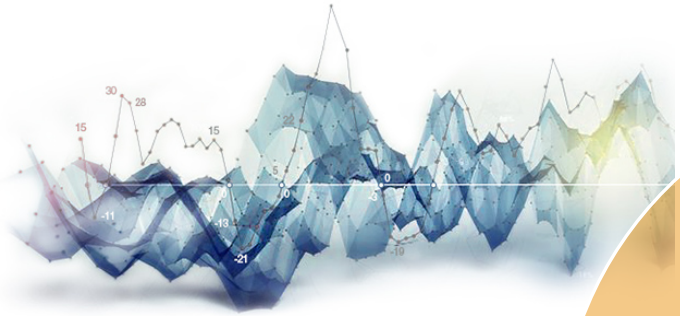
# Multivariate Multiscale Impacts of Genetic Variants on Gene Expression Variability in Humans

JAMES CAI

1/20/2017



Computational Statistics

Data Science

Medical Genetics

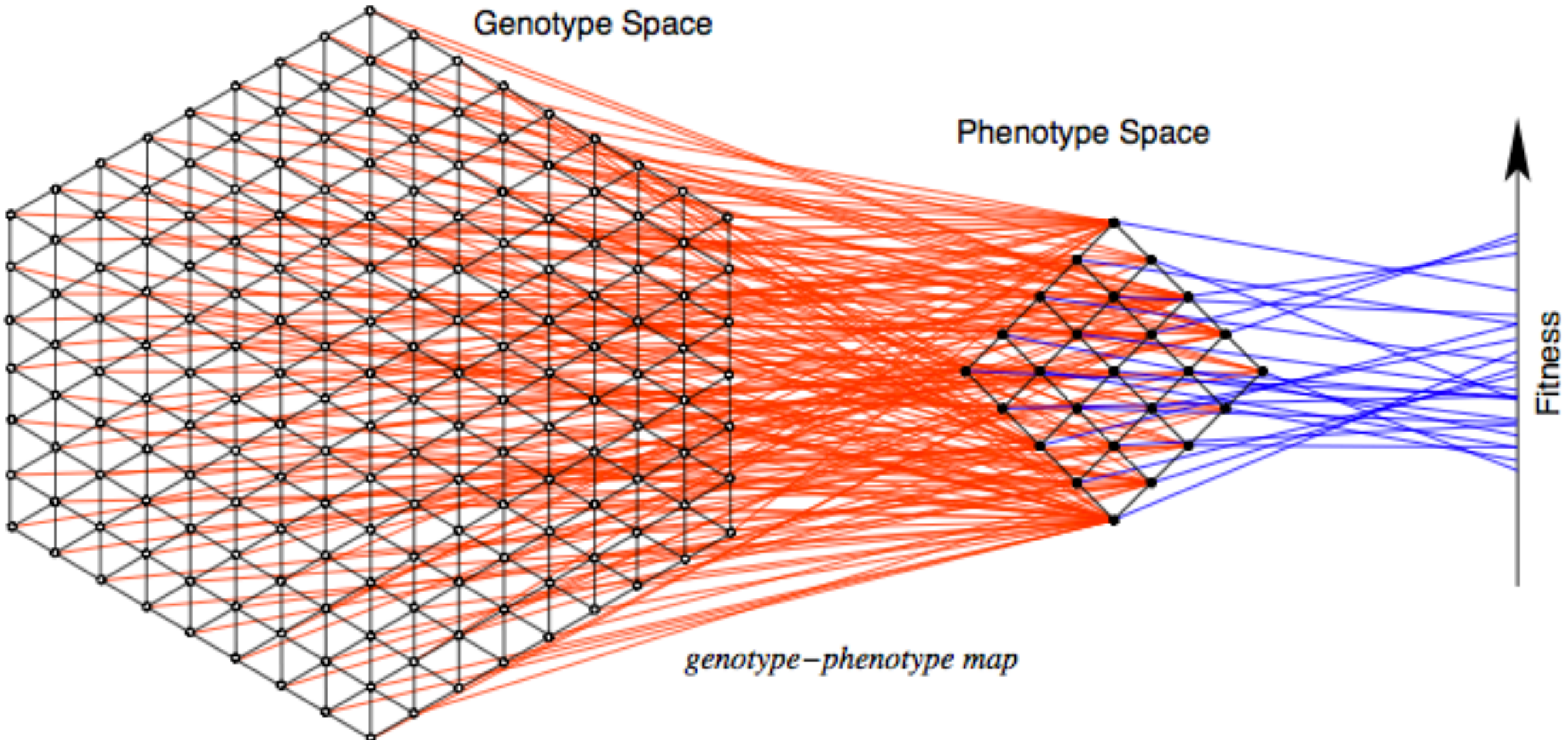
# Outline

Additive, epistatic, and environmental effects through the lens of **evQTLs**

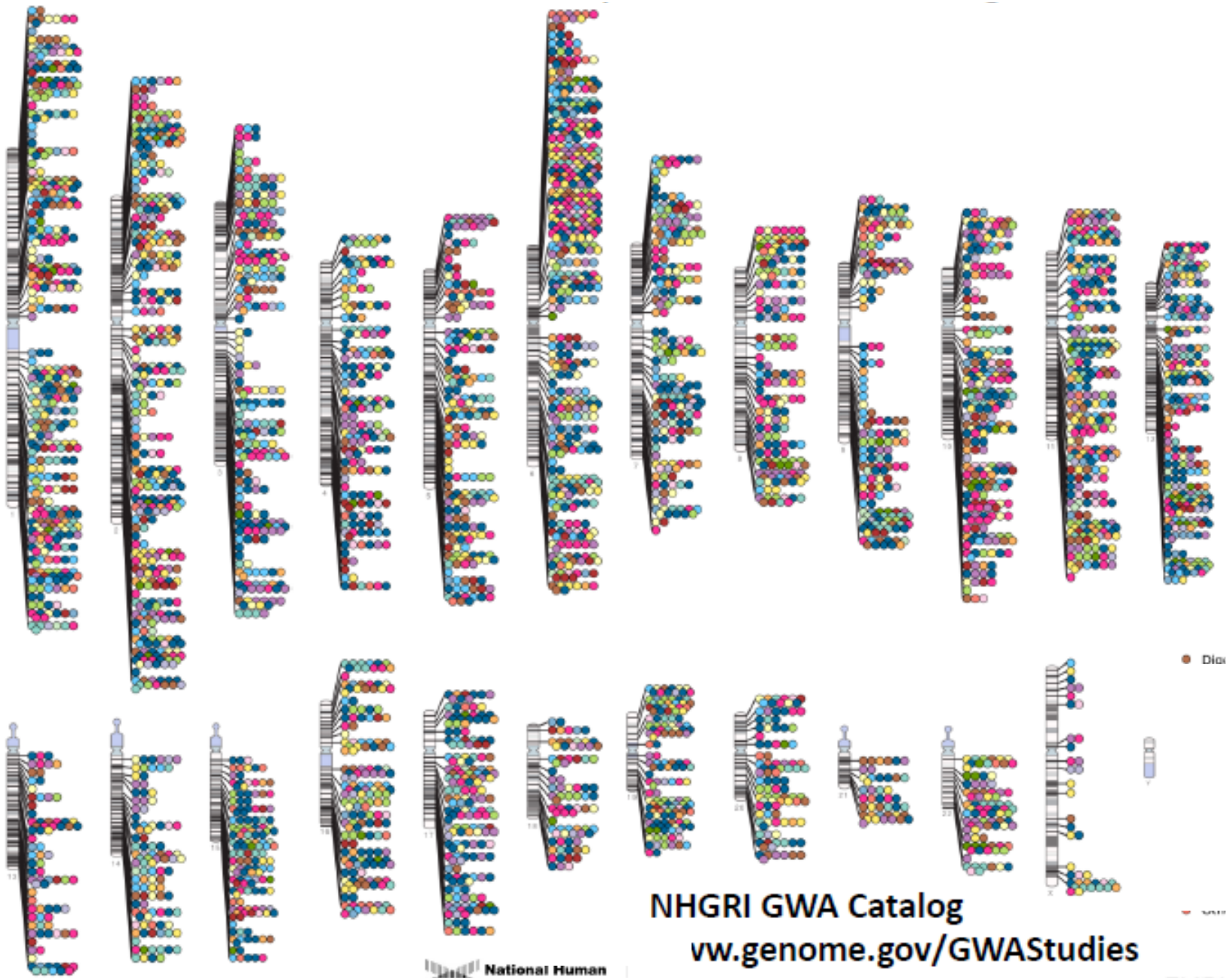
Exploiting **aberrant gene expression** in autism for gene discovery and diagnosis

