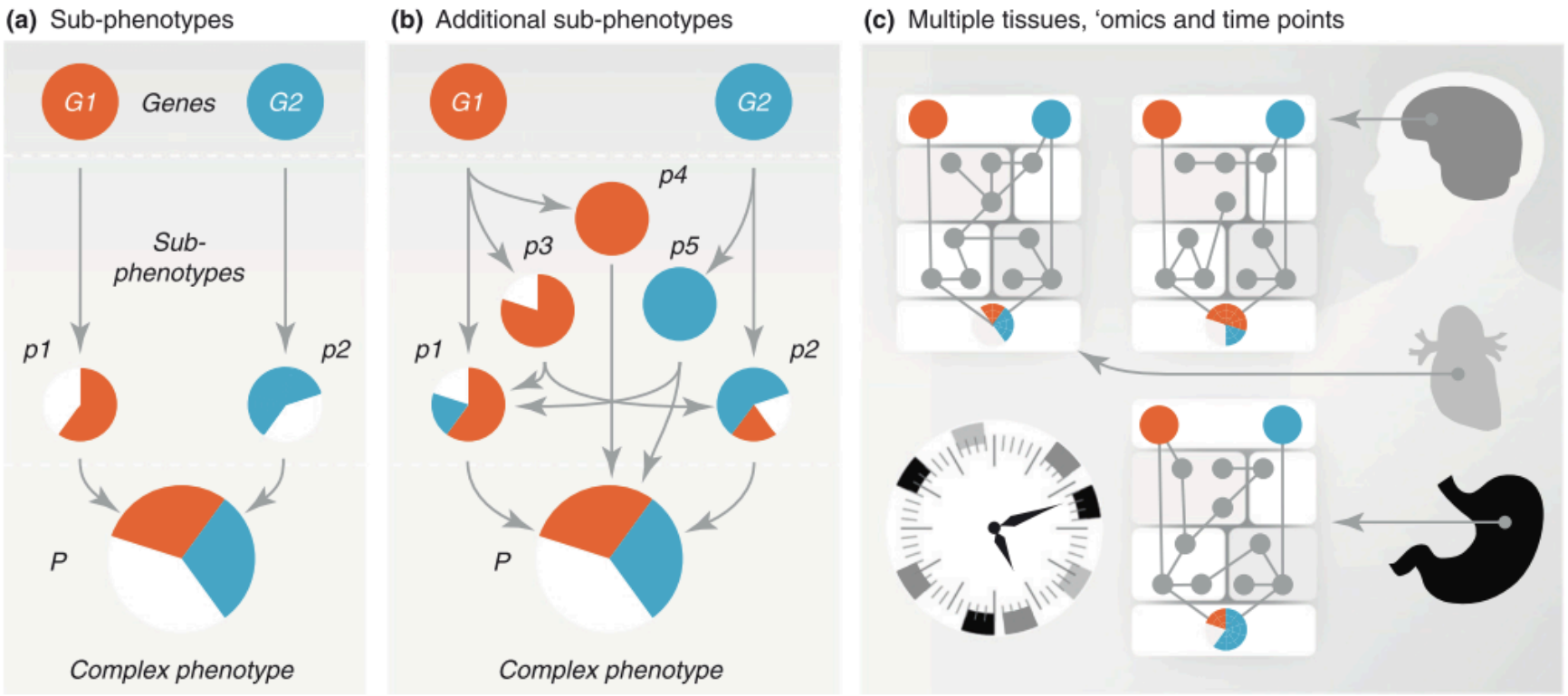
Integrative omics

Michael Inouye
Centre for Systems Genomics
University of Melbourne

Summer Institute in Statistical Genetics 2017
Network & Pathway Analysis of Omics Data
Brisbane

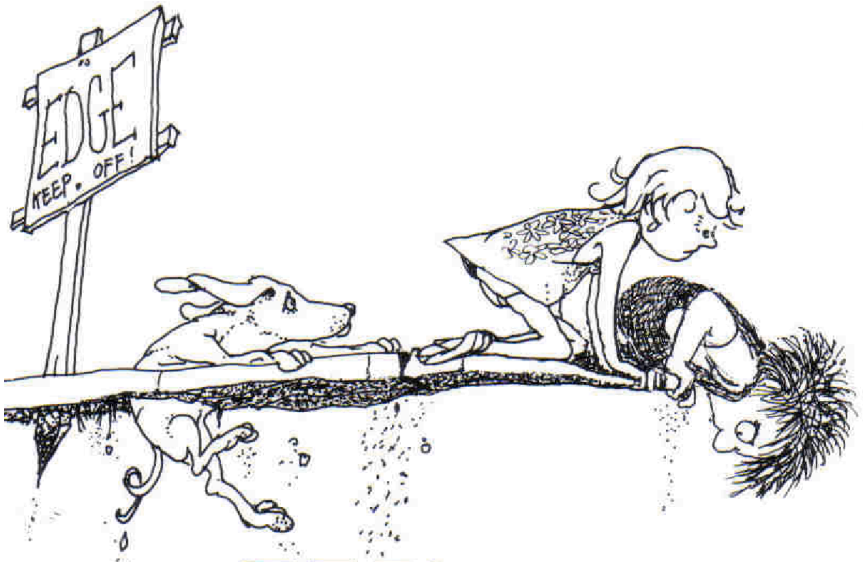
@minouye271
inouyelab.org

Background



Inouye, *Trends Genetics* 2011

Where the sidewalk ends…



Integrative analysis is a relatively undeveloped area

Lots of scope for development and novel ideas

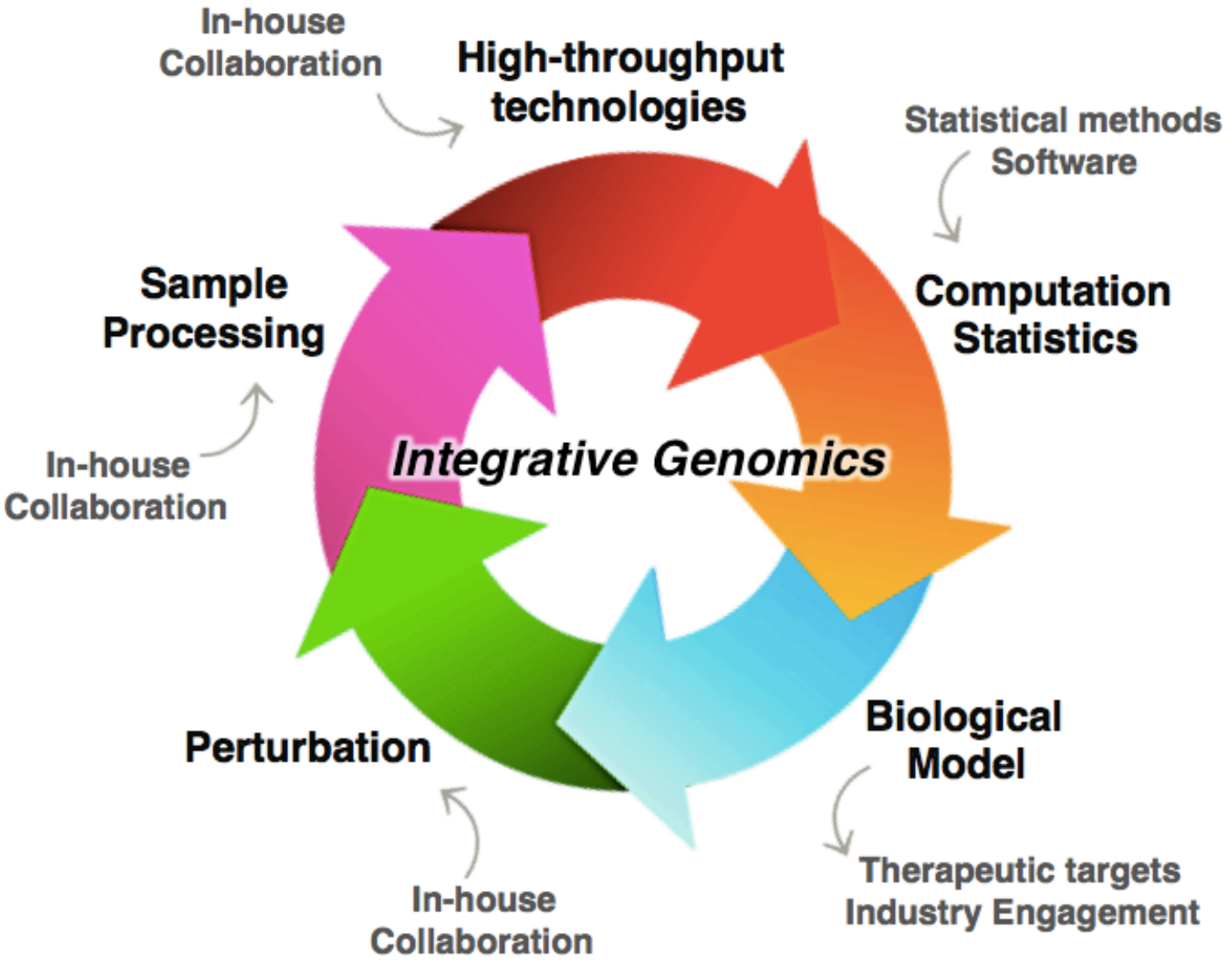
Nothing close to consensus on analytical approaches and strategies

Why integrate?

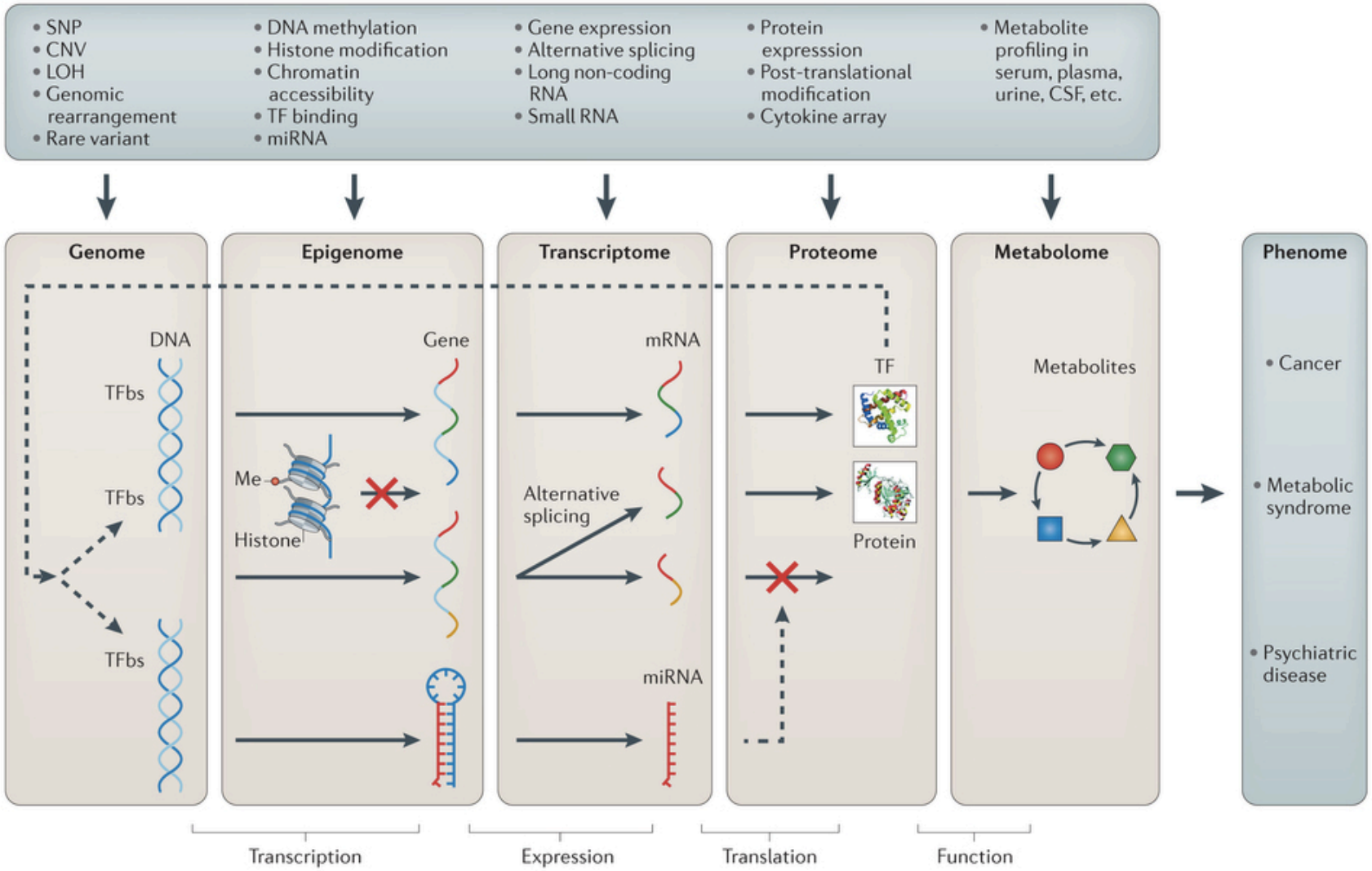
- It's likely that variation within a single omic data type (e.g. genome) will not capture the complexity of the phenotype

- It may not explain all of phenotypic variance nor identify all the causal factors

- Integration may better explain phenotype and identify/characterise (multiple) pathways and intervention points to control phenotype


In-house
Collaboration
High-throughput
technologies
Statistical methods
Software
Computation
Statistics
Sample
Processing
In-house
Collaboration
Integrative Genomics
Perturbation
In-house
Collaboration
Biological
Model
Therapeutic targets
Industry Engagement

Biological framework for multi-omics



Ritchie MD et al, *Nat Rev Genet* 2015

Challenges

- **Large P: High dimensionality**
  - 10K, 100K, 100M variables per sample
- **Small N**
- **Heterogeneous data**
  - Different molecules
  - Different technologies
  - Different sampling strategies
- **Correlation**
- **Computational efficiency/feasibility**

Main things to be aware of

- **Understand the biological models underlying the data**
  - Context and interpretation
- **Know the technology**
  - Batches, biases, error profiles, sensitivities/specificities, missing data
- **Know the sampling strategy(s)**
  - Group-wise (case/control), population-based, enrichments, stimuli?
- **Spend time exploring the data**
  - Without exception, you will see things that require follow up
- **Build analysis pipelines and log all analyses**
- **The data may be complex but your analysis and presentation doesn't have to be**

Role of transcriptome in integrative analyses

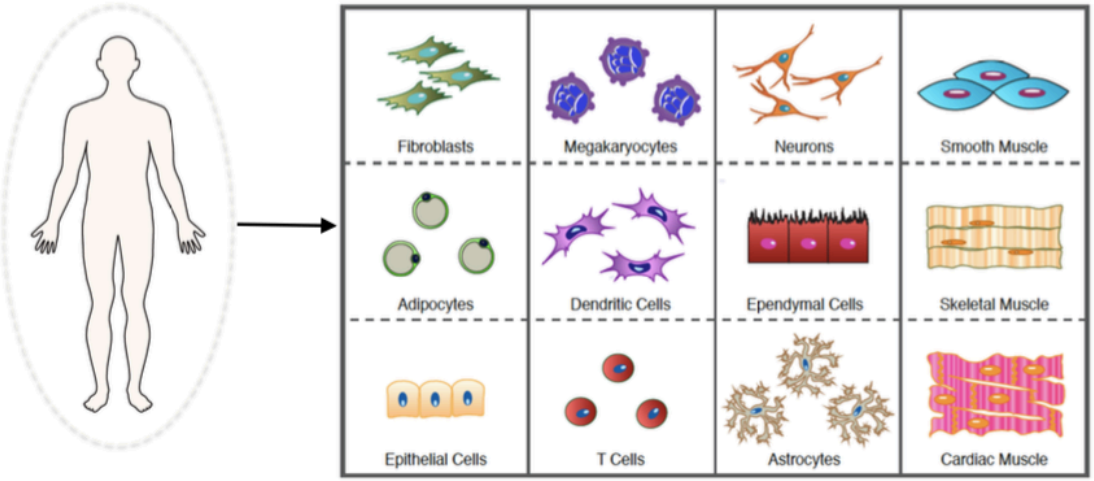
- Insights into biomolecular networks
- Less technical variability than proteomics
- Relatively affordable
- Stable tissues and cell types are (usually) readily available
- Many network methods have been applied to gene expression data in the past
- Gene expression is thus a convenient way to characterise the average biological state of the cell population(s) being assessed

