspots identified dysregulated pathway related to inflammation

- Hypothesis-driven inference can be a powerful approach to dissect regulatory networks

Acknowledgments

Sylvia Richardson

Tim Aitman

Stuart Cook

Jonathan Mangion

Rizwan Sarwar

Extra slides

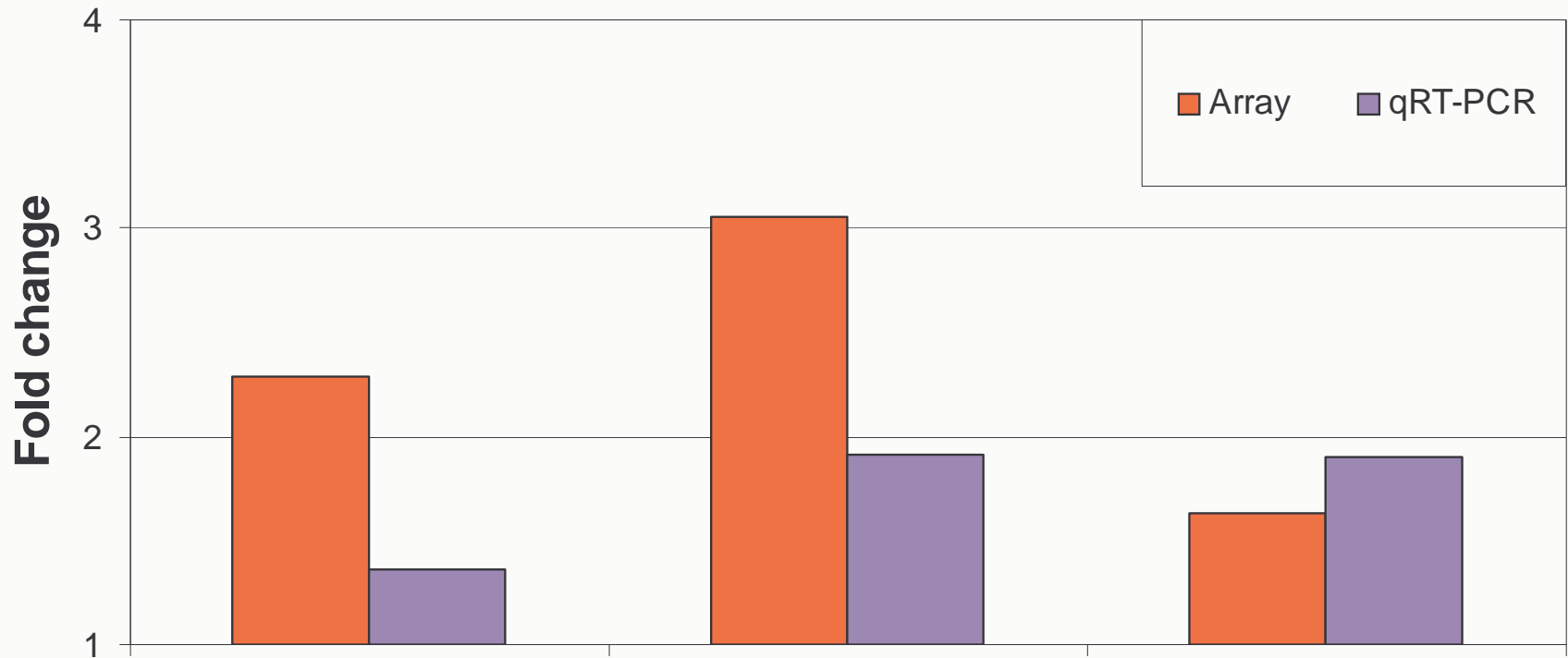# Chr 15 qRT-PCR validation in RI strains

Gene	Rarresin1_pred	Irf7_pred	Stat1
Array **P**	2.28 4.0E-05	3.06 8.6E-05	1.63 1.4E-04
qRT-PCR **P**	1.36 0.039	1.91 0.004	1.90 0.036

Rpt4 and *Irf7* mRNA levels increase in response to interferon

- H9c2 cells (rat cardiac embryonic myoblast)
- Stimulated with recombinant rat interferon for 3 hours
- RNA extracted, assayed by qRT-PCR (SYBR Green I)
- 3 independent expts, 3 biological replicates