# Additive, epistatic, and environmental effects through the lens of **evQTLs**

Effect of **common** genetic variants on gene expression variability



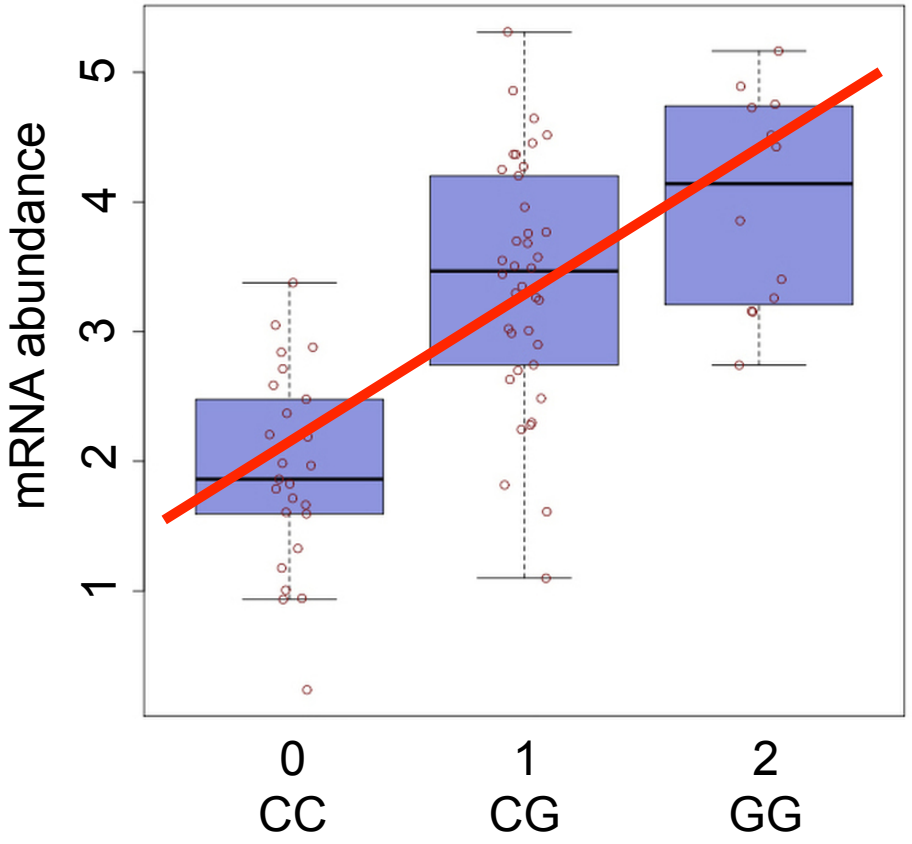

*Biological Evolution and Statistical Physics*, pp. 56–83. Springer-Verlag, Berlin, 2002




NHGRI GWA Catalog
vw.genome.gov/GWAStudies

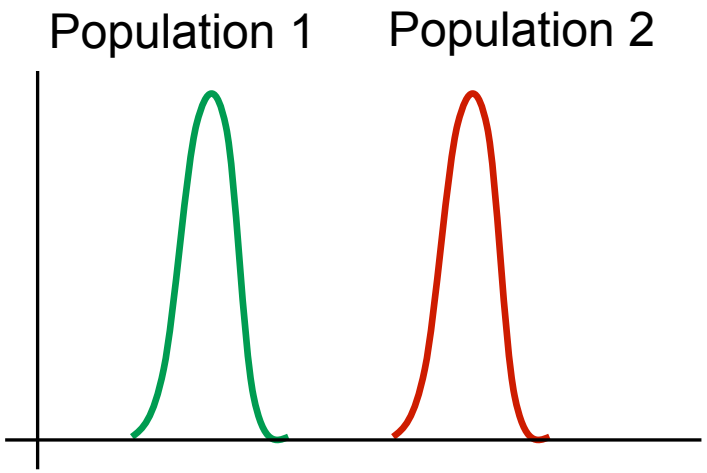
- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurment
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

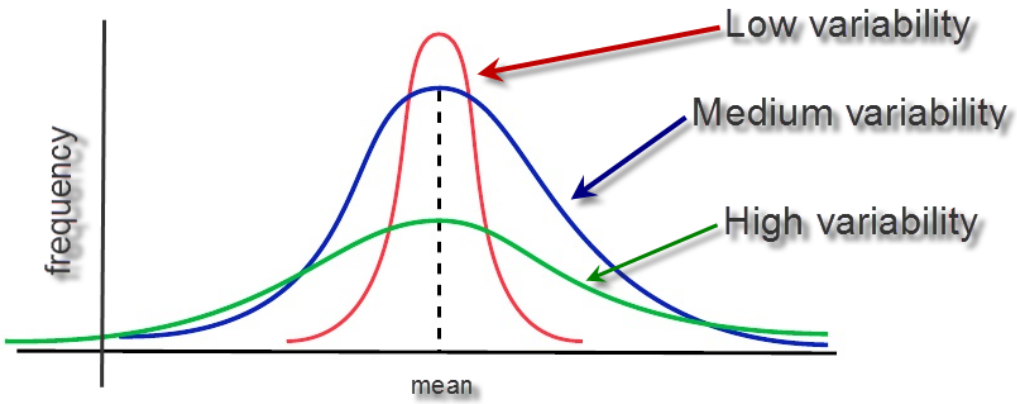
# Expression QTLs (eQTLs)




Gene expression level as an "intermediate phenotype"

$$y_i = \mu + x_i \beta + g_i \alpha + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma^2)$$




# Variation vs. Variability

# New evidence: **phenotypic variability** (variance) is genetically controlled

*FTO* genotype is associated with phenotypic variability of body mass index (**Yang *et al. Nature* 2012**)

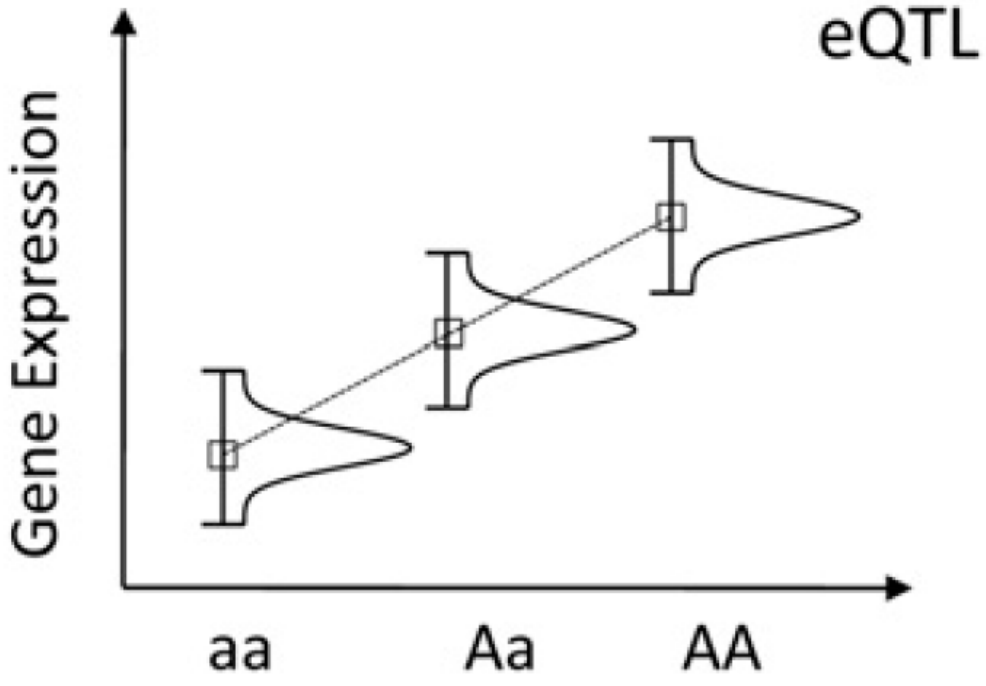
Inheritance beyond plain heritability: variance-controlling genes in *Arabidopsis thaliana* (**Shen *et al*. *PLoS Genet* 2012**)

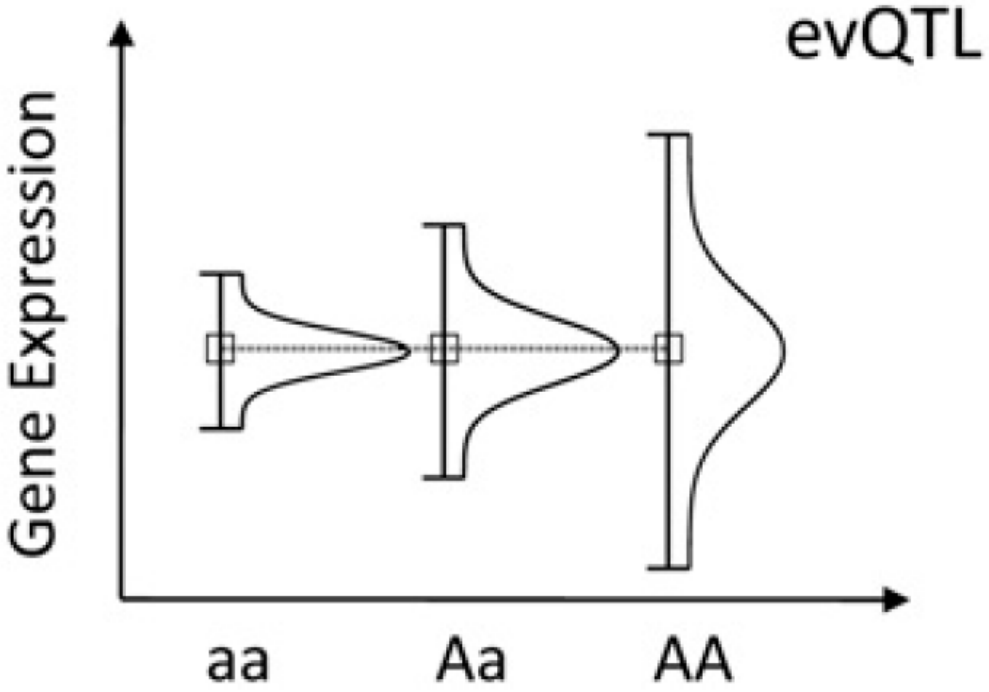
Behavioral idiosyncrasy reveals genetic control of phenotypic variability (**Julien *et al*. *PNAS* 2015**)

Selection on noise constrains variation in a eukaryotic promoter (**Metzger *et al*. *Nature* 2015**)

# Expression variability QTL – evQTL

i.e., genetic loci linked to or associated with expression variance




Hulse & Cai *Genetics* 2013

# Detection of evQTLs

## Linear regression model

$$y_i = \mu + x_i \beta + g_i \alpha + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma^2)$$