A Google Maps for the Human Body

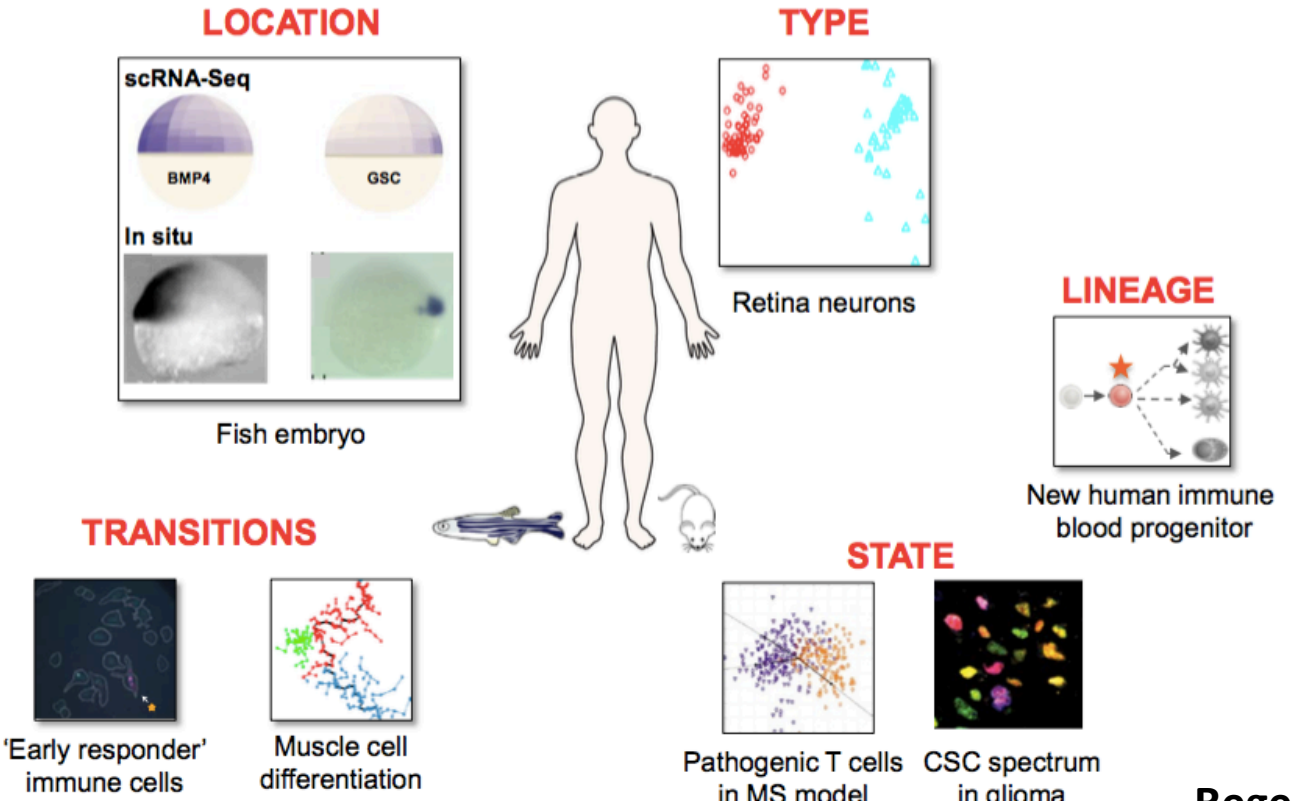
A group of scientists has taken the first important steps towards creating the Human Cell Atlas—a complete inventory of our staggeringly diverse cells.

ED YONG | OCT 14, 2016 | SCIENCE

www.humancellatlas.org



Human adult $2x10^{13}$ cells

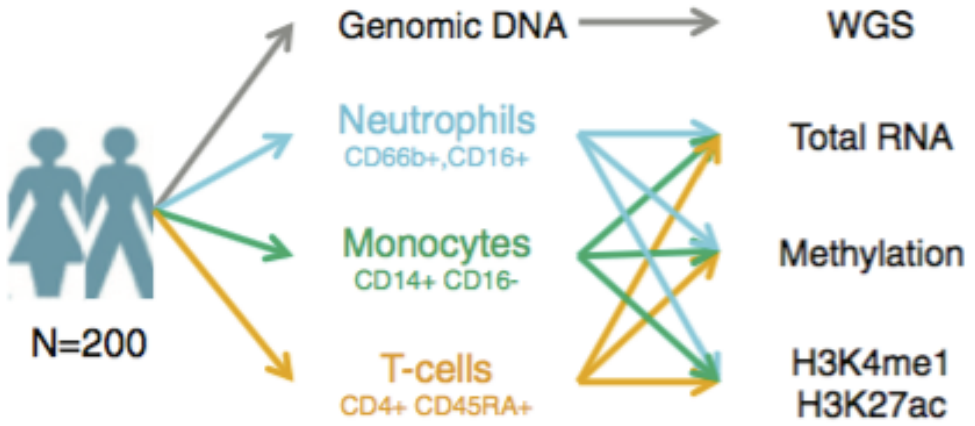


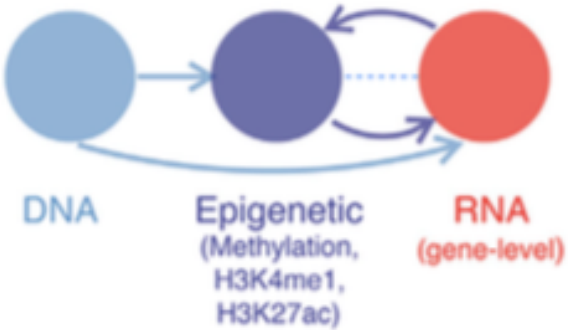
Regev & Teichmann

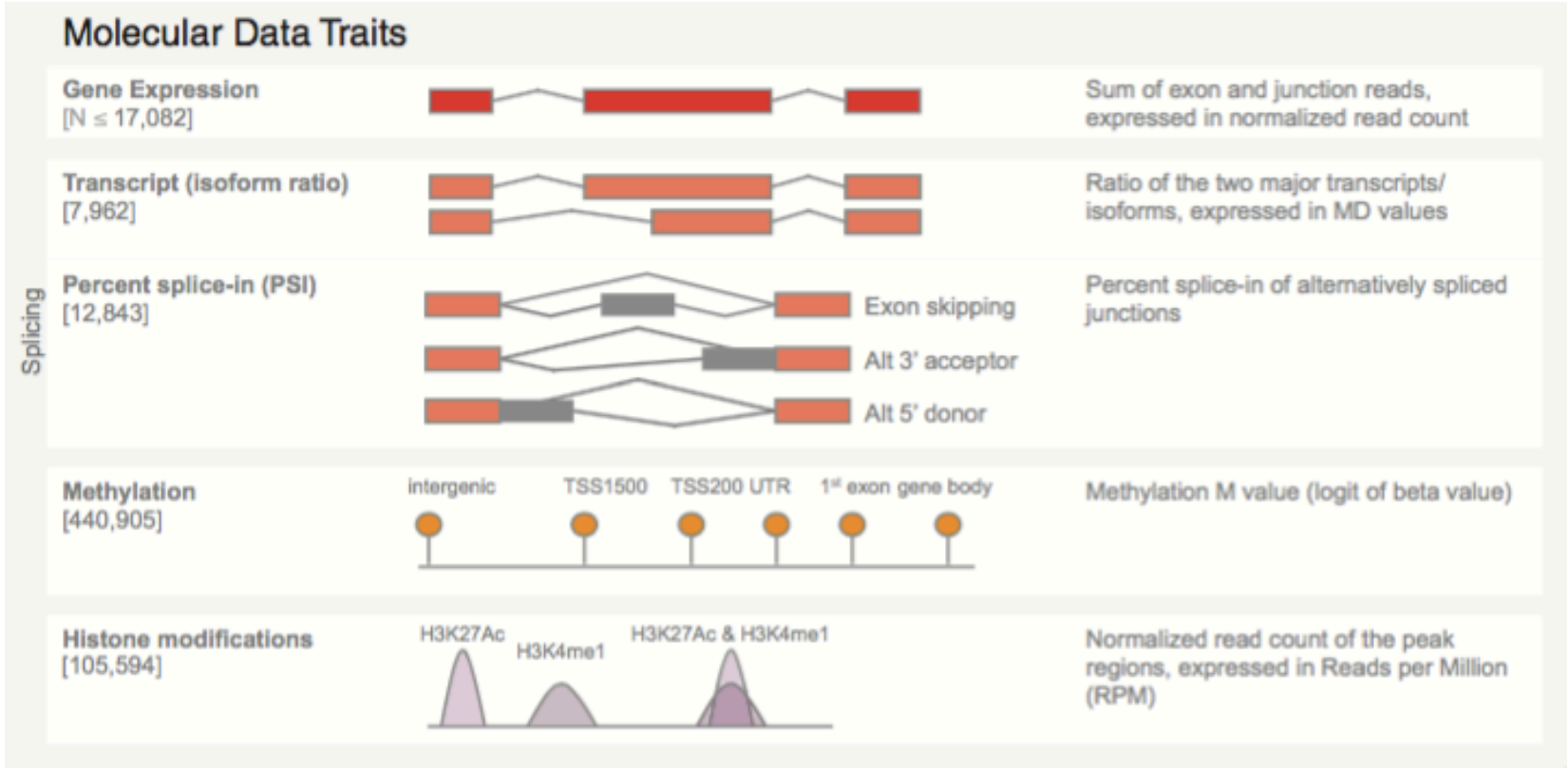
Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells

Lu Chen,[1,2,28] Bing Ge,[3,28] Francesco Paolo Casale,[4,28] Louella Vasquez,[1,28] Tony Kwan,[3] Diego Garrido-Martín,[5,6] Stephen Watt,[1] Ying Yan,[1] Kousik Kundu,[1,2] Simone Ecker,[7,8] Avik Datta,[9] David Richardson,[9] Frances Burden,[2,18] Daniel Mead,[1] Alice L. Mann,[1] Jose Maria Fernandez,[7] Sophia Rowlston,[2,18] Steven P. Wilder,[10] Samantha Farrow,[2,18] Xiaojian Shao,[3] John J. Lambourne,[3,2,18] Adriana Redensek,[3] Cornelis A. Albers,[13,16] Vyacheslav Amstislavskiy,[14] Sofie Ashford,[2,18] Kim Berentsen,[15] Lorenzo Bomba,[1] Guillaume Bourque,[3] David Bujold,[3] Stephan Busche,[3] Maxime Caron,[3] Shu-Huang Chen,[3] Warren Cheung,[3] Oliver Delaneau,[12] Emmanouil T. Dermitzakis,[12] Heather Elding,[1] Irina Colgiu,[17] Frederik O. Bagger,[2,4,18] Paul Flicek,[9] Ehsan Habibi,[15] Valentina Iotchkova,[1,11] Eva Janssen-Megens,[15] Bowon Kim,[15] Hans Lehrach,[14] Ernesto Lowy,[9] Amit Mandoli,[15] Filomena Matarese,[15] Matthew T. Maurano,[19] John A. Morris,[3] Vera Pancaldi,[7] Farzin Pourfarzad,[20] Karola Rehnstrom,[2,18] Augusto Rendon,[2,21] Thomas Risch,[14] Nilofar Sharifi,[15] Marie-Michelle Simon,[3] Marc Sultan,[14] Alfonso Valencia,[7] Klaudia Walter,[1] Shuang-Yin Wang,[15] Mattia Frontini,[2,18,22] Stylianos E. Antonarakis,[12] Laura Clarke,[9] Marie-Laure Yaspo,[14] Stephan Beck,[8] Roderic Guigo,[5,6,23] Daniel Rico,[7,24] Joost H.A. Martens,[15] Willem H. Ouwehand,[1,2,18,22,25] Taco W. Kuijpers,[2,20,26] Dirk S. Paul,[8,27] Hendrik G. Stunnenberg,[15] Oliver Stegle,[4] Kate Downes,[2,18] Tomi Pastinen,[3,*] and Nicole Soranzo[1,2,22,25,29,*]

Cell 167, 1398–1414 (2016)







Variance Decomposition

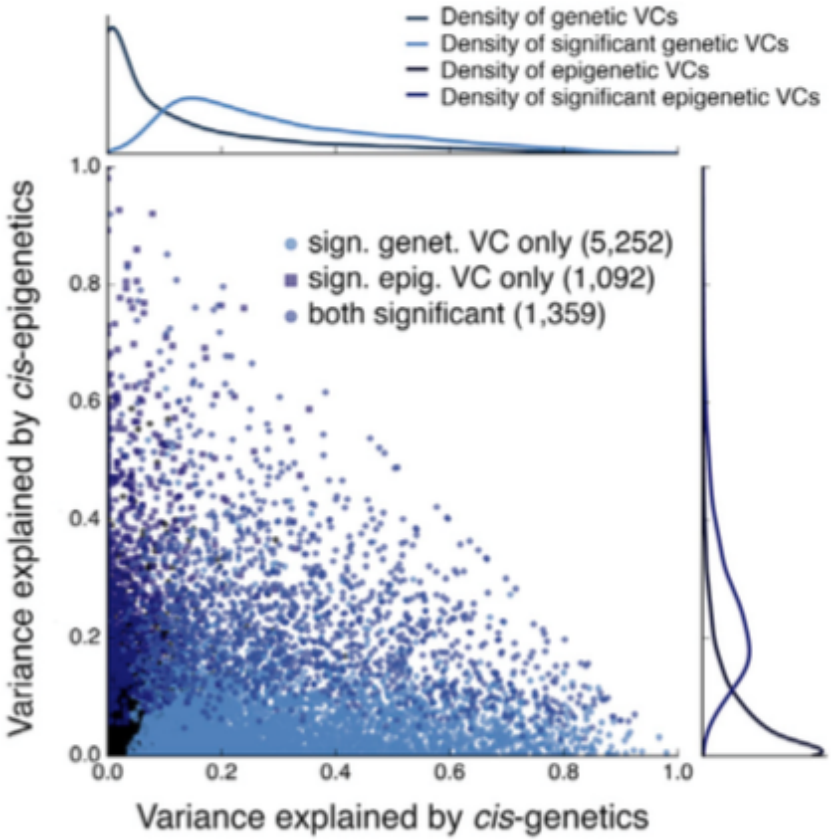
Quantitative Trait Loci

Alleles-specific Analysis

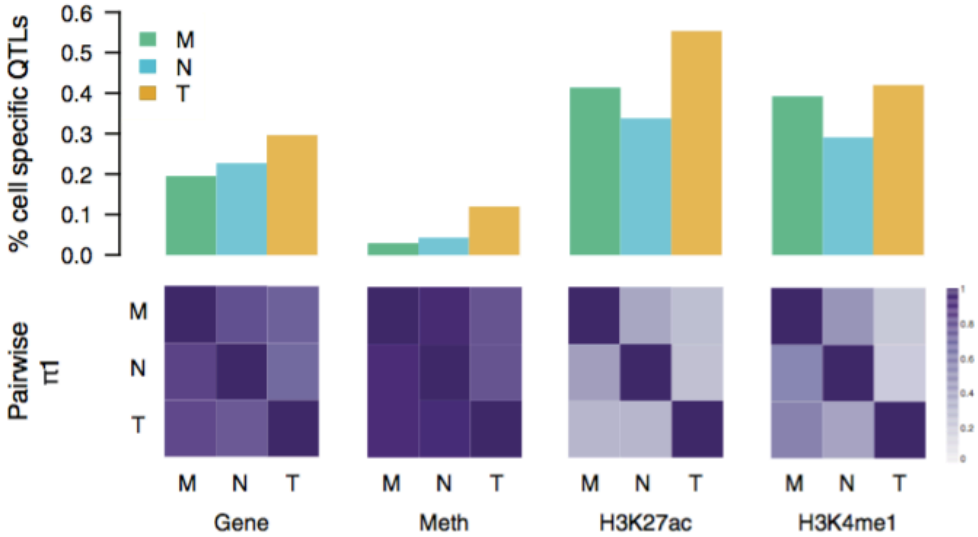
Disease Integration

Chen et al, *Cell* 2016

Is a gene's (monocyte) transcription dominated by genetic or epigenetic effects?