## Double generalized linear model (DGLM)

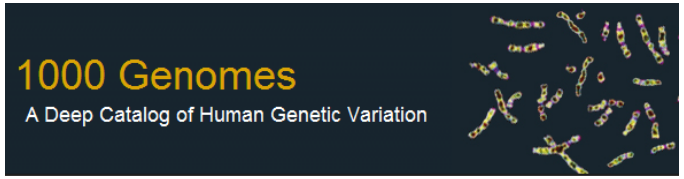
$$y_i = \mu + x_i \beta + g_i \alpha + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma^2 \exp(g_i \theta))$$

Smyth *J R Statist Soc B* 1989, Rönnegård & Valdar *Genetics* 2011

# Genome scan for evQTLs

**Data Sets:**

1. Genotype data from the **1000G** project




2. RNA-seq data from the **Geuvadis** project


**Detect *cis*-evQTLs**
**(DGLM method)**
28,494,473 tests performed
(15,124 genes vs. common SNPs
[MAF>15%])

↓

17,949 significant SNP-gene pairs
(Bonferroni corrected $P<0.05$)
1,304 genes

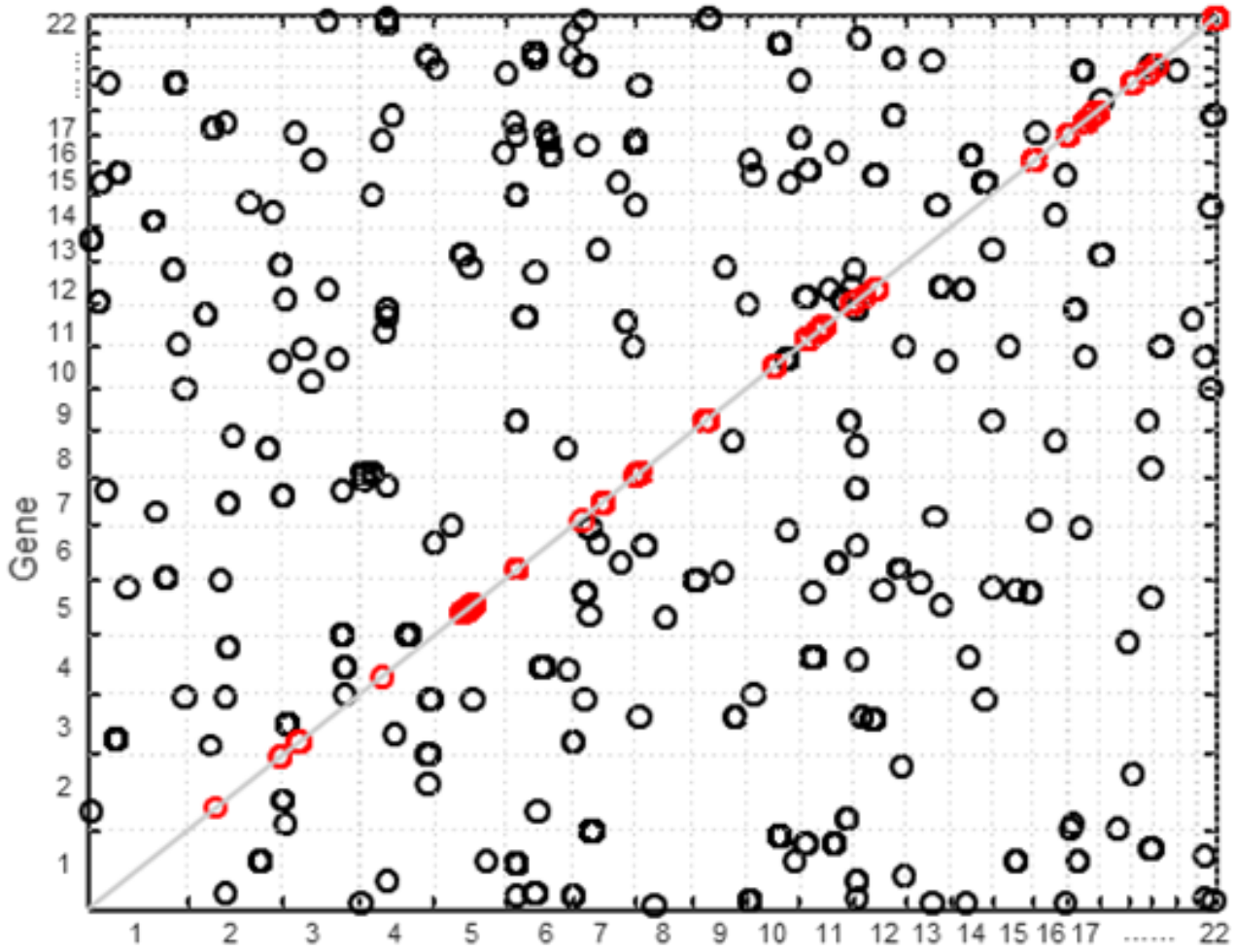
Yang et al. (Cai) *Hum Mol Genet* 2016

**Detect *cis*- and *trans*-evQTLs (SVLM method)**
43,637,670,424 tests performed (15,124 genes vs. 2,885,326 common SNPs [MAF>15%])

1,513 significant SNP-gene pairs (Benjamini-Hochberg FDR<0.1)
268 genes

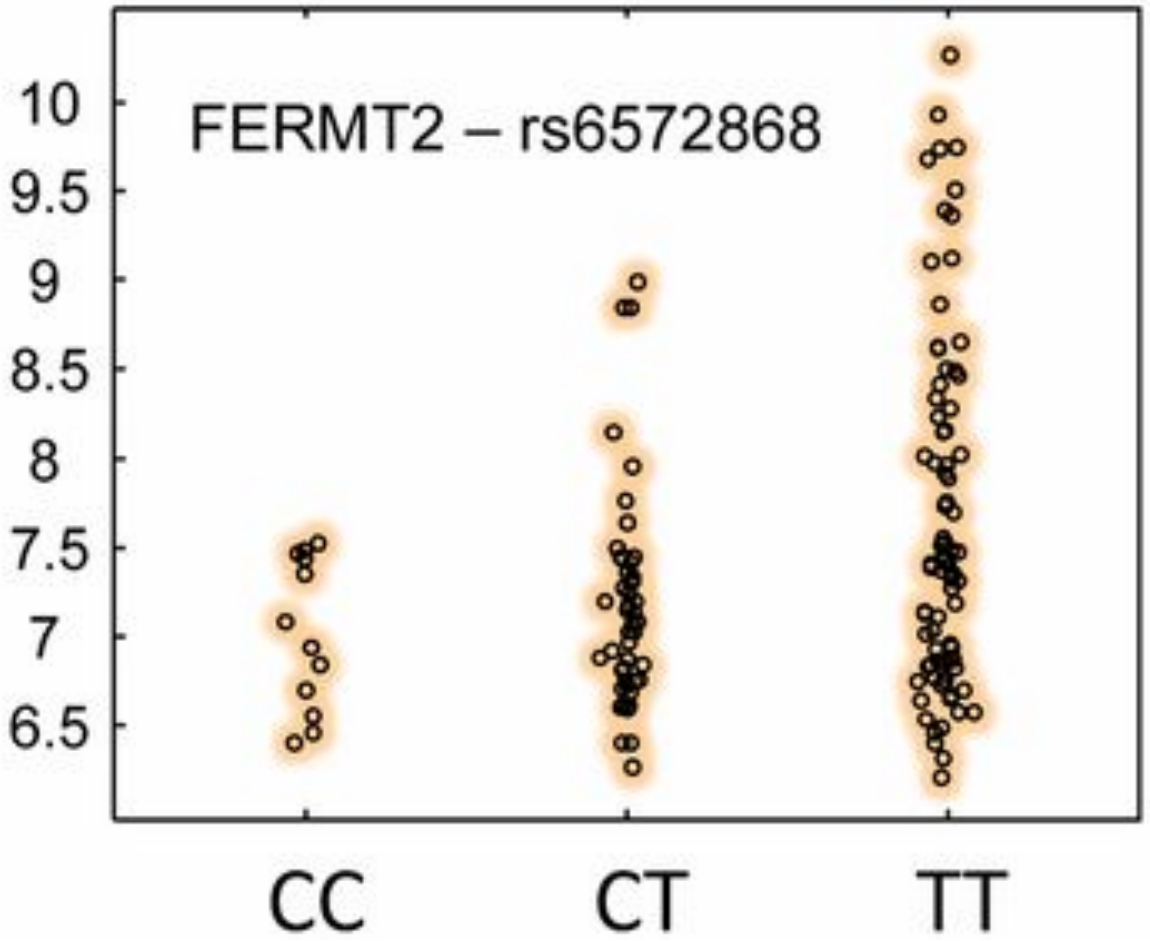
33 *cis*-evQTL genes

235 *trans*-evQTL genes



# Expression variability QTL – evQTL

i.e., genetic loci linked to or associated with expression variance




Hulse & Cai *Genetics* 2013

GTEx Portal

2016-07-08

New Multi-tissue eQTL Comparison Visualization

Read More >>

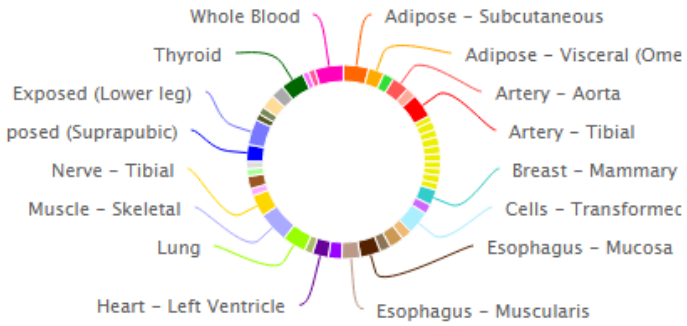
## Current Release

Latest Version: V6p

Dataset Summary Statistics Report

Browse eQTL Tissues

Total samples in all eQTL tissues: 7051

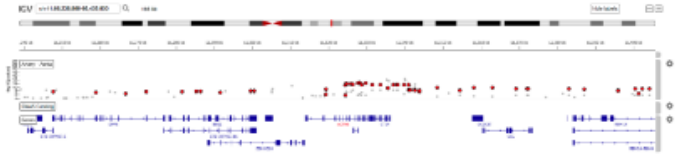



## Genetic Association

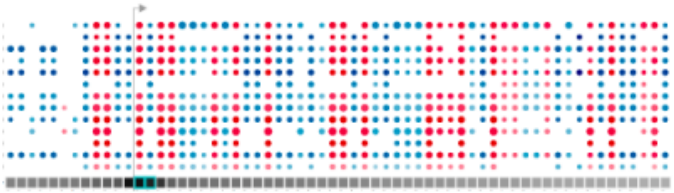
Single Tissue eQTLs