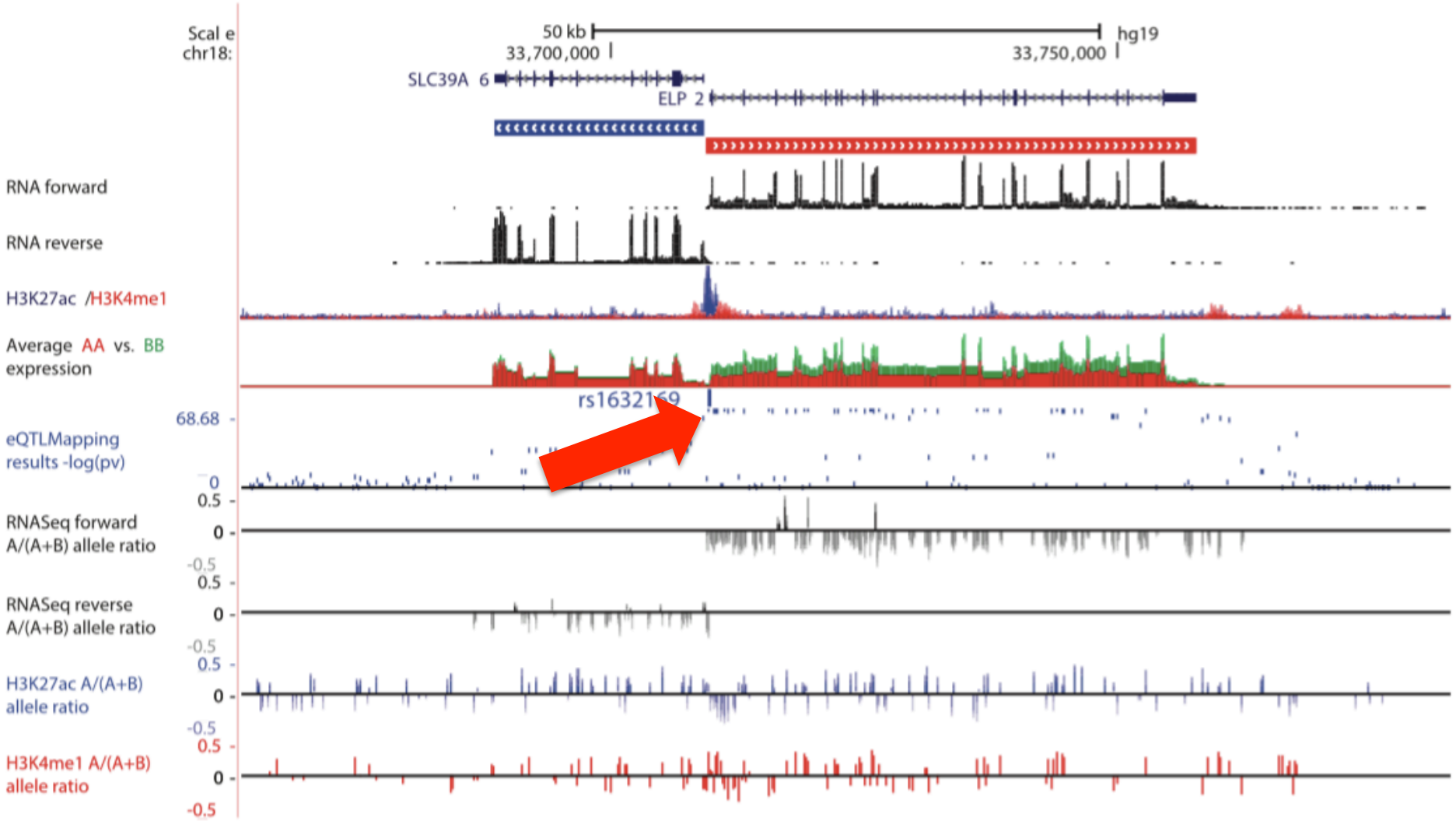


Shared QTLs between cell types



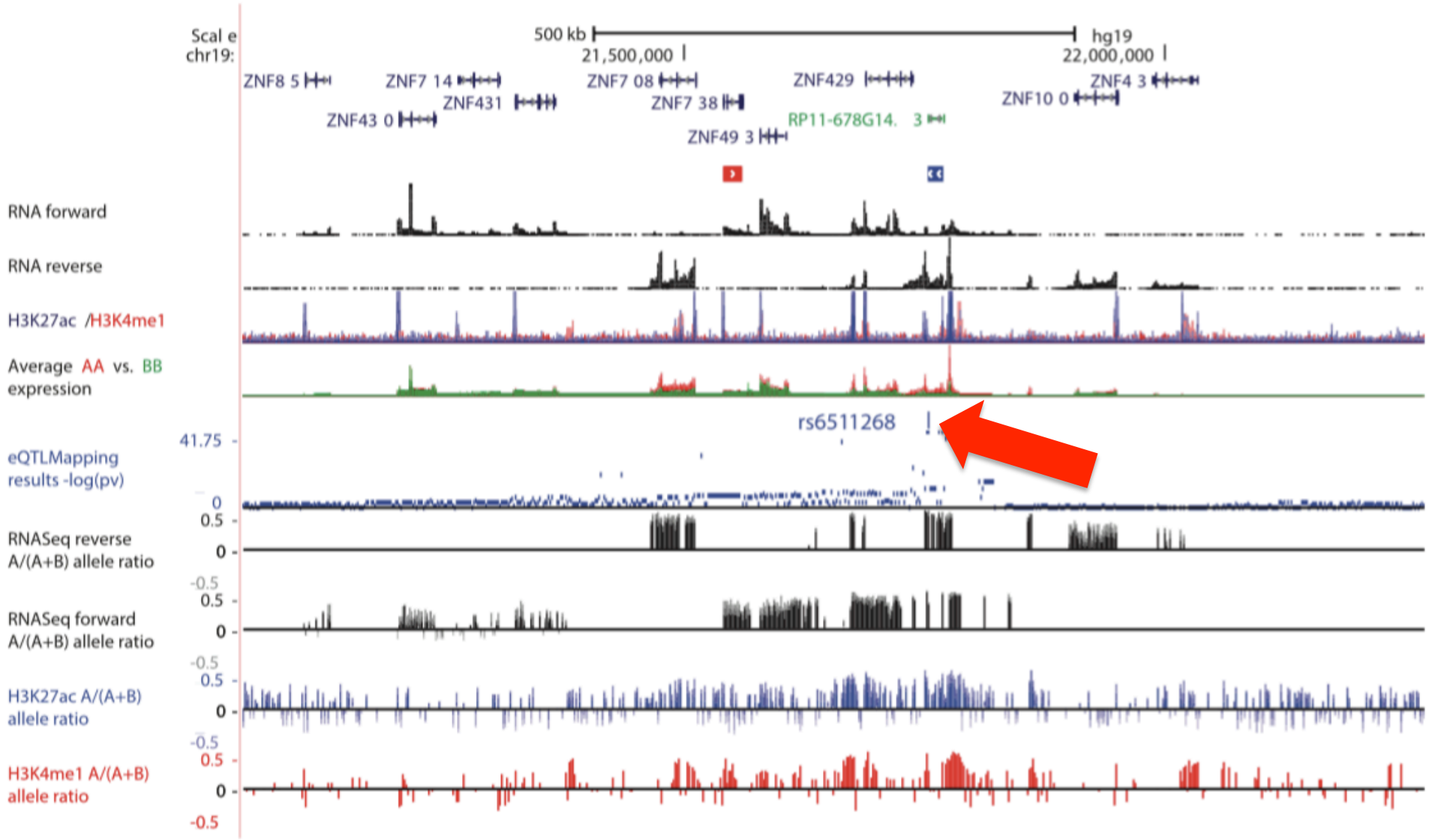
Chen et al, *Cell* 2016

eSNP effects at a bidirectional promoter for *SLC39A* and *ELP*



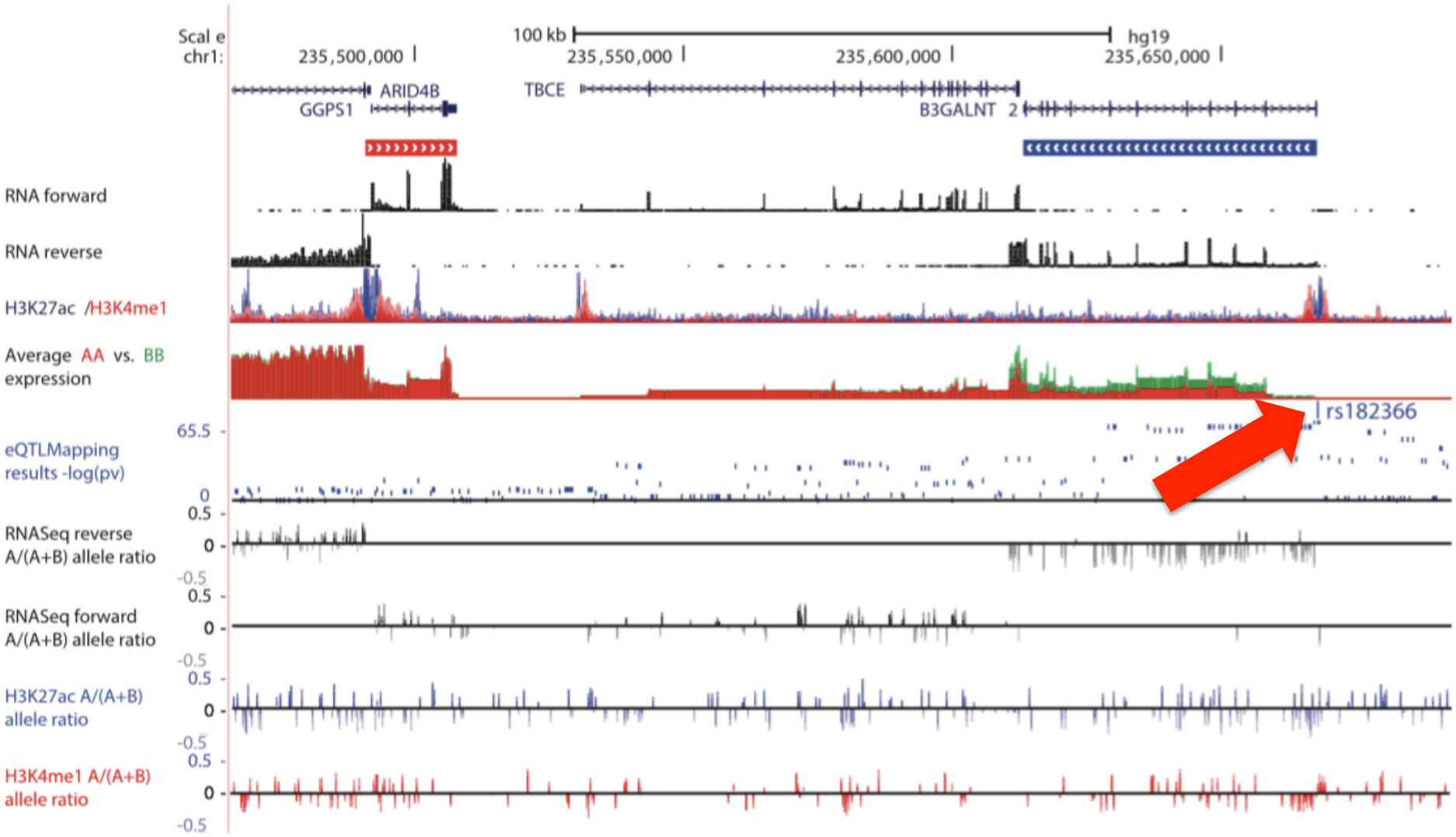
Chen et al, *Cell* 2016

eSNP effects on chromatin and forward/reverse strand expression



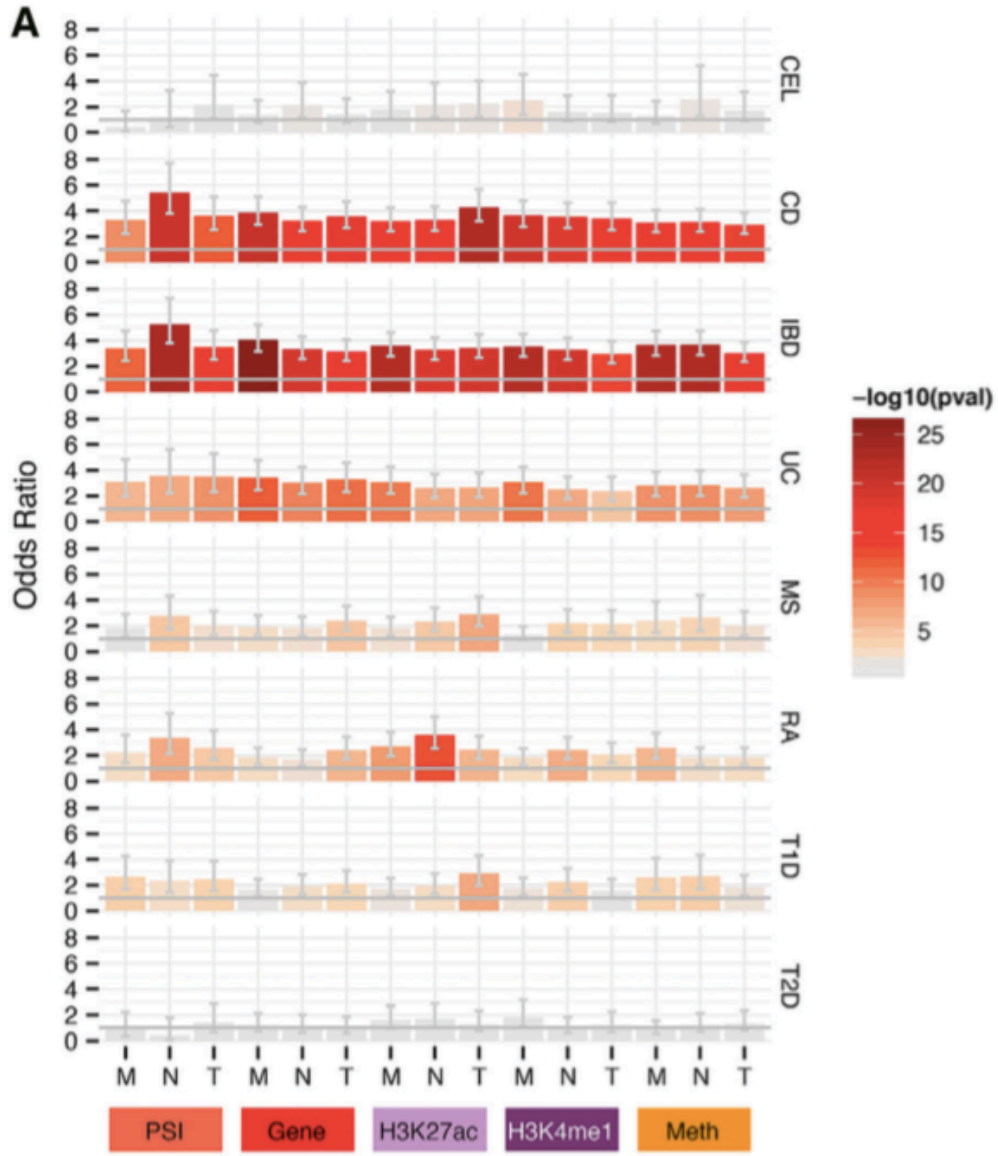
Chen et al, *Cell* 2016

eSNP effects *B3GALNT* & *ARID4B* promoters but only *B3GALNT* expression



Chen et al, *Cell* 2016

Enrichment of cell type specific QTLs at autoimmune loci



Integrative analysis of genomic, transcriptomic & metabolomic variation

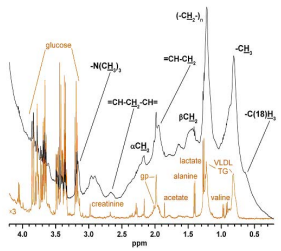
DILGOM: 590 randomly sampled individuals