Search eQTL by gene or SNP ID

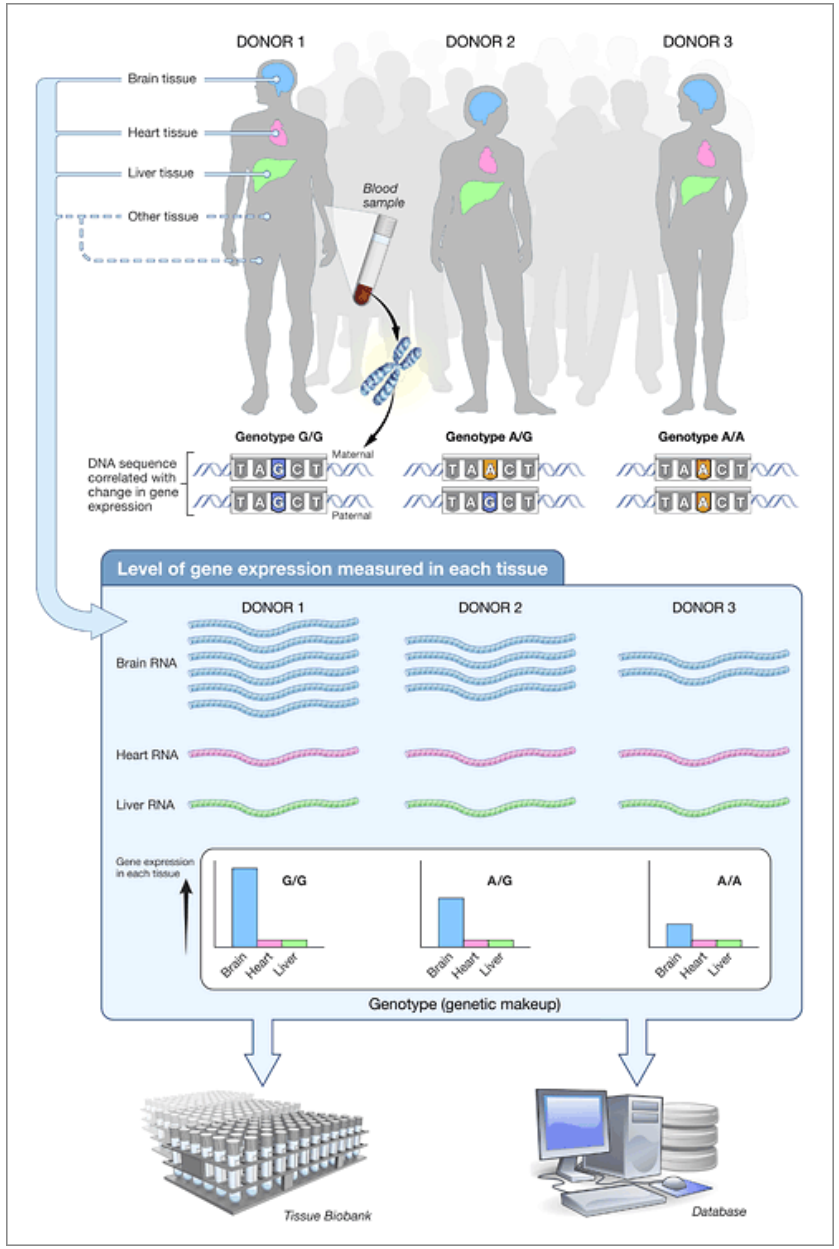
eQTL IGV Browser

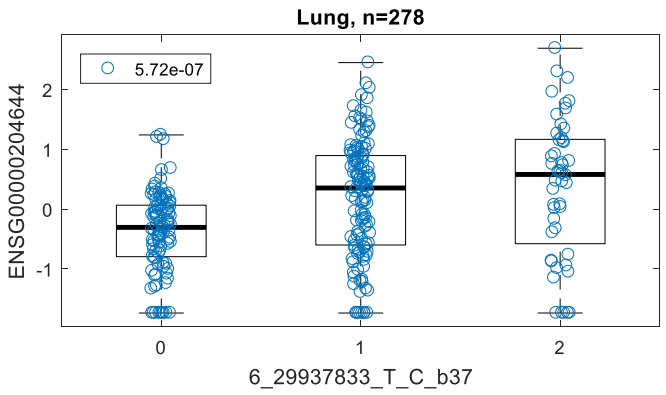


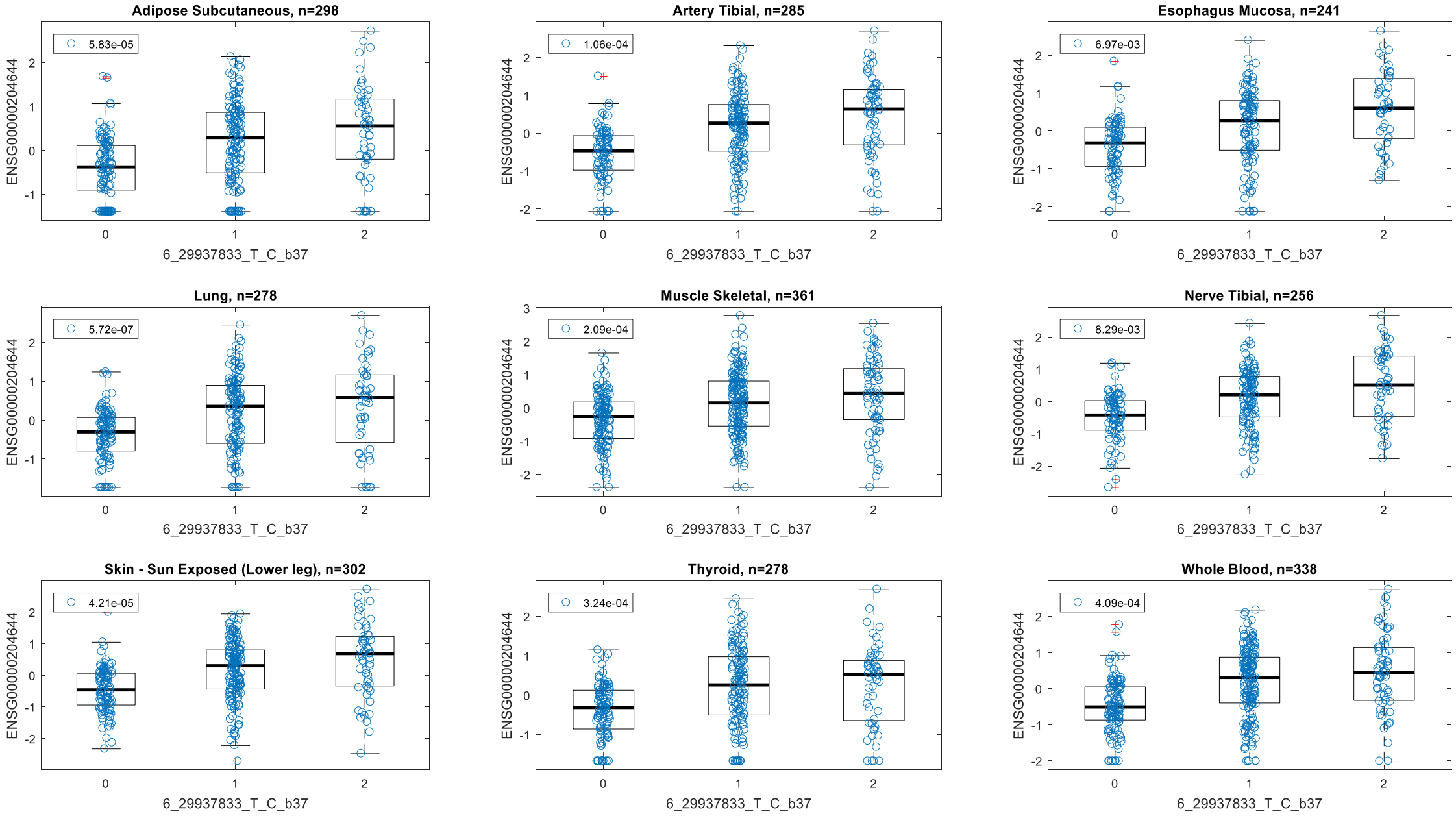
Gene eQTL Visualizer



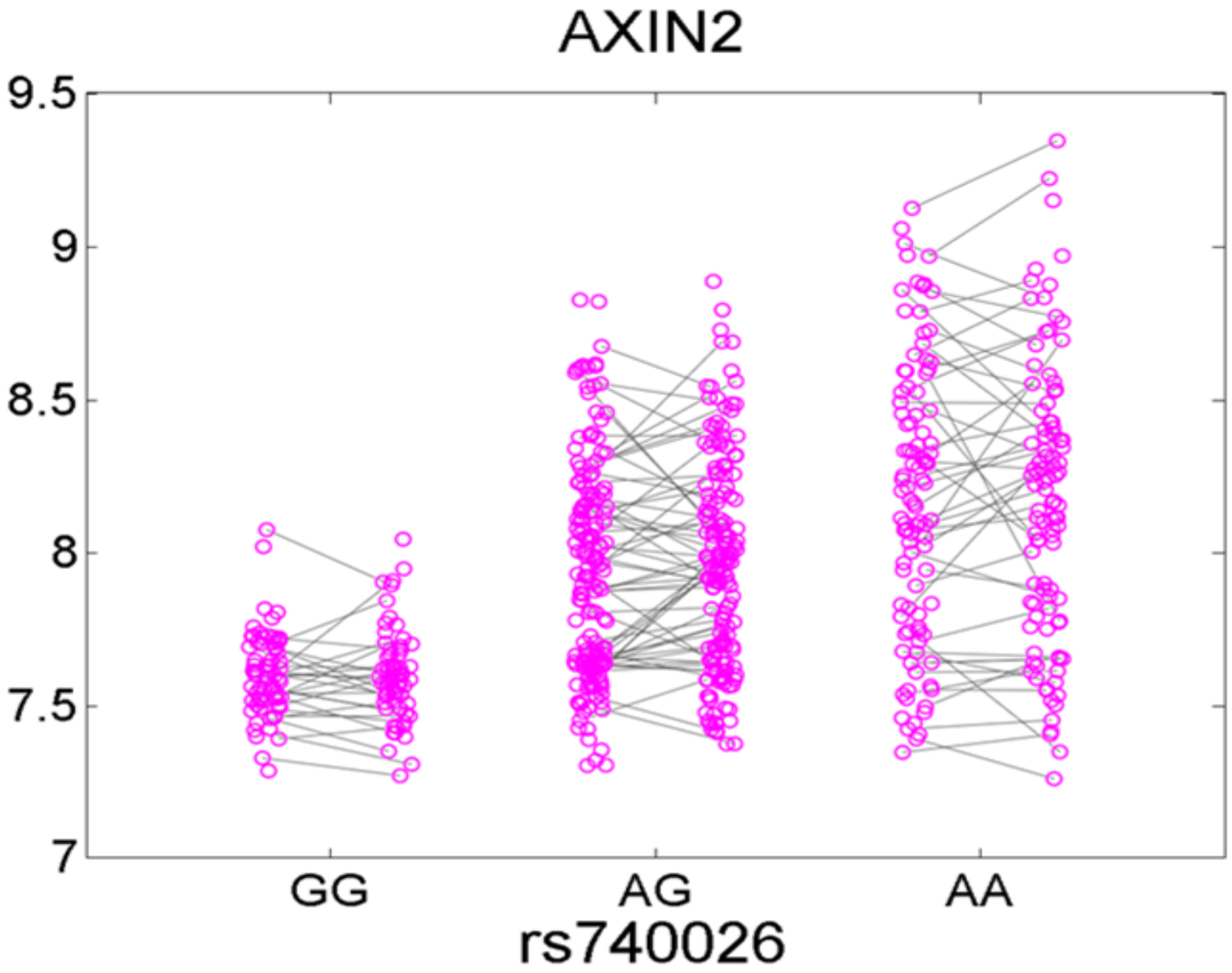
View eQTL data of a gene...



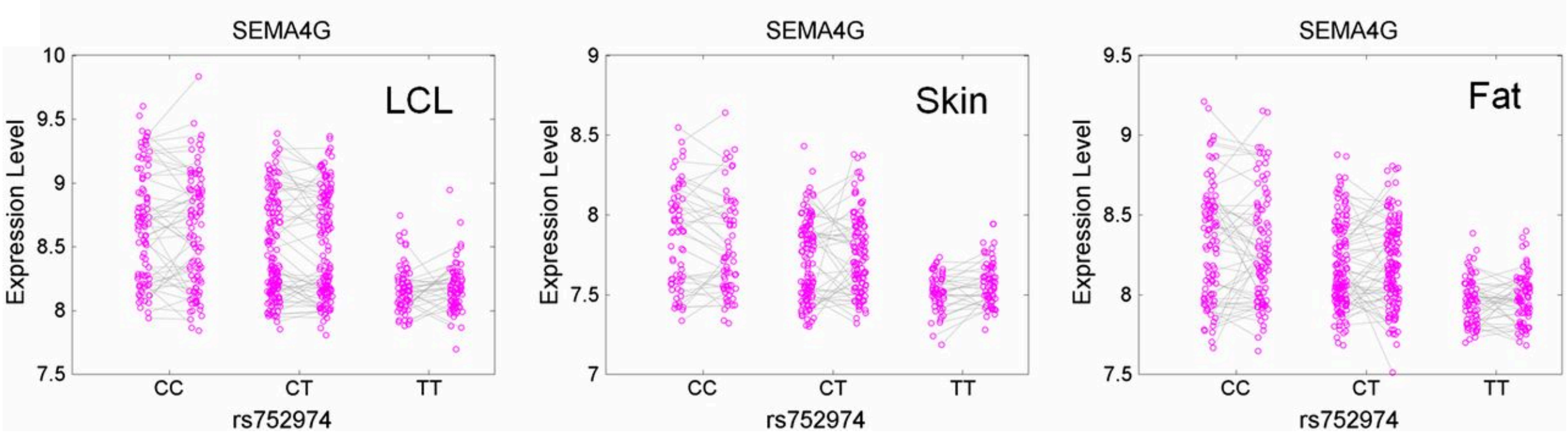

Lung, n=278
5.72e-07
ENSG00000204644
2
1
0
-1
0
1
2
6_29937833_T_C_b37


Adipose Subcutaneous, n=298
5.83e-05
Artery Tibial, n=285
1.06e-04
Esophagus Mucosa, n=241
6.97e-03
Lung, n=278
5.72e-07
Muscle Skeletal, n=361
2.09e-04
Nerve Tibial, n=256
8.29e-03
Skin - Sun Exposed (Lower leg), n=302
4.21e-05
Thyroid, n=278
3.24e-04
Whole Blood, n=338
4.09e-04
ENSG00000204644
6_29937833_T_C_b37


Tim Spector




Wang et al. (Cai) *Genetics* 2014



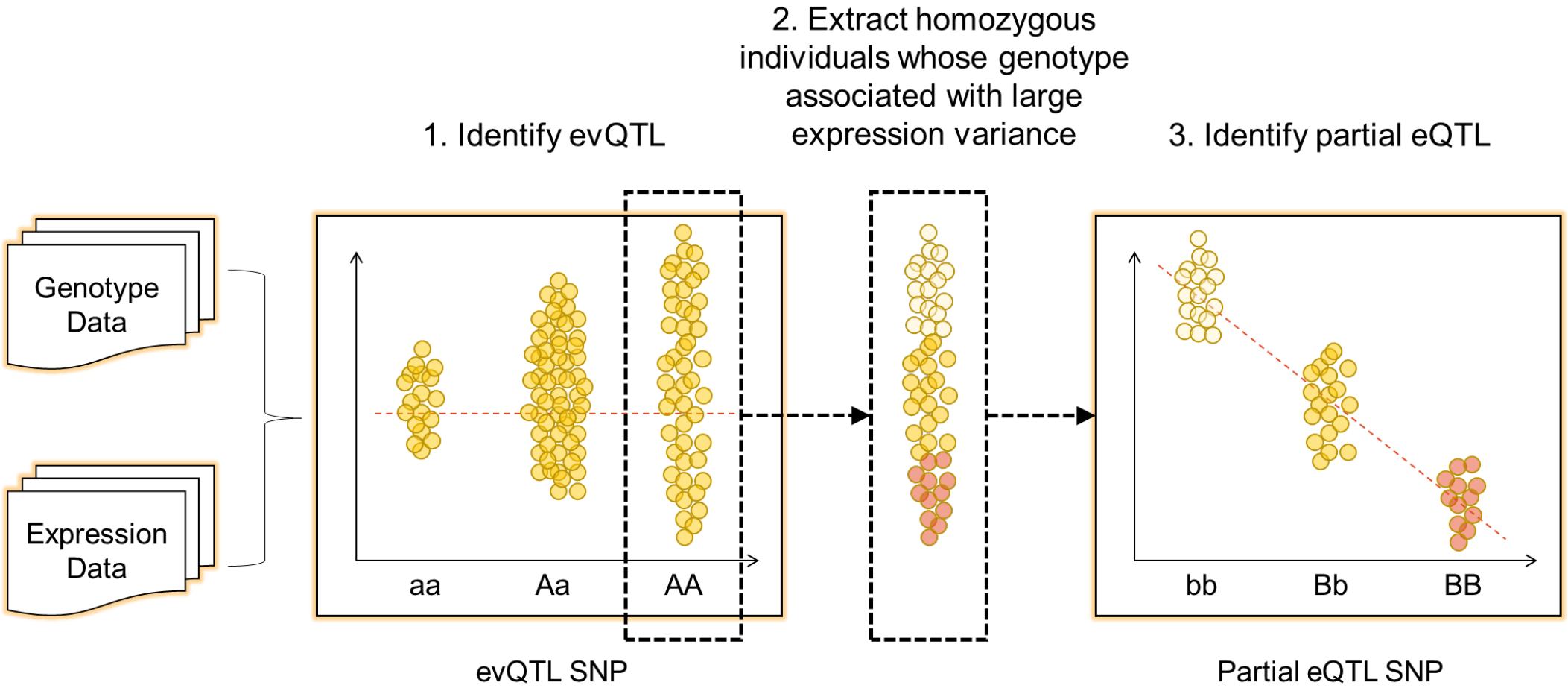

Wang et al. (Cai) *Genetics* 2014

# **Two distinct models** explaining the creation of evQTLs

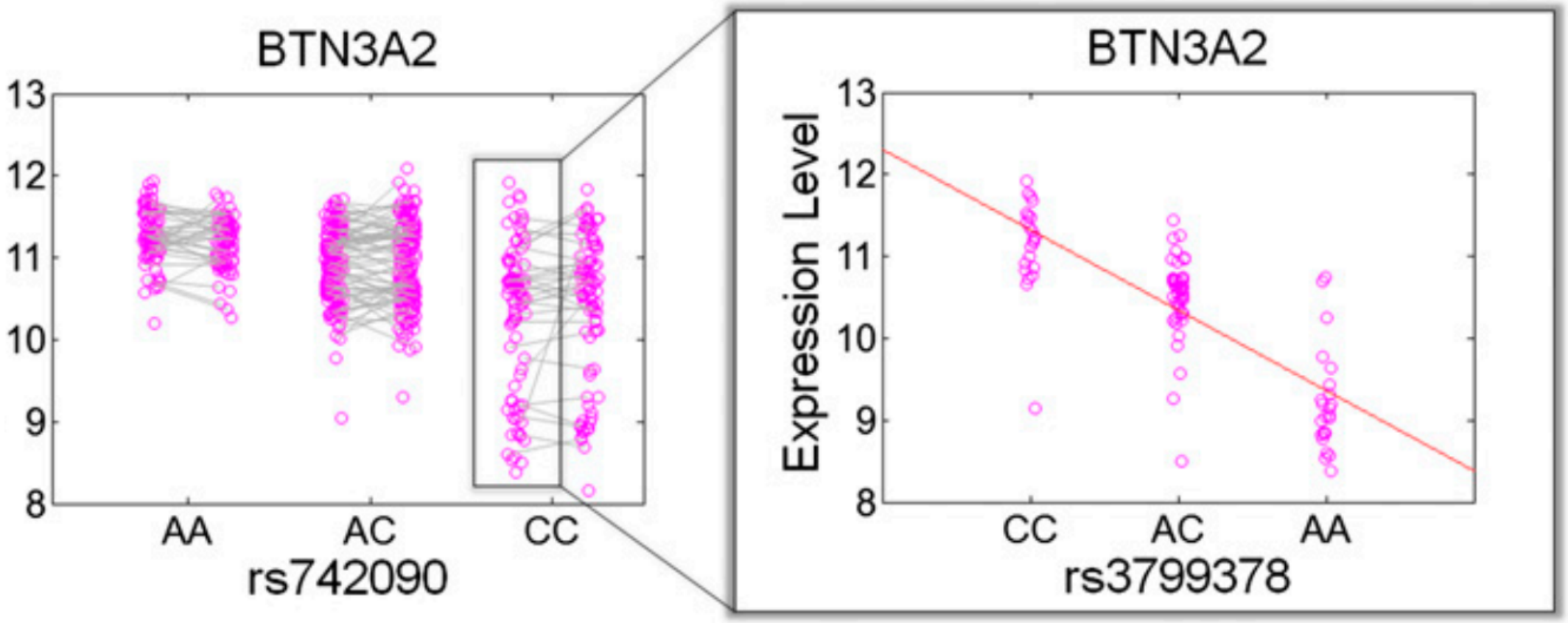
**GxG (epistasis):** the interaction between genotypes