Fasting whole blood

Genome

Transcriptome

Metabolome



+ clinical data

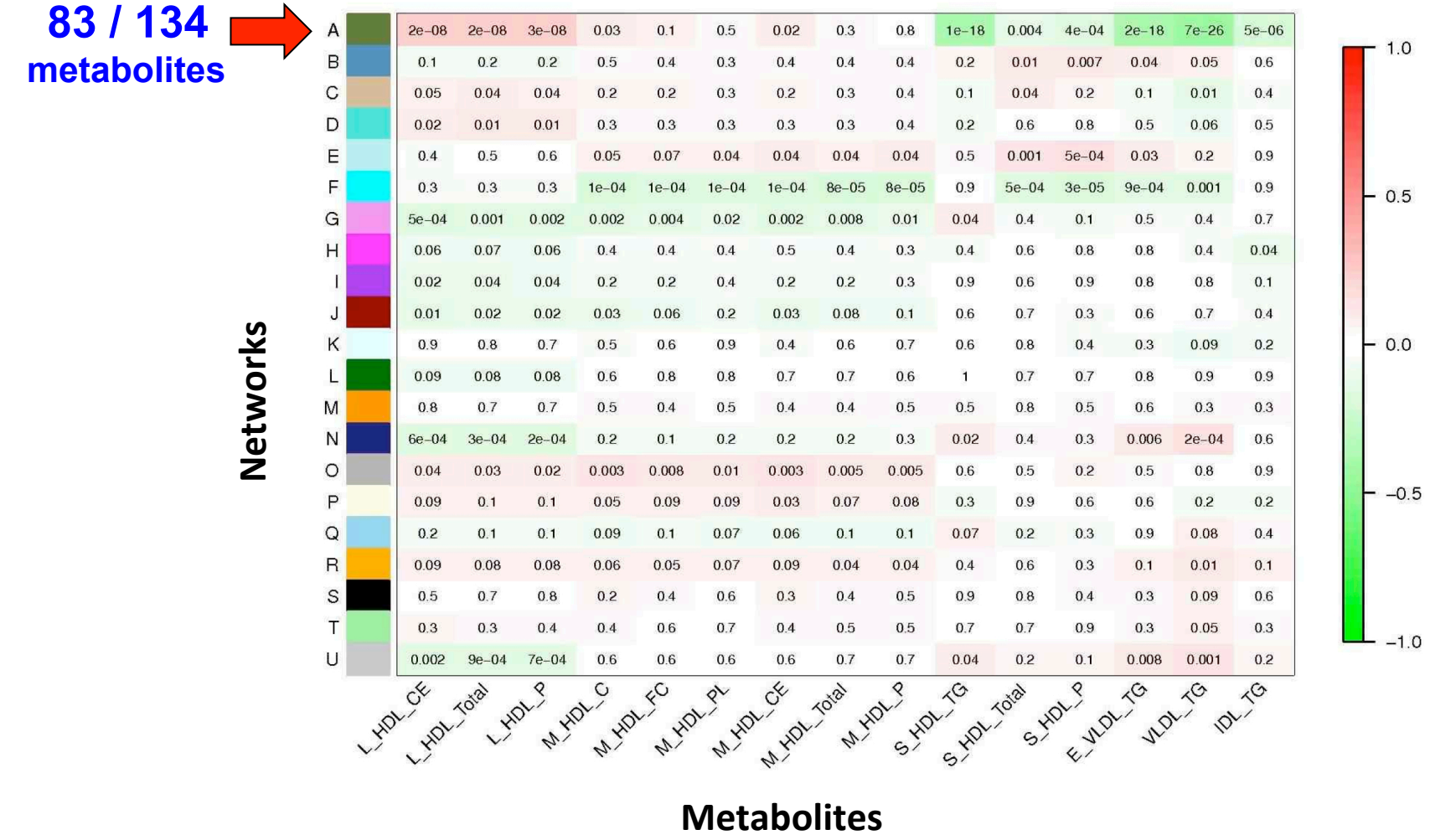
wellcome trust sanger institute

UNIVERSITY OF HELSINKI

Inouye *PLoS Genetics* 2010

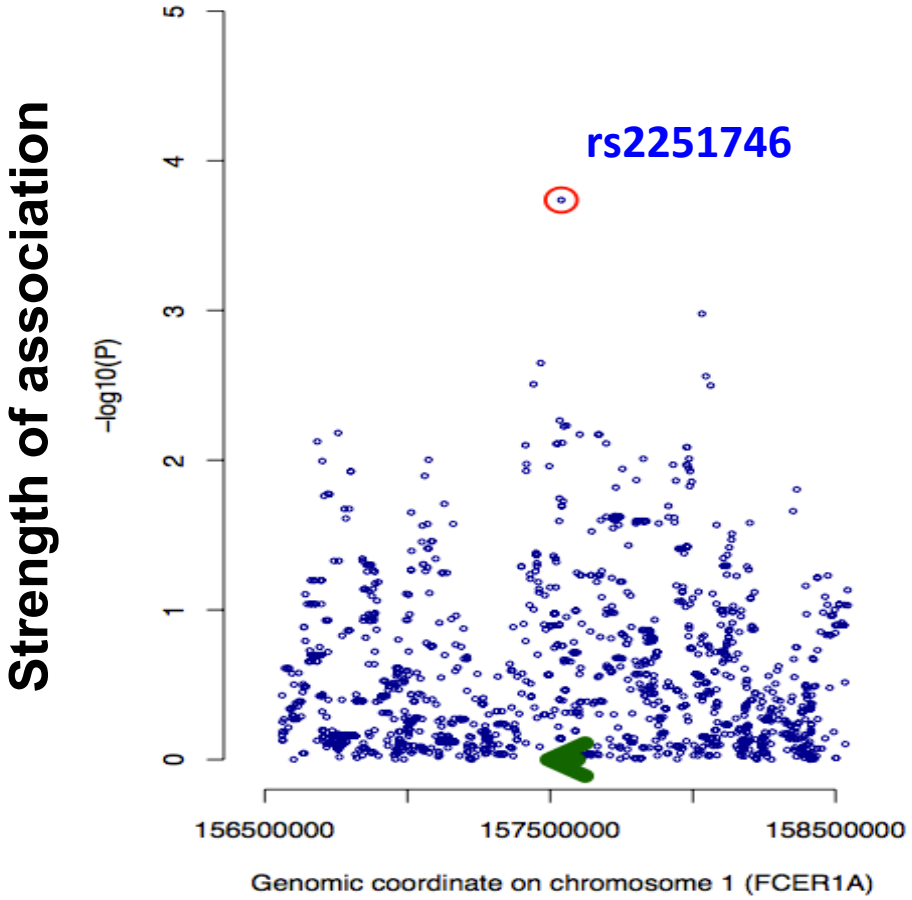
Relationships between gene networks and metabolome

83 / 134 metabolites



Inouye* & Kettunen* et al; *Molecular Systems Biology,* 2010

Does genetic variation influence LL module?



FCER1A $P = 1.83 \times 10^{-4}$
LL module $P = 4.28 \times 10^{-6}$

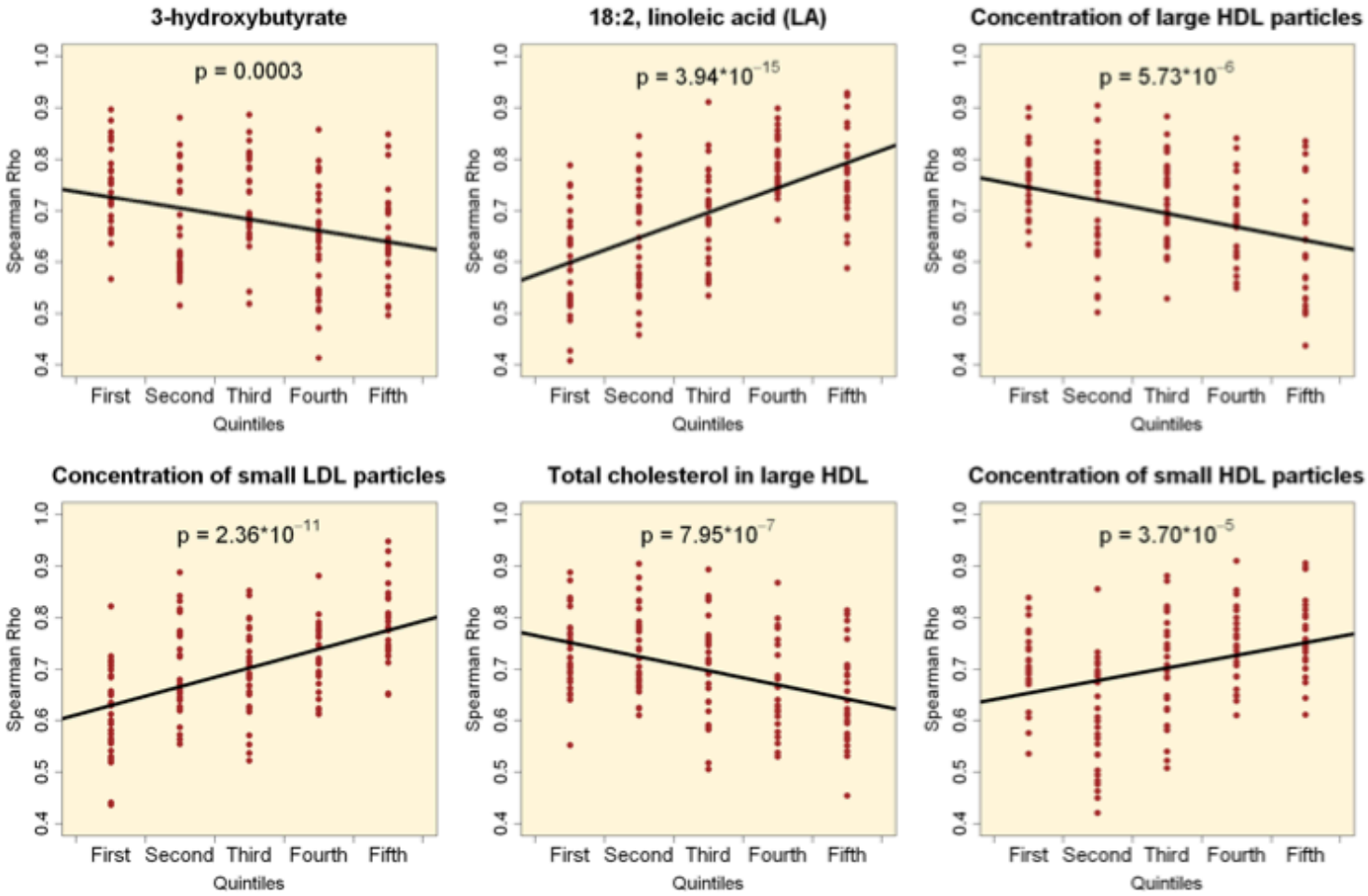
OPEN ACCESS Freely available online PLoS GENETICS

Genome-Wide Scan on Total Serum IgE Levels Identifies *FCER1A* as Novel Susceptibility Locus

Stephan Weidinger[1,2]*, Christian Gieger[3,4], Elke Rodriguez[2], Hansjörg Baurecht[2,5], Martin Mempel[1,2], Norman Klopp[3], Henning Gohlke[3], Stefan Wagenpfeil[5,6], Markus Ollert[1,2], Johannes Ring[1], Heidrun Behrendt[2], Joachim Heinrich[3], Natalija Novak[7], Thomas Bieber[7], Ursula Krämer[8], Dietrich Berdel[9], Andrea von Berg[9], Carl Peter Bauer[10], Olf Herbarth[11], Sibylle Koletzko[12], Holger Prokisch[13,14], Divya Mehta[13,14], Thomas Meitinger[13,14], Martin Depner[12], Erika von Mutius[12], Liming Liang[15], Miriam Moffatt[16], William Cookson[16], Michael Kabesch[12], H.-Erich Wichmann[3,4], Thomas Illig[3]

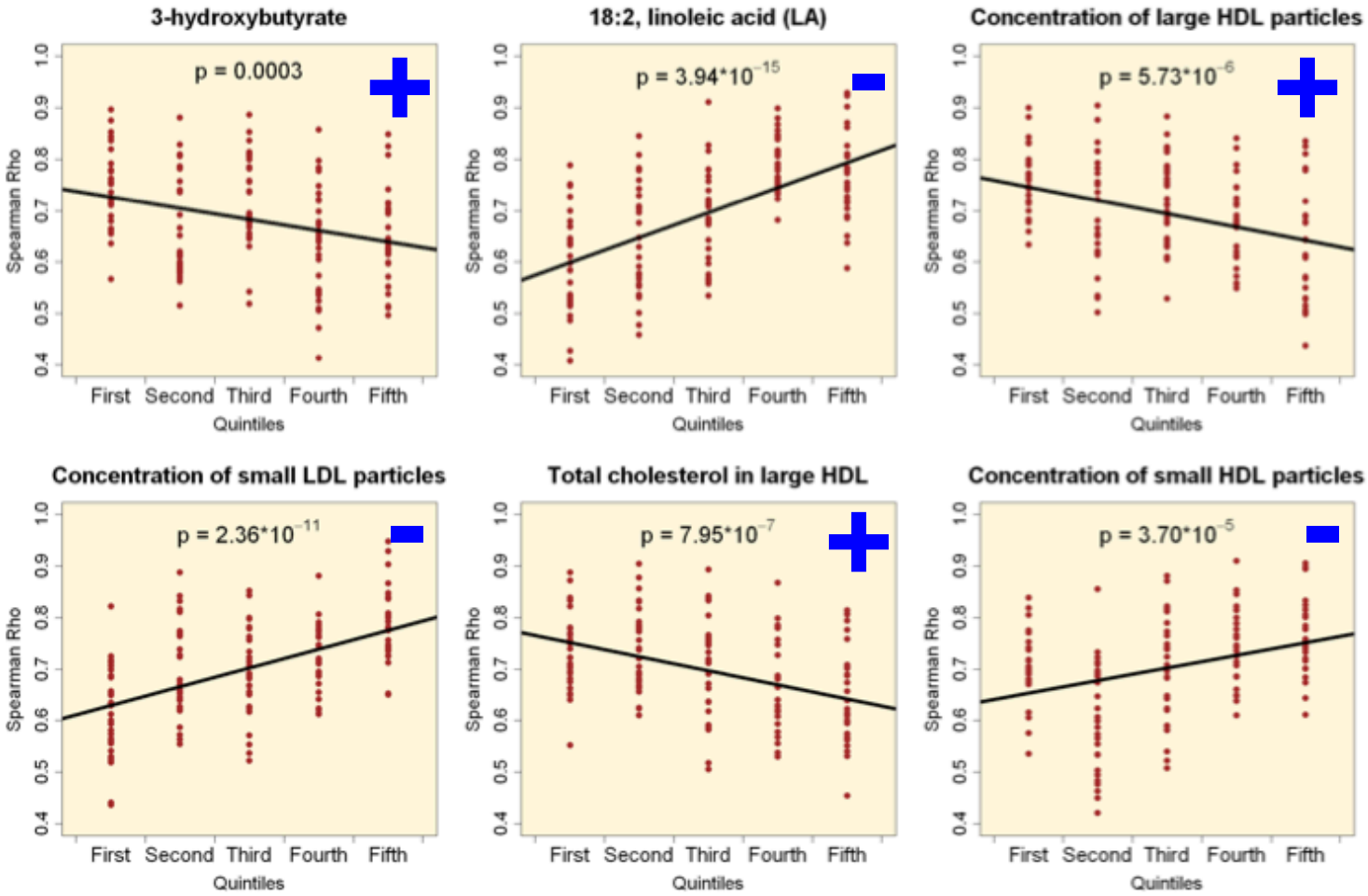
Inouye et al; *PLoS Genetics*, 2010

LL module appears reactive, do metabolites affect its connectivity?