**GxE (destabilization):** the interaction between genotype and environment

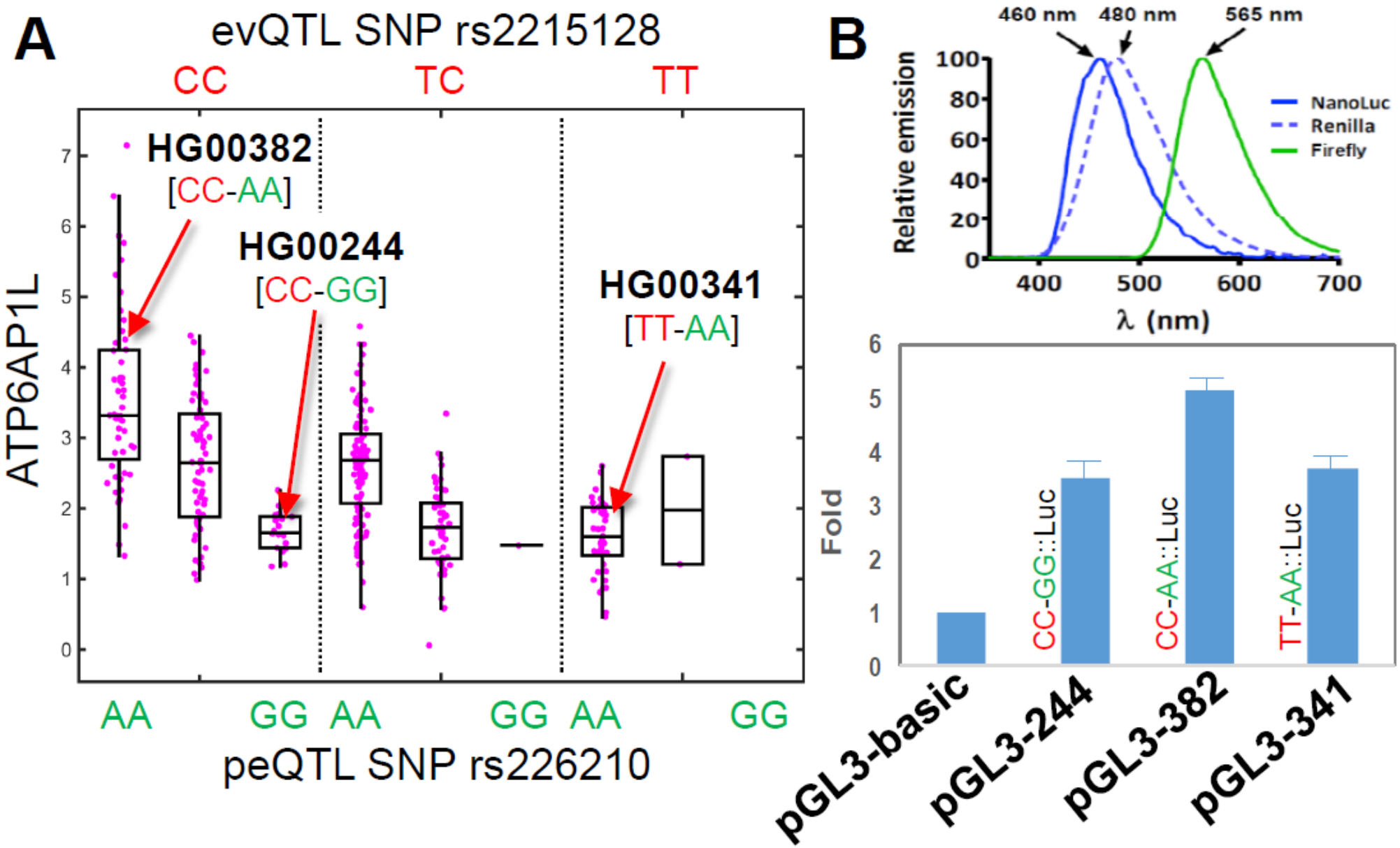
# GxG (epistasis) model



# GxG (epistasis) model



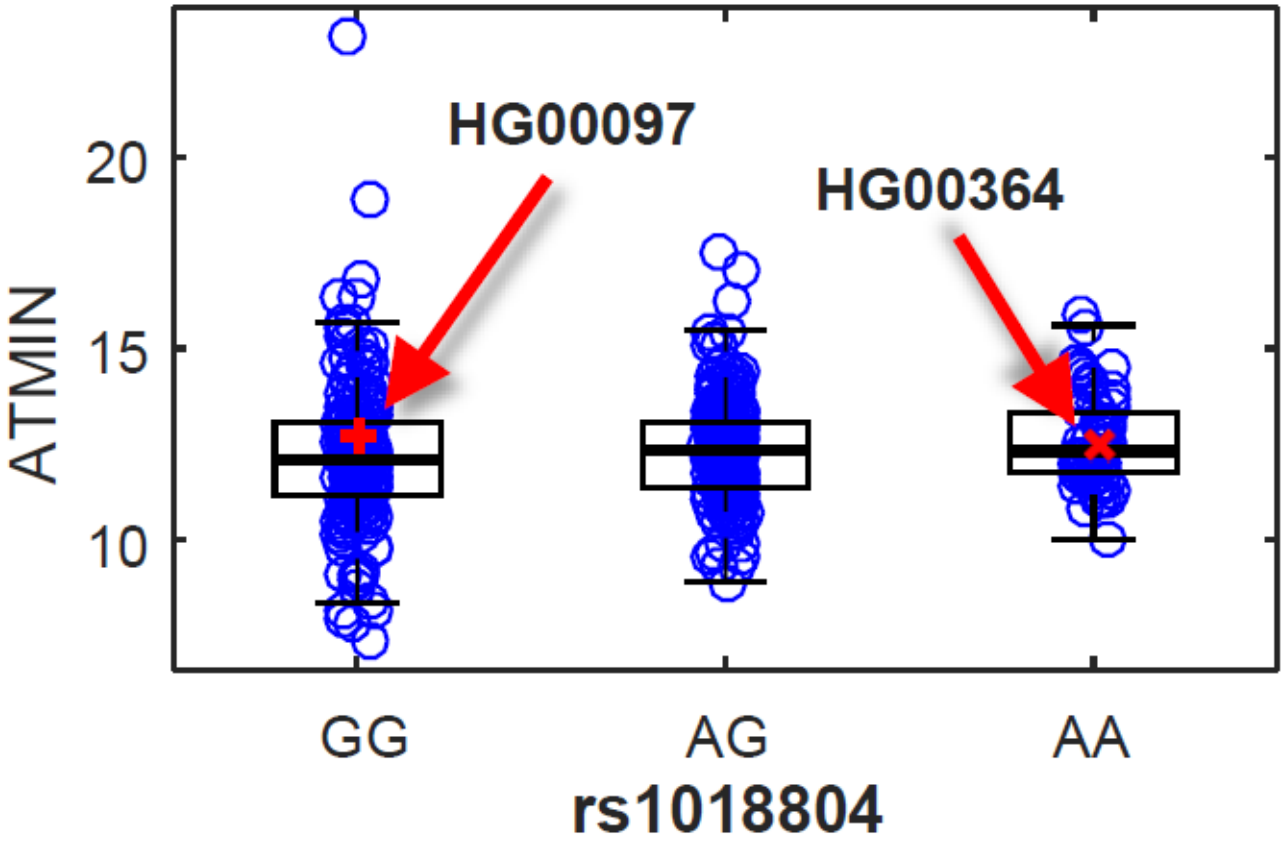

Wang *et al*. (Cai) *Genetics* 2014



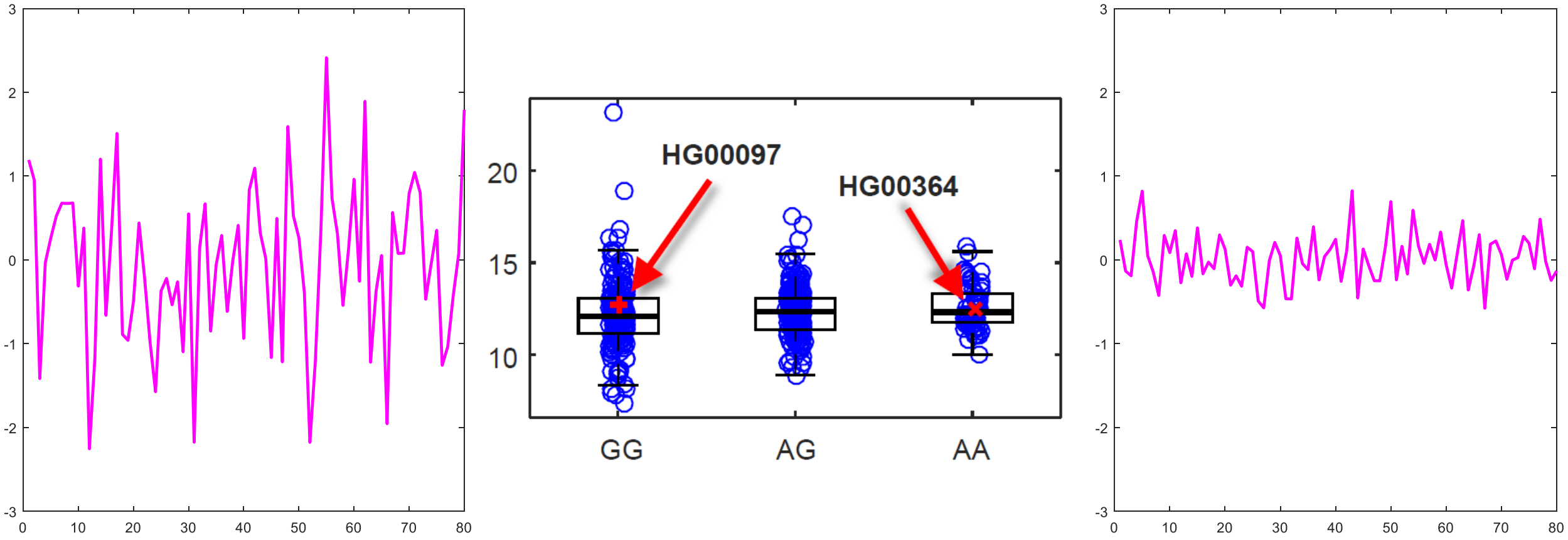

Unpublished

# GxE (destabilization) model – repetitive qPCR

Select **two cell lines** from groups with large and small expression variability.