Inouye* & Kettunen* et al; *Molecular Systems Biology,* 2010

Potential negative feedback loop

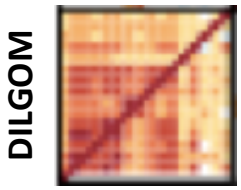


Inouye* & Kettunen* et al; *Molecular Systems Biology,* 2010

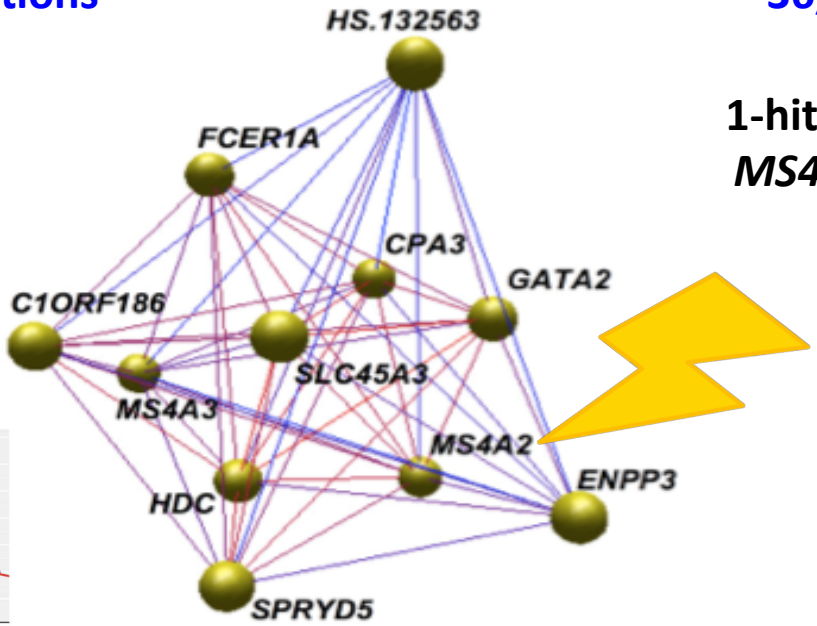
IgE signaling subnetwork at the transcriptome - metabolome interface

Independent replication of subnet and metabolite associations

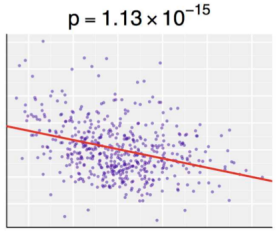
Gene-gene correlations



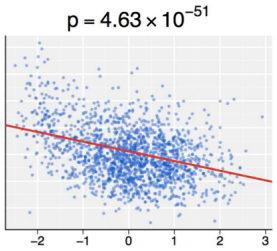
Young Finns Study



DILGOM
N=518

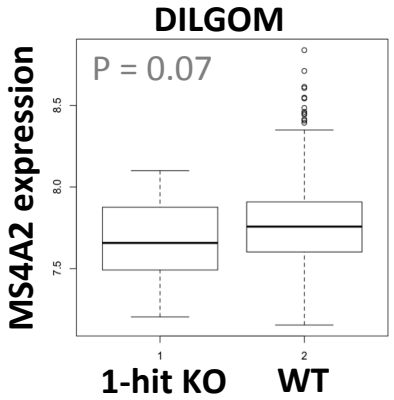


YFS
N=1,400



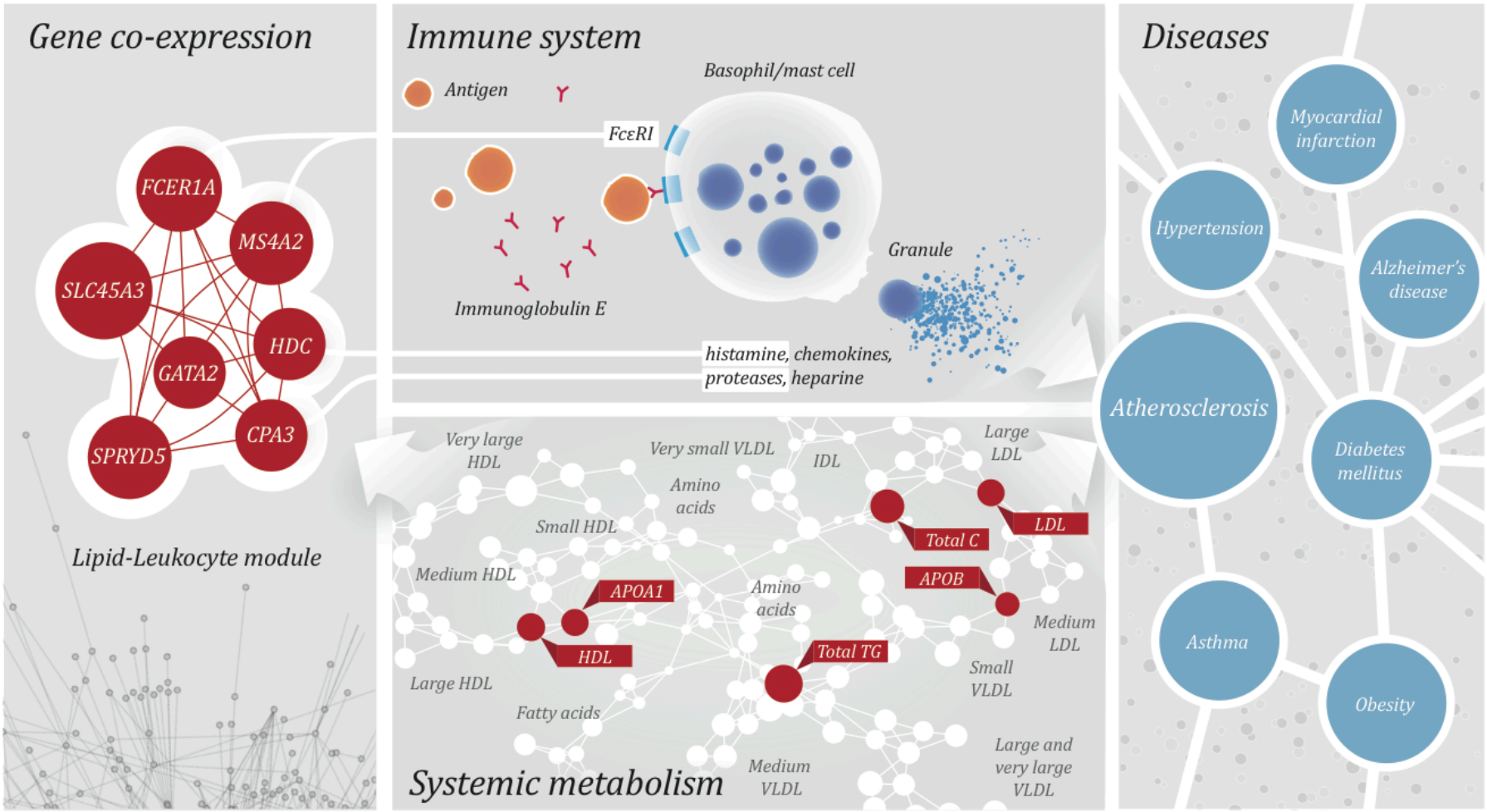
Natural human knockouts 36,000 exomes (SISu)

1-hit KO *MS4A2* → ↑ **Circulating triglycerides**

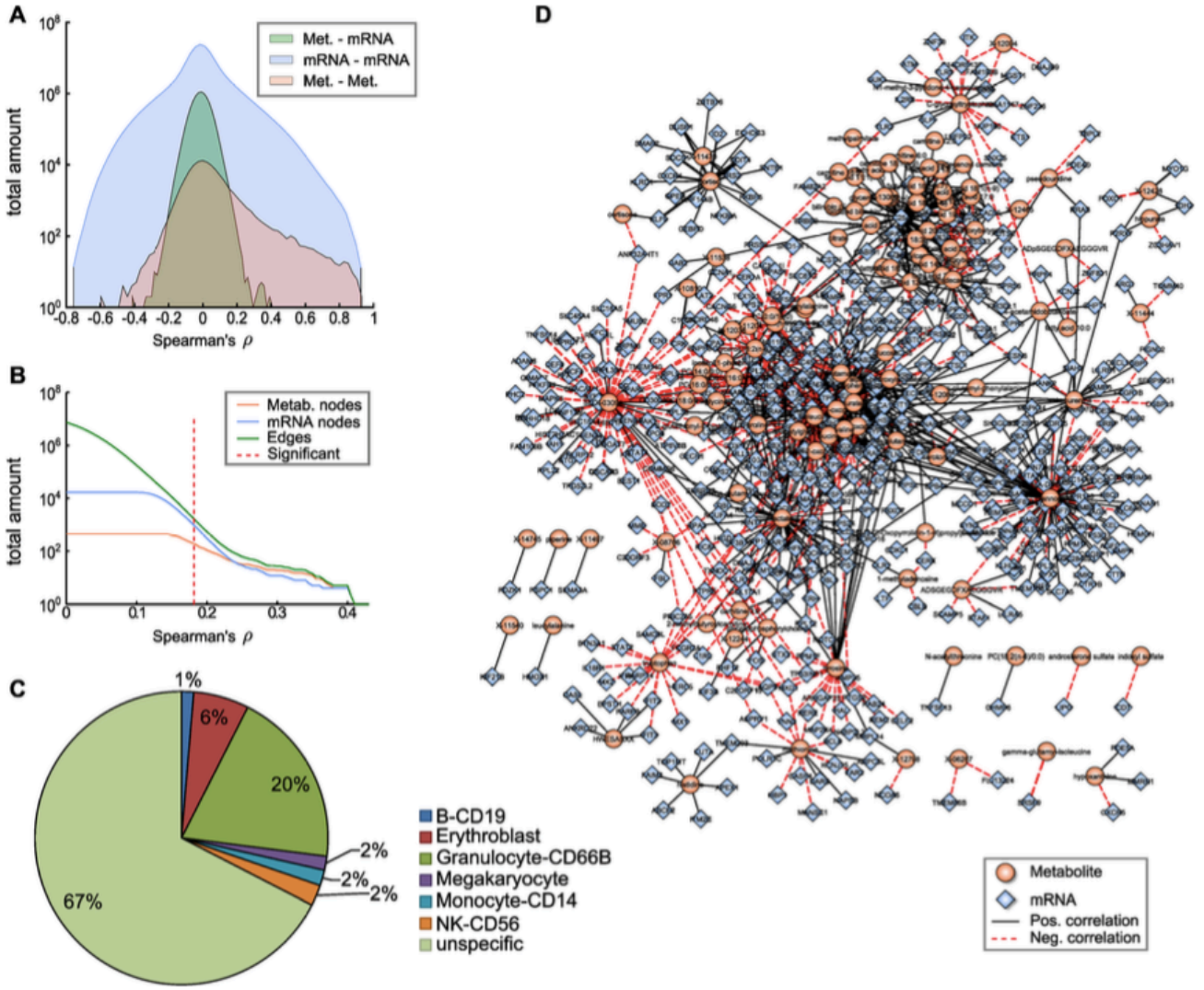


Lim, ..., Daly, Palotie *PLoS Genetics* 2014

Constructing a working biological model

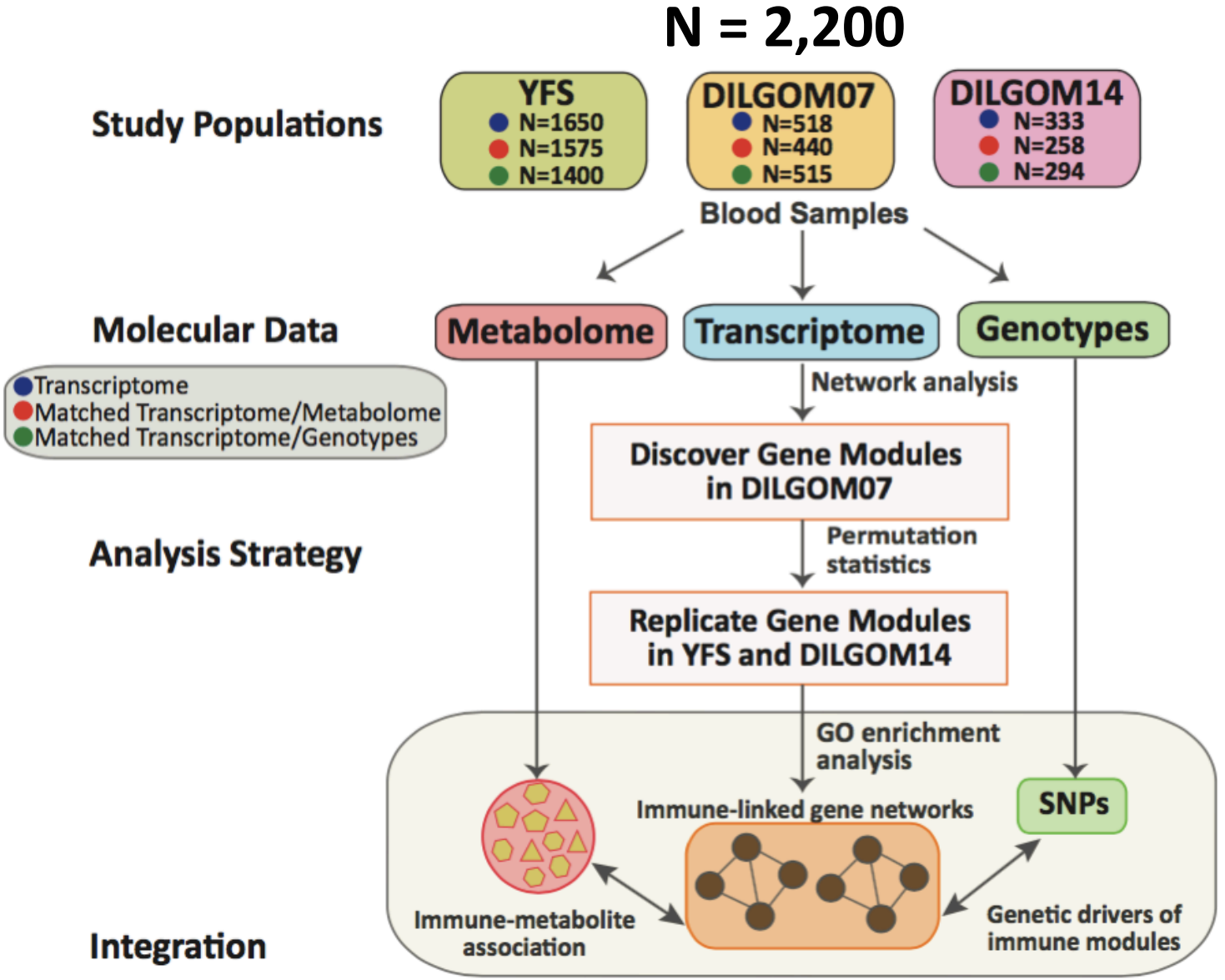


External validation



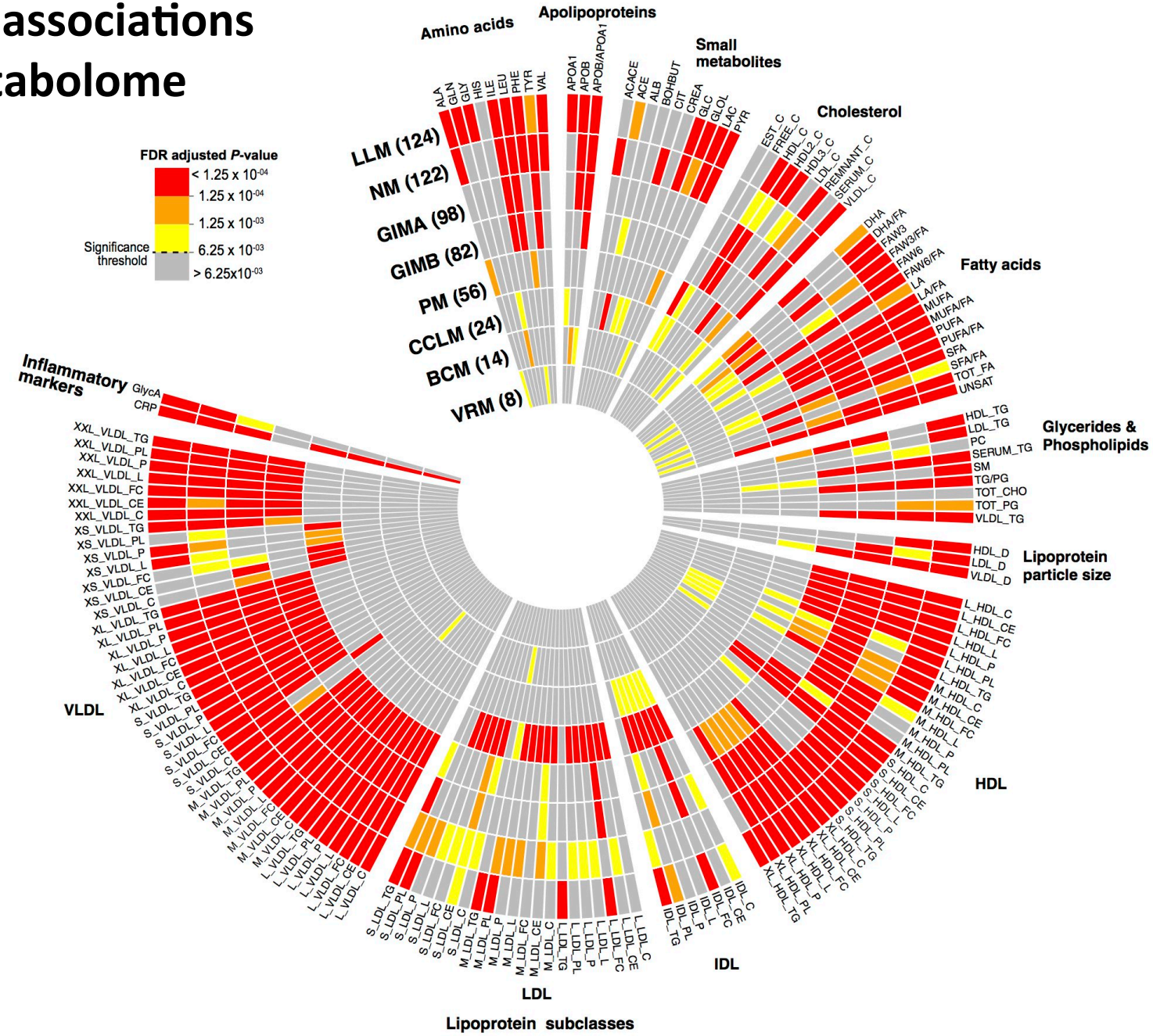
KORA cohort
N=700

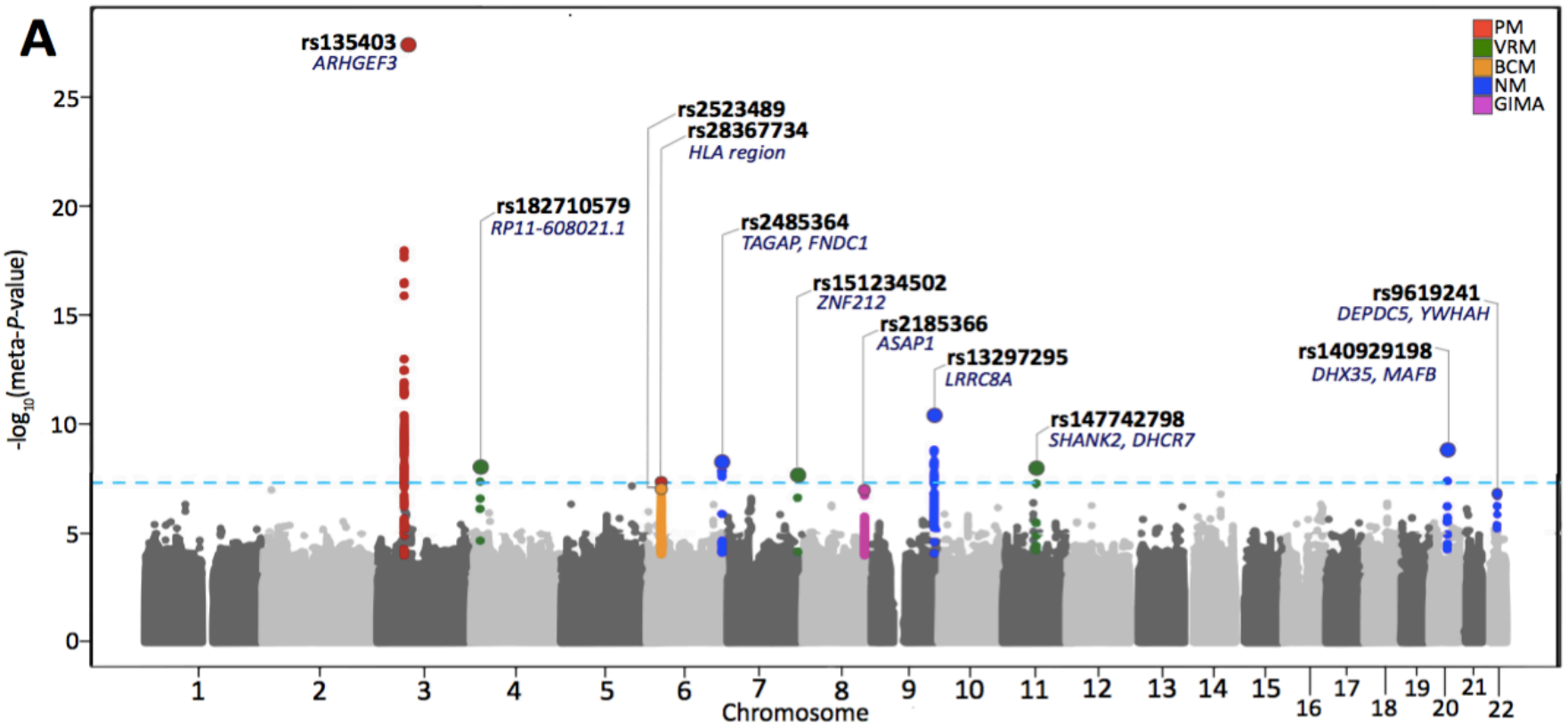
Bartel & Theis, PLoS Genetics 2015

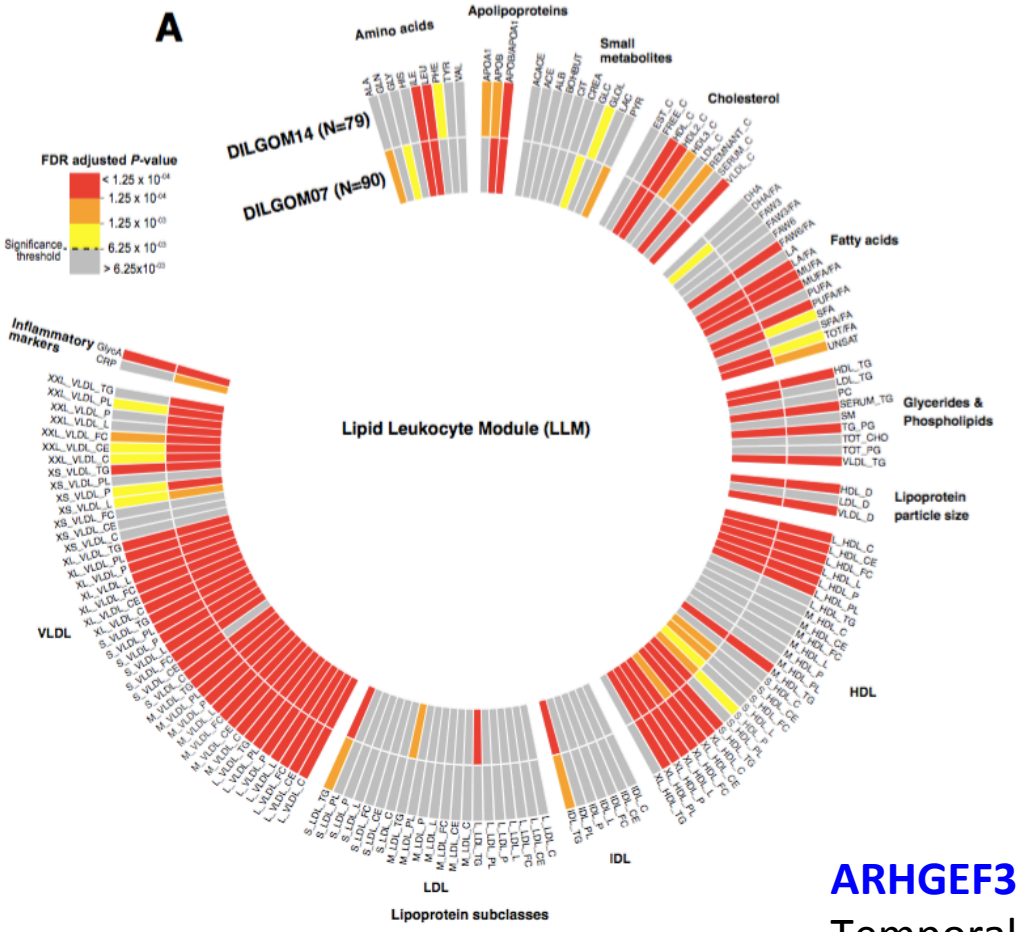


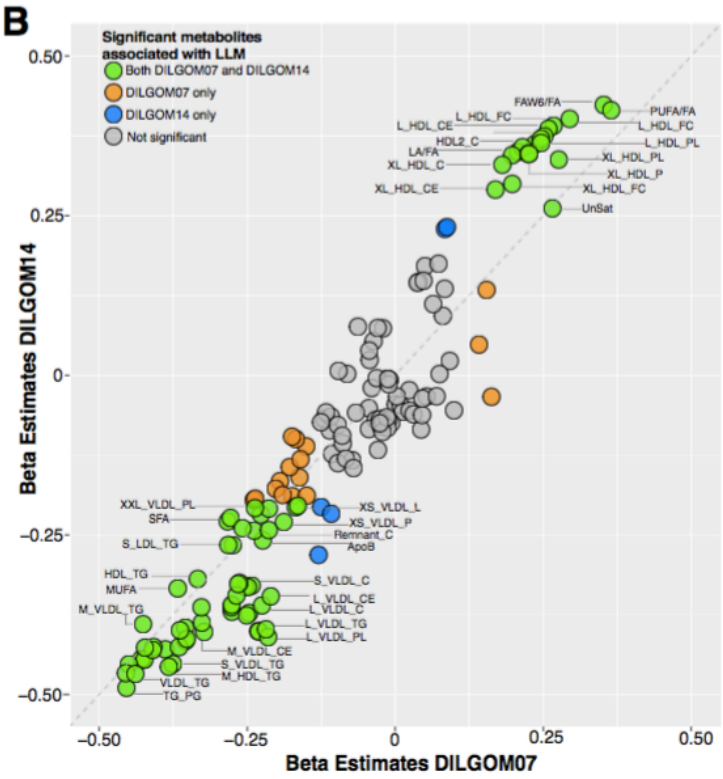
Nath, et al BioRxiv doi:10.1101/089839

Blood transcriptional network associations with metabolome




A
rs135403
ARHGEF3
rs2523489
rs28367734
HLA region
rs182710579
RP11-608021.1
rs2485364
TAGAP, FNDC1
rs151234502
ZNF212
rs2185366
ASAP1
rs13297295
LRRC8A
rs147742798
SHANK2, DHCR7
rs9619241
DEPDC5, YWHAH
rs140929198
DHX35, MAFB
PM
VRM
BCM
NM
GIMA
-log10(meta-P-value)
0
5
10
15
20
25
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome





ARHGEF3 – Platelet module

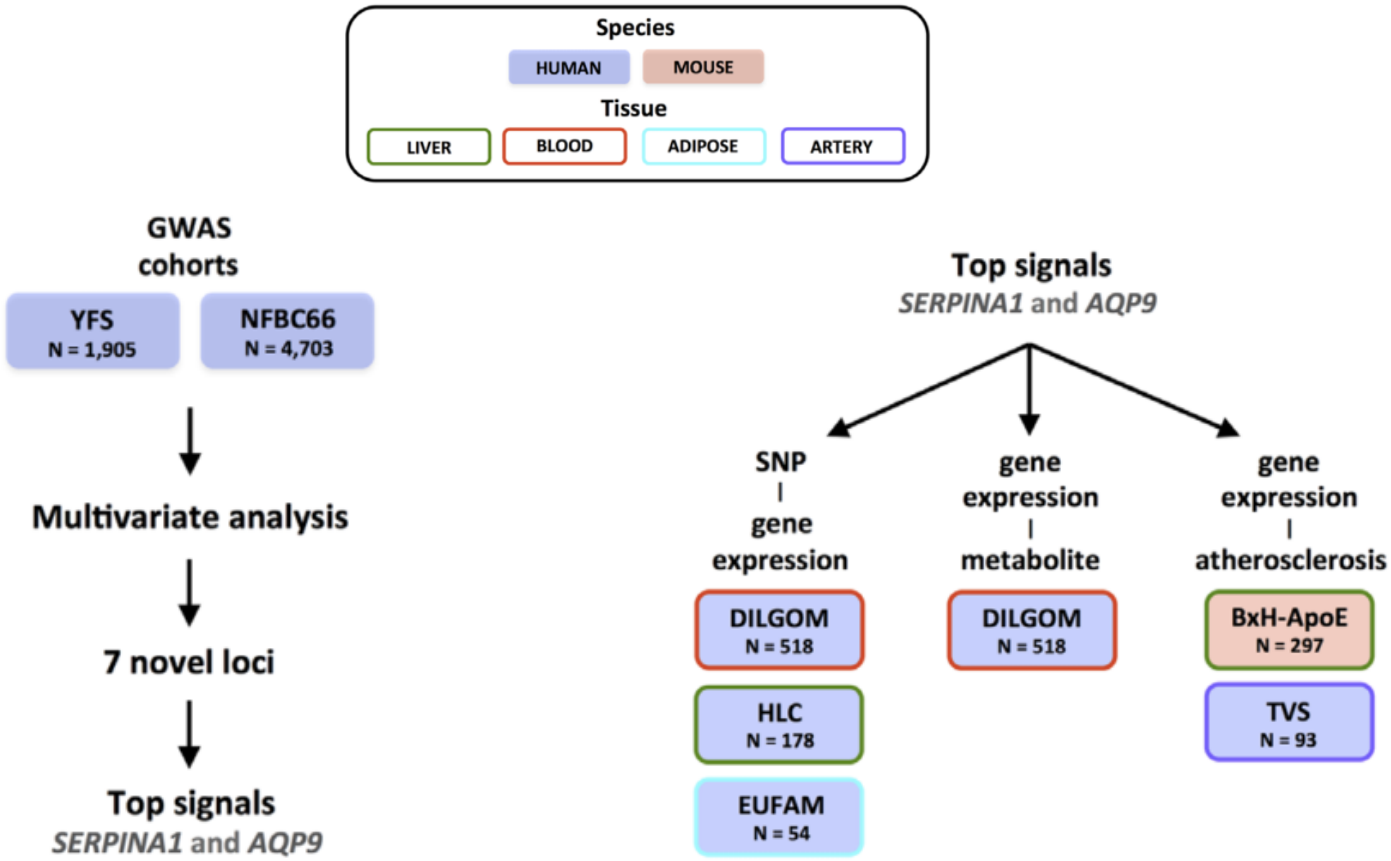
Temporally stable QTL

OPEN ACCESS Freely available online

PLOS | GENETICS

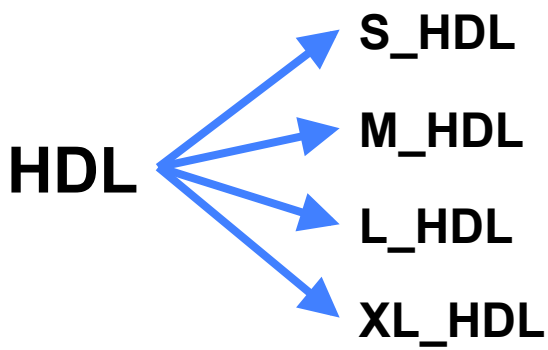
Novel Loci for Metabolic Networks and Multi-Tissue Expression Studies Reveal Genes for Atherosclerosis

Michael Inouye[1,2*], Samuli Ripatti[3,4,5], Johannes Kettunen[3,4], Leo-Pekka Lyytikäinen[6], Niku Oksala[6,7], Pirkka-Pekka Laurila[3,4,8], Antti J. Kangas[9], Pasi Soininen[9,10], Markku J. Savolainen[9,11,12], Jorma Viikari[13], Mika Kähönen[14], Markus Perola[4], Veikko Salomaa[4], Olli Raitakari[15], Terho Lehtimäki[6], Marja-Riitta Taskinen[16], Marjo-Riitta Järvelin[11,17,18], Mika Ala-Korpela[9,10,12,17], Aarno Palotie[3,5,8,19], Paul I. W. de Bakker[19,20,21,22]



Genetics of metabolism

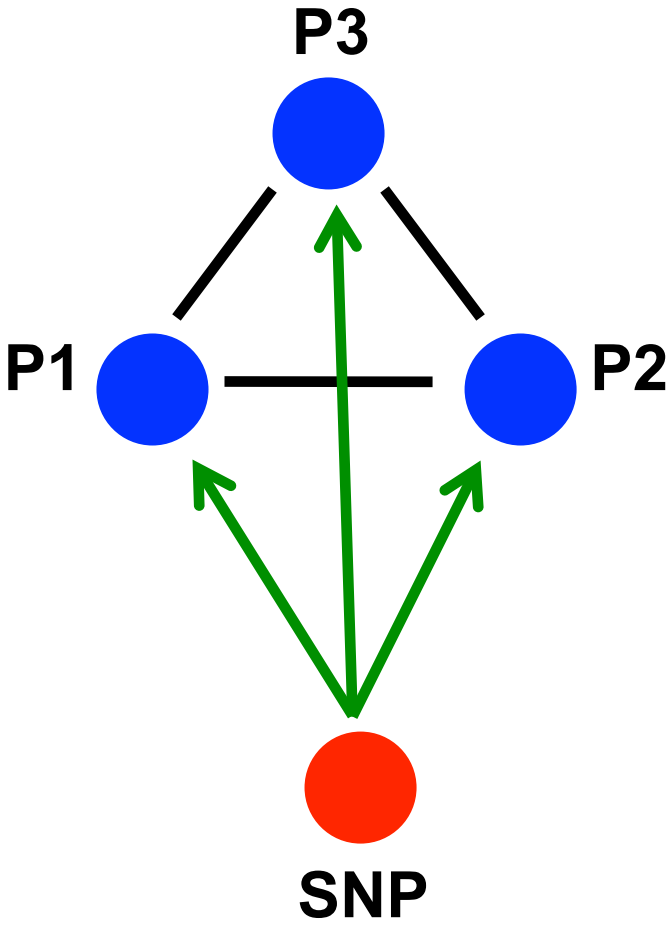
- **GWAS have found 100s of loci for blood metabolites**
  - Metabolic & cardiovascular disease, etc.
- **However, many are 'total' measures…**
  - Cholesterol
  - High-Density Lipoprotein (HDL)
  - Low-Density Lipoprotein (LDL)
- ***Fine-mapping phenotypes***
  - E.g. height decomposed to bone-lengths
  - 'Total' metabolic measures to metabolomics
  - The deeper we phenotype, the better we understand pathways

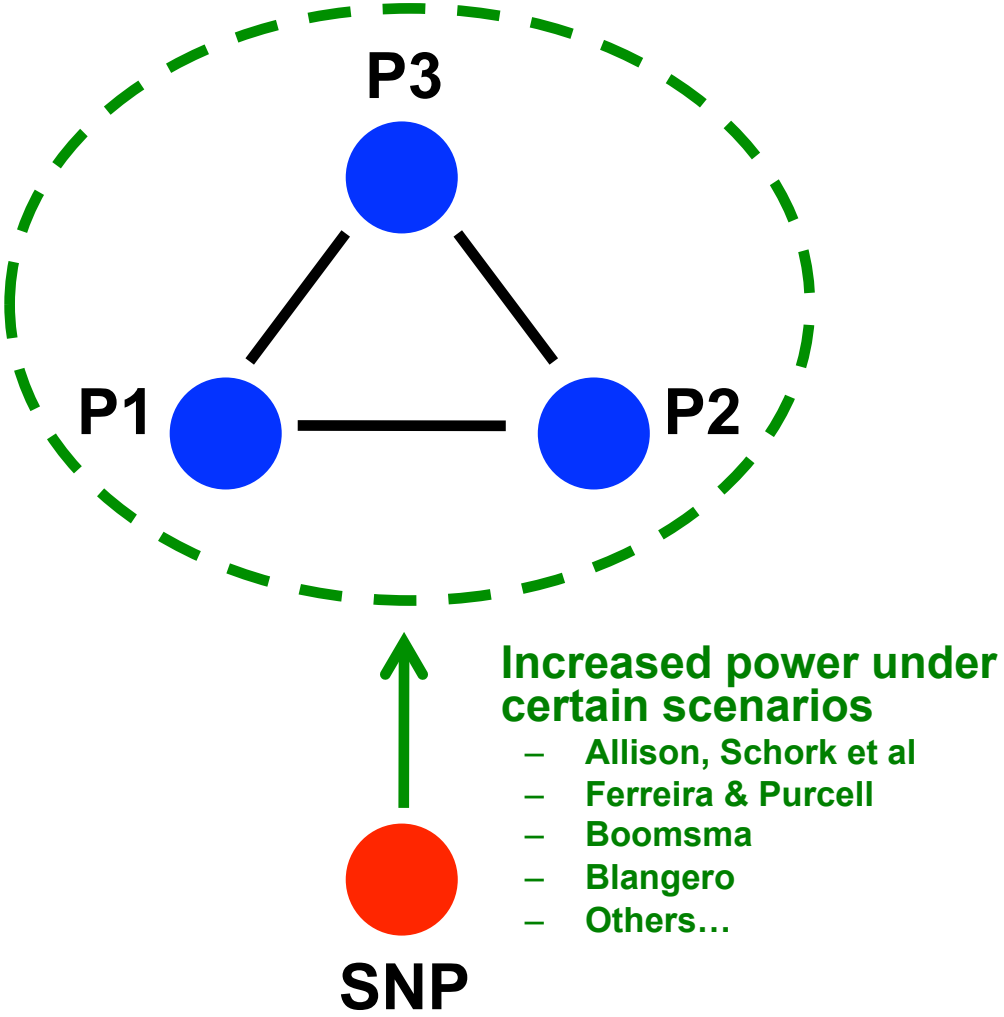


GWAS paradigm

- **1 SNP and 1 phenotype at a time**
  - Assumed independence of phenotypes
- **Phenotypes do not act in isolation**
  - Pleiotropy is common
- **A proportion of variance in one phenotype can be explained by another phenotype**

Can we leverage relationships among phenotypes?

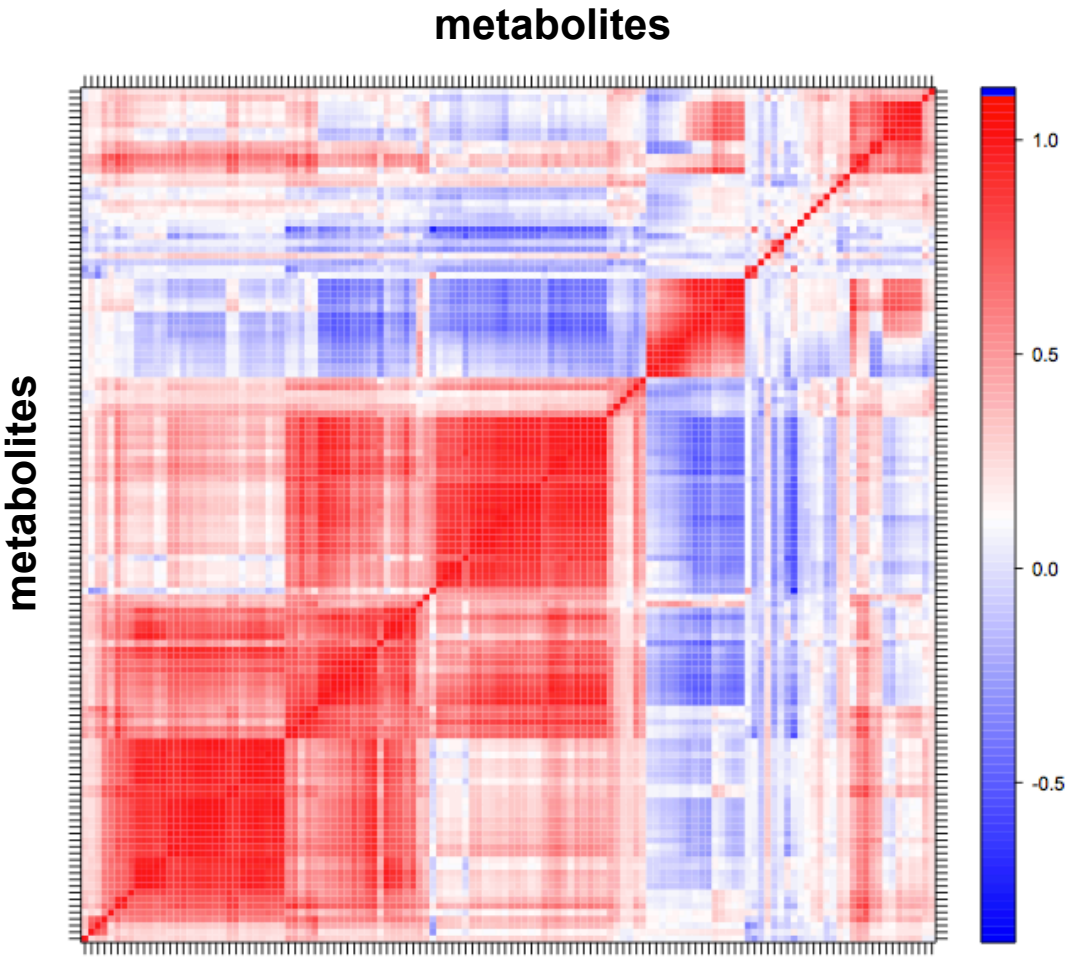




Increased power under certain scenarios

- Allison, Schork et al
- Ferreira & Purcell
- Boomsma
- Blangero
- Others…

Correlation structure of serum metabolome



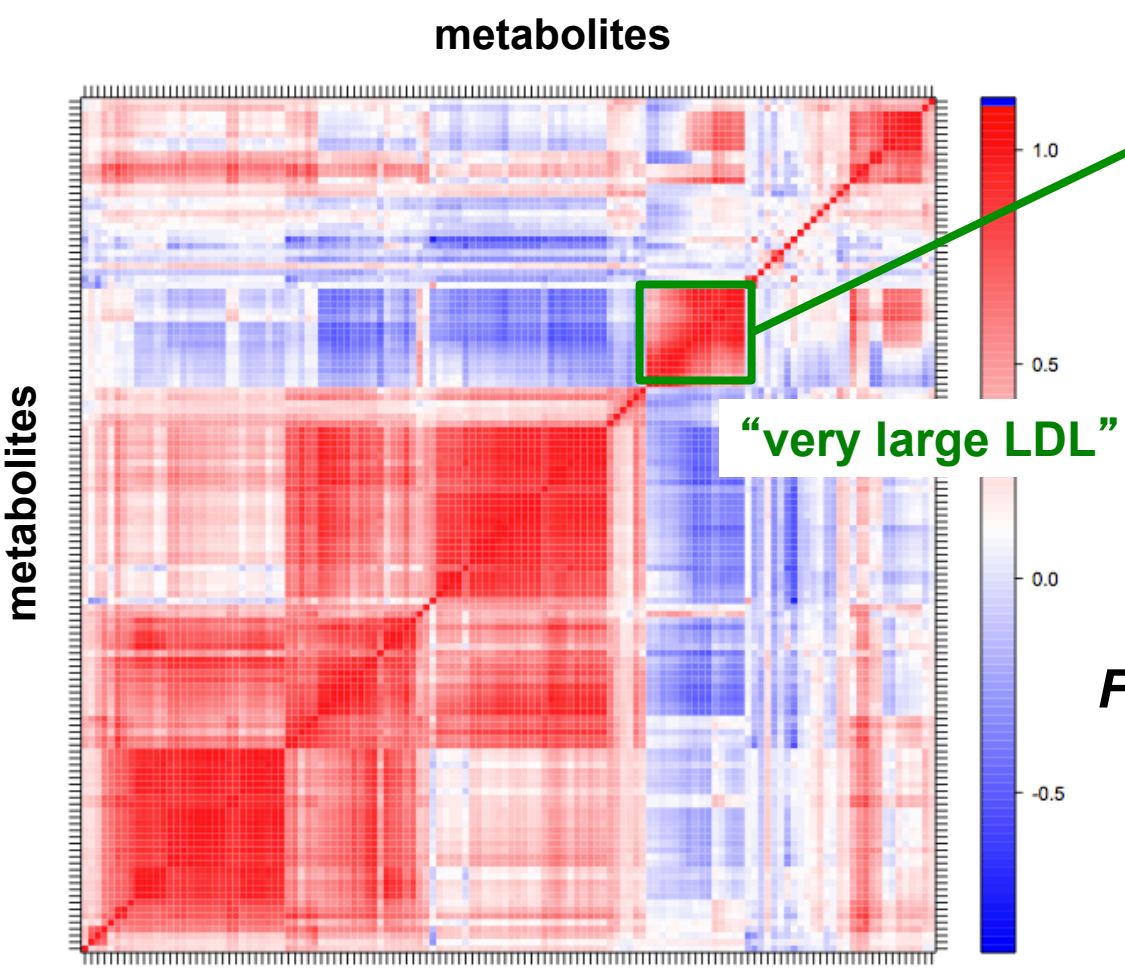
130 metabolites across 6,600 individuals

- **Lipoproteins**
- **Lipids**
- **Small molecules**
- **Amino acids**
- **Inflammatory markers**

How do we deal with such high-dimensional data?

- ***Multiple testing***
- ***Statistical power***

How do we leverage phenotype correlations?



Model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$$

SNP allele (0, 1, 2)

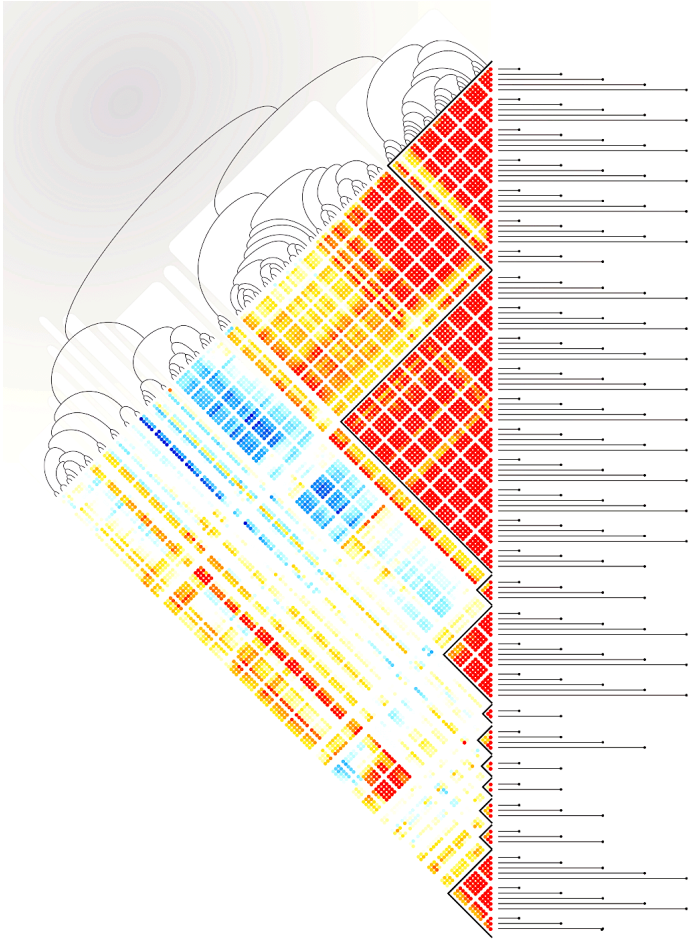
CCA – Wilks' lambda

$\Lambda = 1 - \rho^2$

***F* – approximation (Rao, 1973)**

See Galesloot et al *PLOS One* 2014 for comparison of MV methods (PLINK, BIMBAM, SNPTEST, Multiphen, etc)

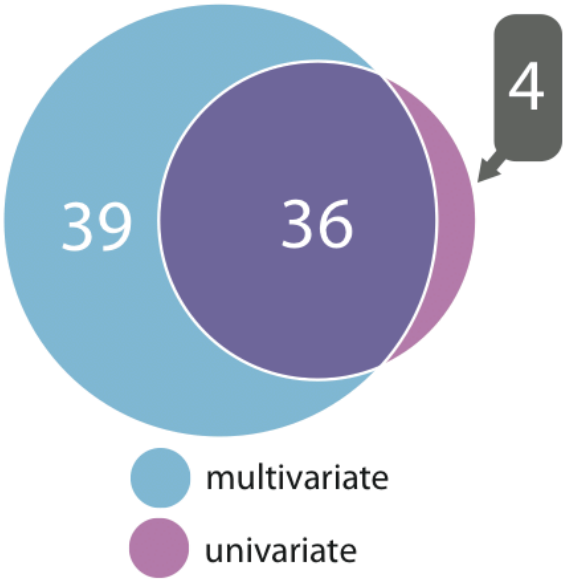
Determining metabolite networks



1 – apoB lipoproteins
2 – BC and aromatic amino acids
large TG-rich VLDL
3 – large HDL
4 – small HDL
5 – polyunsaturated lipids
6 – ketone bodies
7 – glucose-alanine cycle
8 – renal function
9 – FA chain length/composition
10 – LDL diameter and FA composition
11 – urea & acetate

Comparison of associations

Genetic associations



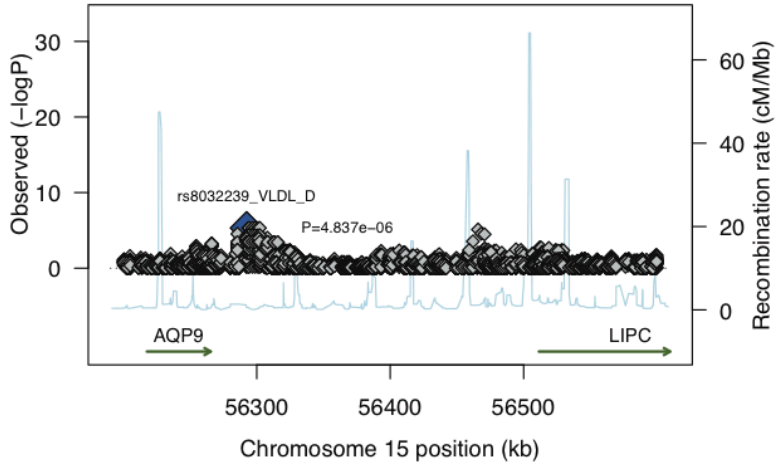
Total loci detected

	Univariate	Multivariate
YFS	3	8 (5)
NFBC66	15	25 (19)
Joint	23	34 (31)

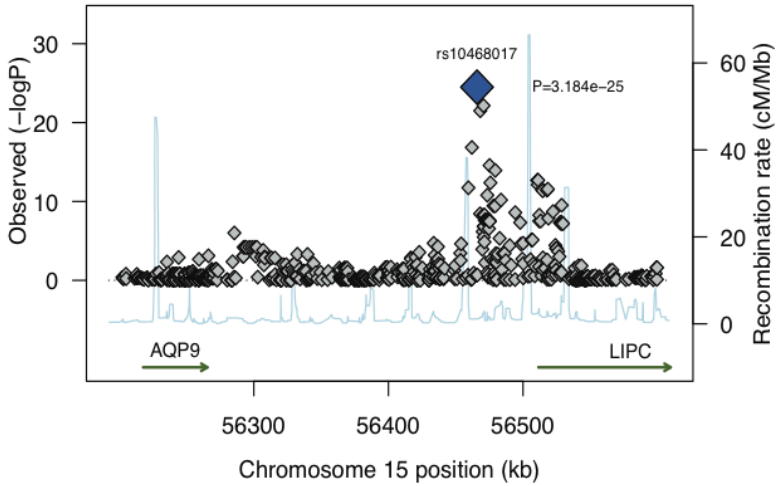
Example: hepatic triglyceride lipase (*LIPC*)

Univariate

All 26 metabolites in metabolite network 1

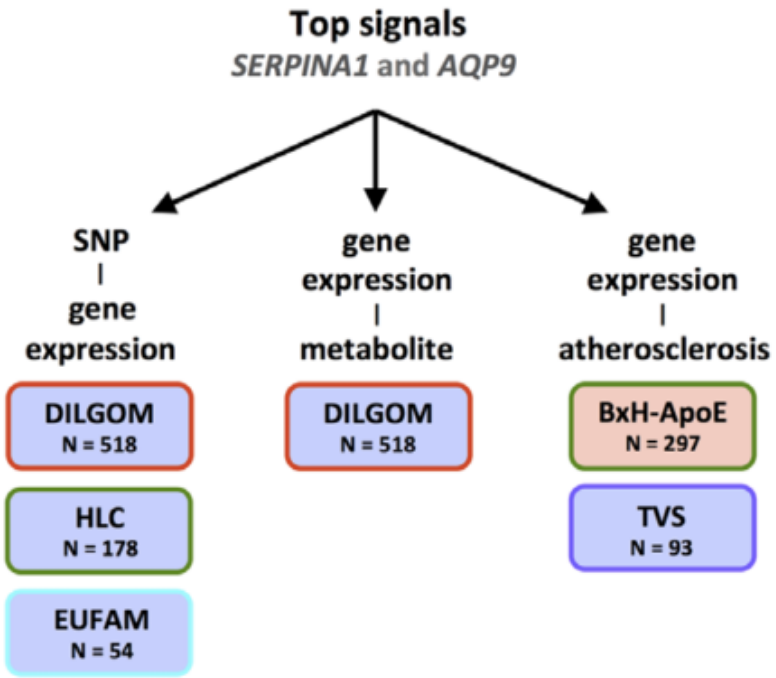


Multivariate

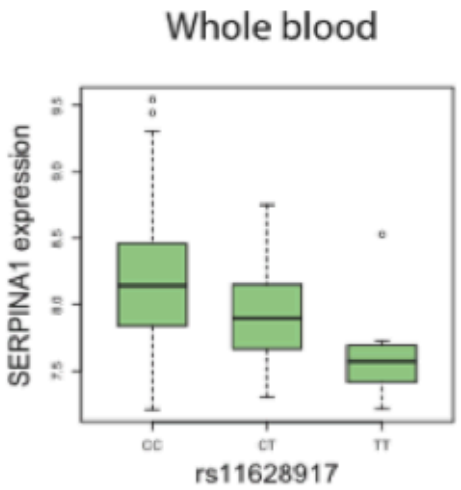


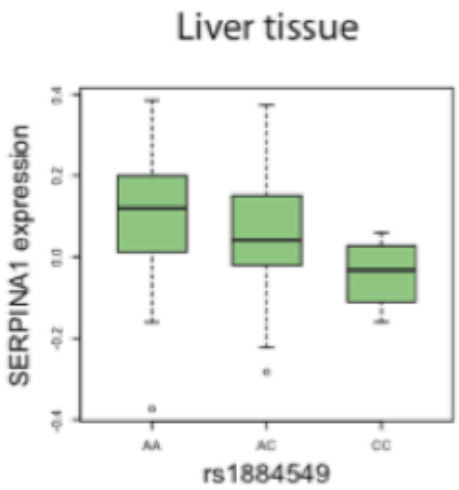
34 significant loci total
7 novel

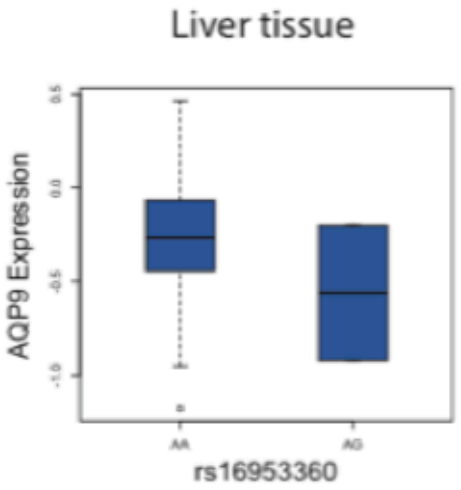
Metabolite networks	Top SNP	Chr	Pos	Top Pvalue	Top metabolite	Gene
1,2	rs1303	14	93914596	5x10^-48	IDL-C	SERPINA1
1,2,3,4	rs16939881	15	56259271	3x10^-27	XL-HDL-TG	AQP9



SNPs for metabolic networks also drive *AQP9* and *SERPINA1* expression







DILGOM
N = 518
P < 10^-10
R^2 = 0.07

SNP associated with metabolic networks 1, 2

HLC
N = 178
P = 4x10^-3
R^2 = 0.04

SNP associated with metabolic networks 1

HLC
N = 178
P = 5x10^-3
R^2 = 0.04

SNP associated with metabolic networks 1,2,3,4

SERPINA1

Identification of genetic variants influencing the human plasma proteome

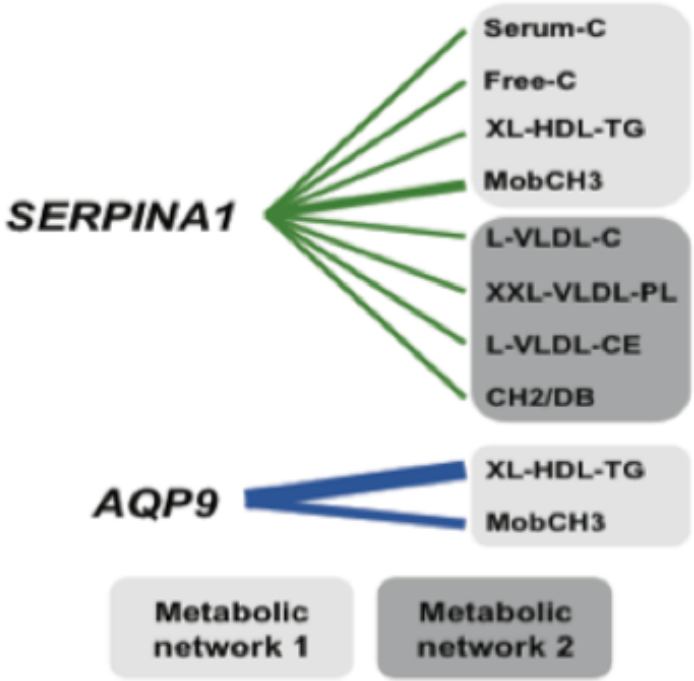
Åsa Johansson[a,b], Stefan Enroth[a], Magnus Palmblad[c], André M. Deelder[c], Jonas Bergquist[d], and Ulf Gyllensten[a,1]

[a]Department of Immunology, Genetics, and Pathology, Rudbeck Laboratory, SciLifeLab, Uppsala University, 75185 Uppsala, Sweden; [b]Uppsala Clinical Research Center, Uppsala University, 75237 Uppsala, Sweden; [c]Center för Proteomics and Metabolomics, Leiden University Medical Center, 2333 ZC, Leiden, The Netherlands; and [d]Department of Chemistry–Biomedical Centre, Analytical Chemistry, SciLifeLab, Uppsala University, 75124 Uppsala, Sweden

Edited* by Richard N. Zare, Stanford University, Stanford, CA, and approved February 11, 2013 (received for review October 8, 2012)

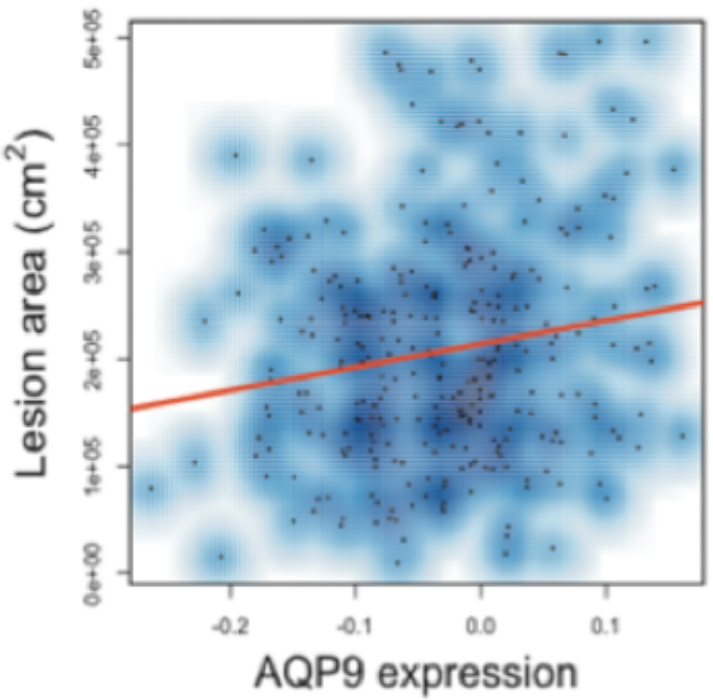
PNAS 2013

AQP9 and *SEPRINA1* expression is associated with metabolites



Metabolite	Chr	Position	Pvalue	Beta (95% CI)	Expressed Gene
XL-HDL-TG	15	56265176	8.48E-09	-0.61 (-0.82 - -0.41)	*AQP9*
MobCH3	15	56265176	7.16E-05	-0.43 (-0.63 - -0.22)	*AQP9*
MobCH3	14	93914570	6.46E-05	-0.51 (-0.75 - -0.26)	*SERPINA1*
L-VLDL-CE	14	93914570	2.47E-04	-0.49 (-0.76 - -0.23)	*SERPINA1*
XXL-VLDL-PL	14	93914570	2.48E-04	-0.51 (-0.78 - -0.24)	*SERPINA1*
L-VLDL-C	14	93914570	2.63E-04	-0.48 (-0.73 - -0.22)	*SERPINA1*
XL-HDL-TG	14	93924923	3.16E-04	-0.37 (-0.58 - -0.17)	*SERPINA1*
Free-C	14	93924923	3.98E-04	-0.37 (-0.57 - -0.16)	*SERPINA1*
Serum-C	14	93924923	4.00E-04	-0.37 (-0.57 - -0.17)	*SERPINA1*
CH2/DB	14	93924789	4.21E-04	1.14 (0.51 - 1.77)	*SERPINA1*

Liver *AQP9* associated with atherosclerosis in mouse model



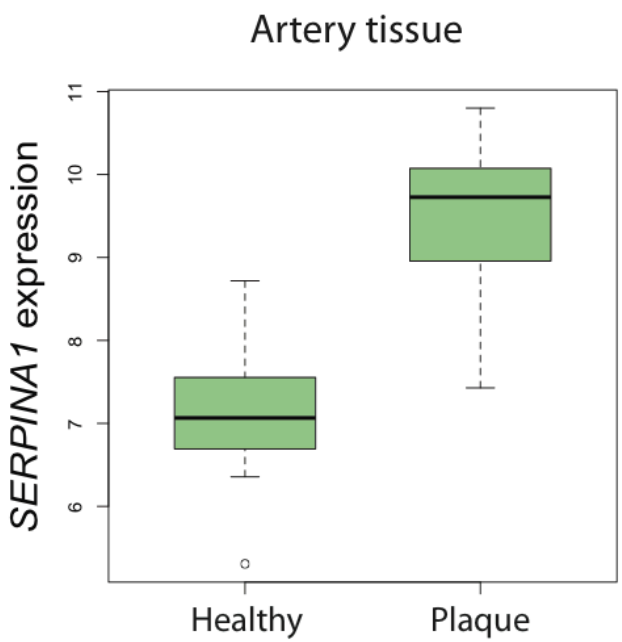
BxH-ApoE (N = 297):

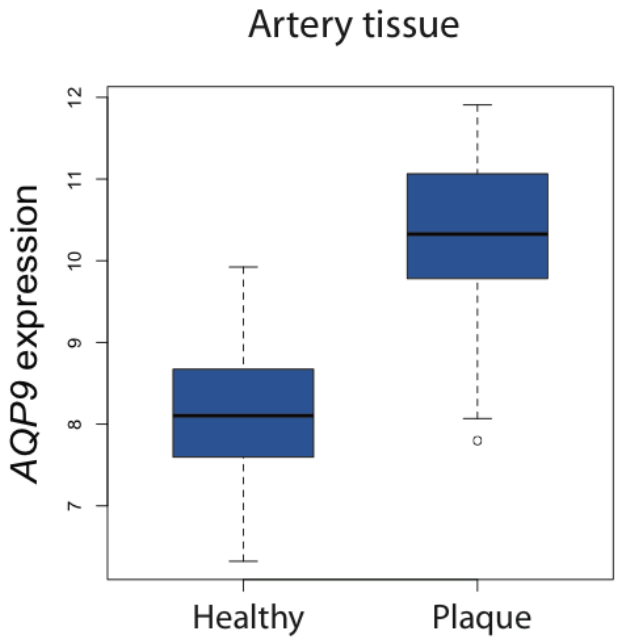
- derived from backcross of highly susceptible to atherosclerosis (C57BL/6J *ApoE*$^{-/-}$) and highly resistant (C3H/HeJ *ApoE*$^{-/-}$).
- Fed on high-fat, western diet for 16 weeks then euthanized at 24 weeks.

***P* = 5x10^-3**

Samples in top decile of *AQP9* expression have on average 30% larger lesion area than those in bottom decile

AQP9 & *SERPINA1* in human aorta





Summary

- **Integrative omics is a highly promising and evolving field with many challenges to be addressed**
- **Transcriptome and scRNA-seq are rapidly advancing in size and scope**
- **Global patterns vs intriguing specific examples**
- **Transcriptome-metabolome interactions are extensive (at least in blood)**
- **Leverage networks for statistical power (with care)**

Accessible resources for integrative genomics

- SageBase (via Sage BioNetworks)
- UK BioBank
- ImmGen
- ImmVar
- ENCODE
- THL Biobank
- TwinsUK
- iHMP / HMP2
- GTEx
- Epigenomics Roadmap Project
- Collaborative Cross (~outbred mice)
- Coming Soon: Precision Medicine Initiative