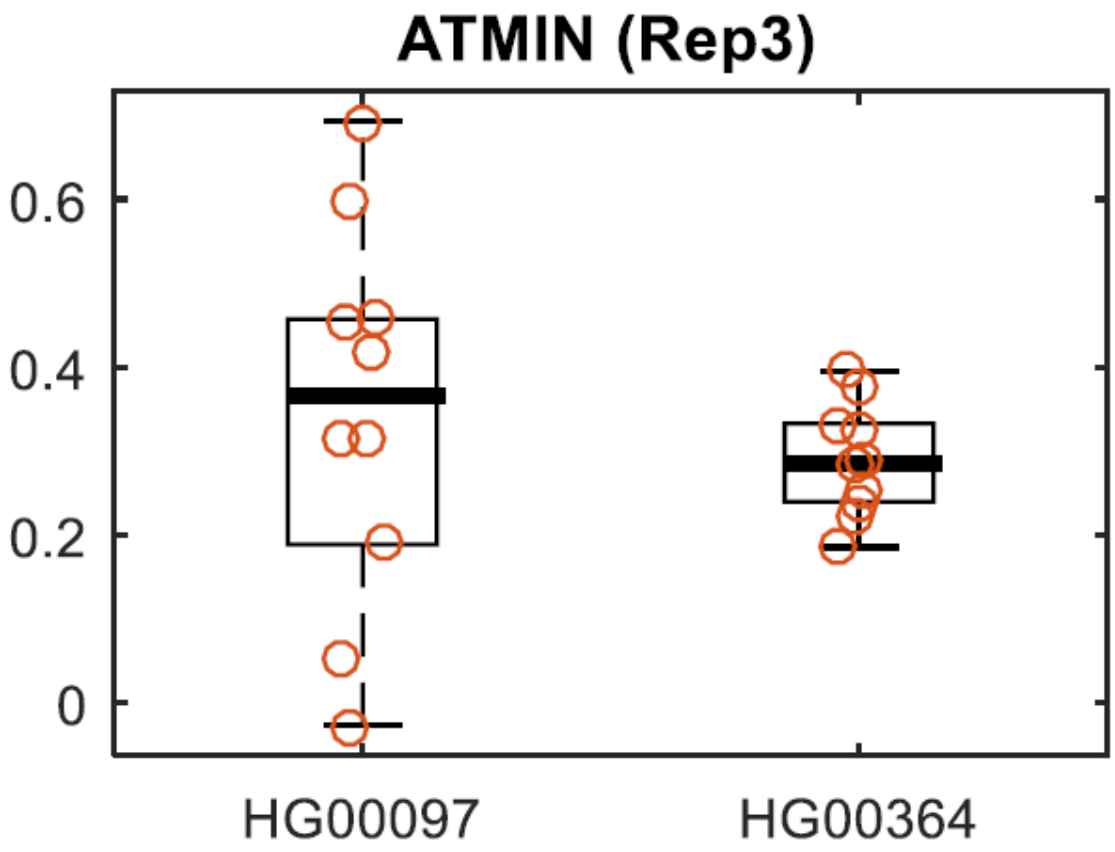

Yang *et al.* (Cai) *Hum Mol Genet* 2016



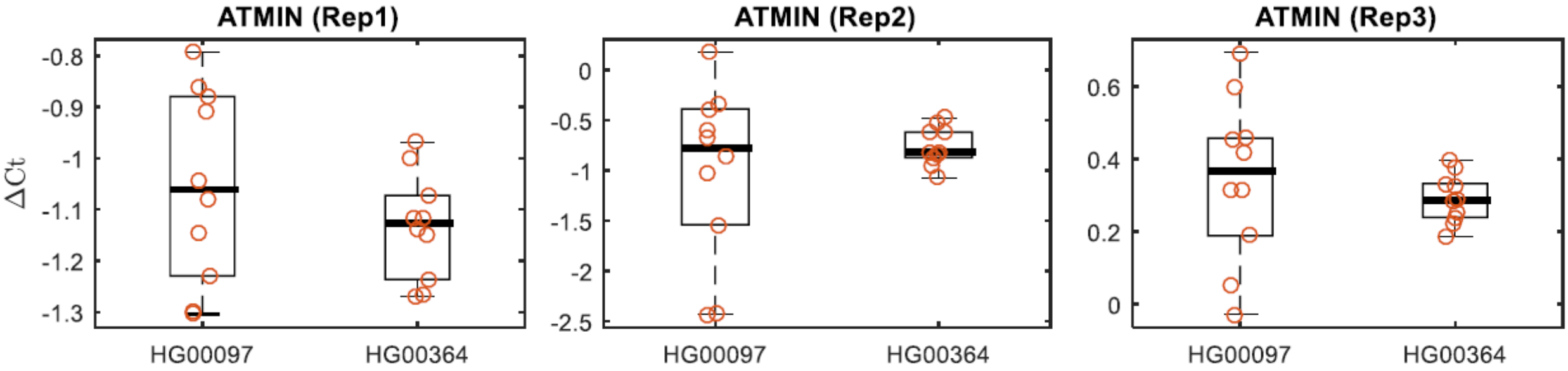

Yang *et al.* (Cai) *Hum Mol Genet* 2016

# GxE (destabilization) model – repetitive qPCR

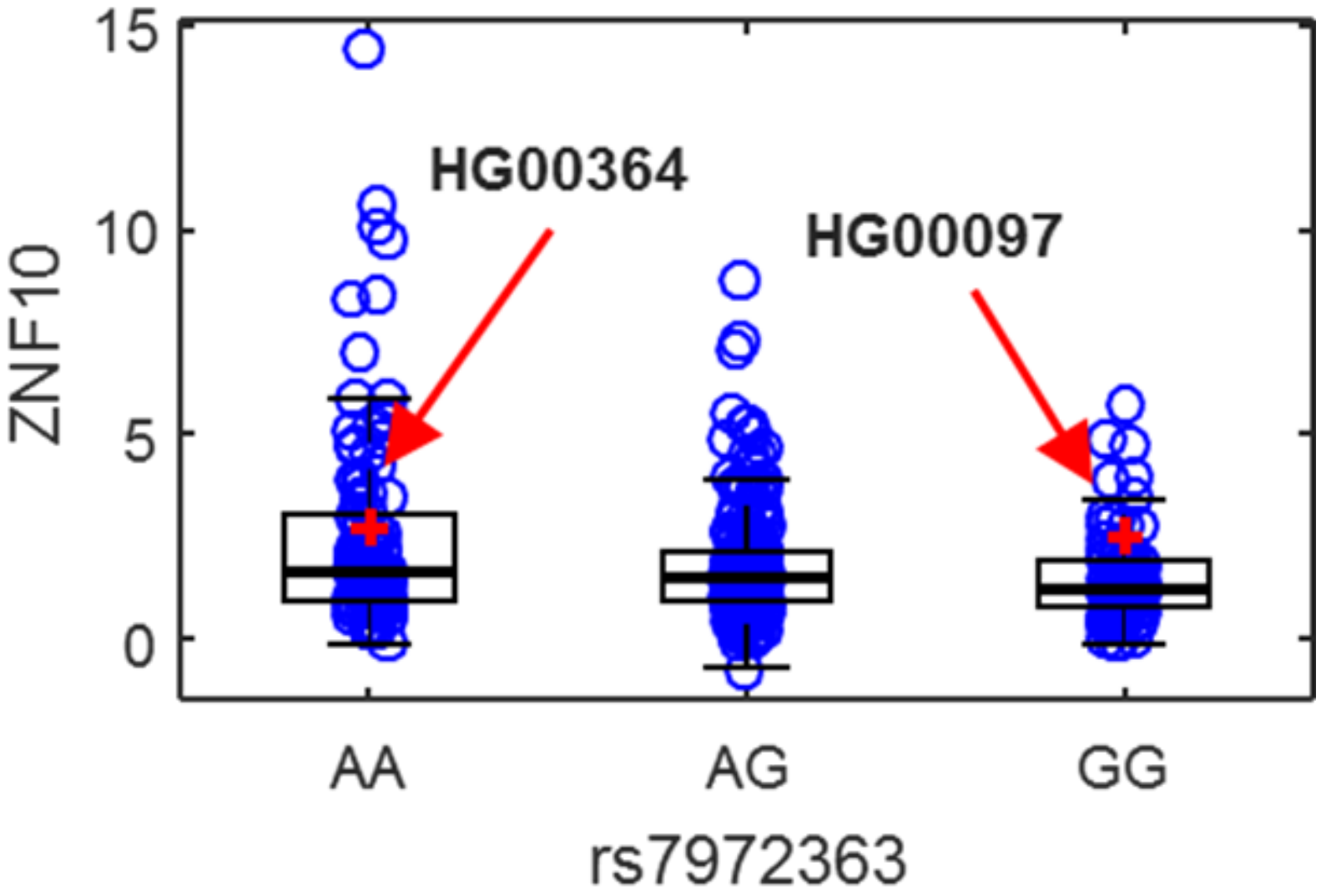
**qRT-PCR** assay was repeated 10 times for each sample.



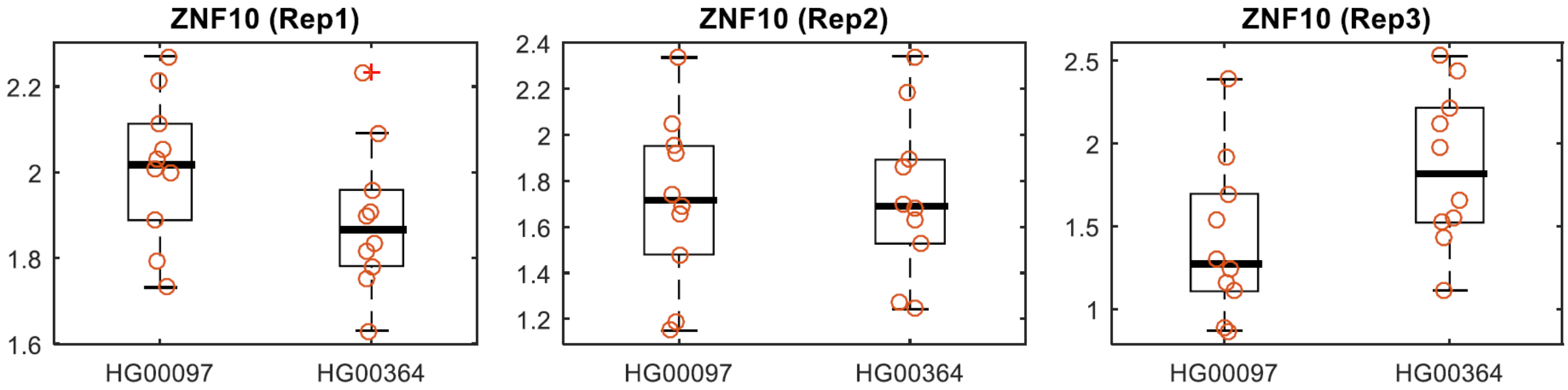
# **GxE** (**destabilization**) model – repetitive qPCR



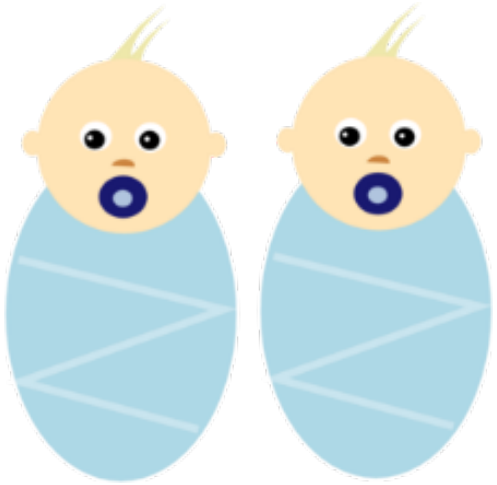
# An evQTL explained by the **GxG** (**epistasis**) model

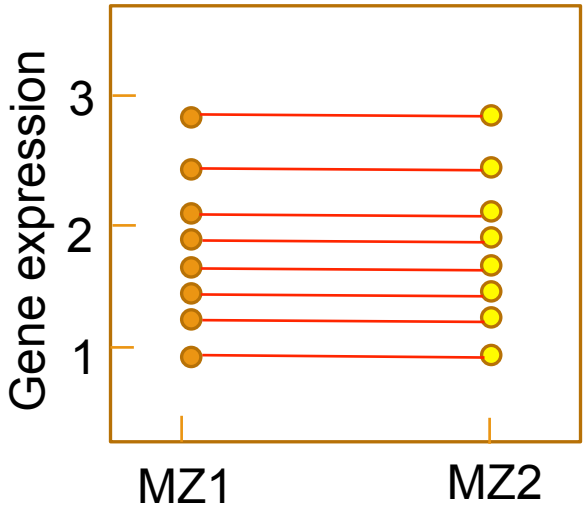


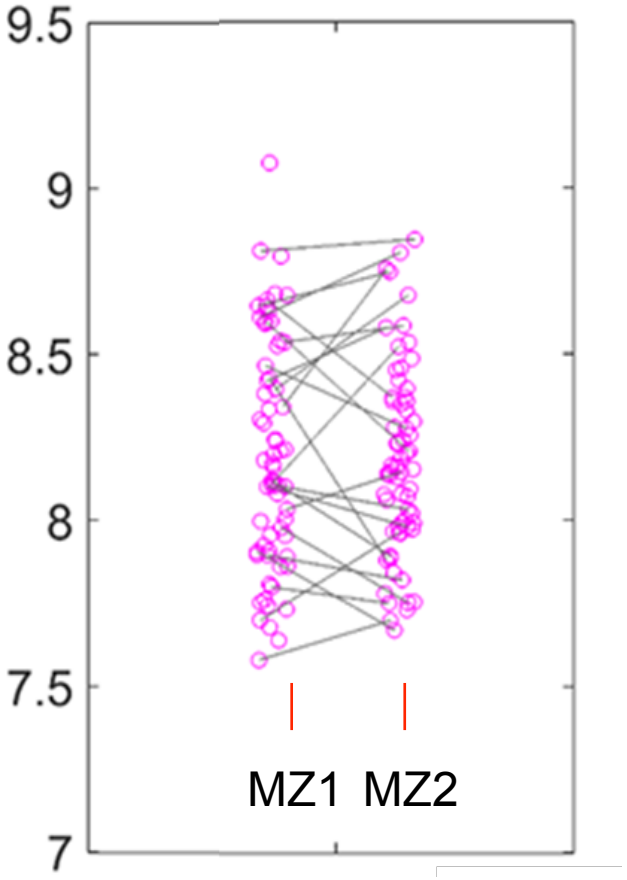
# An evQTL explained by the **GxG** (**epistasis**) model



# **GxE** (**destabilization**) model – discordant expression between monozygotic (MZ) twins

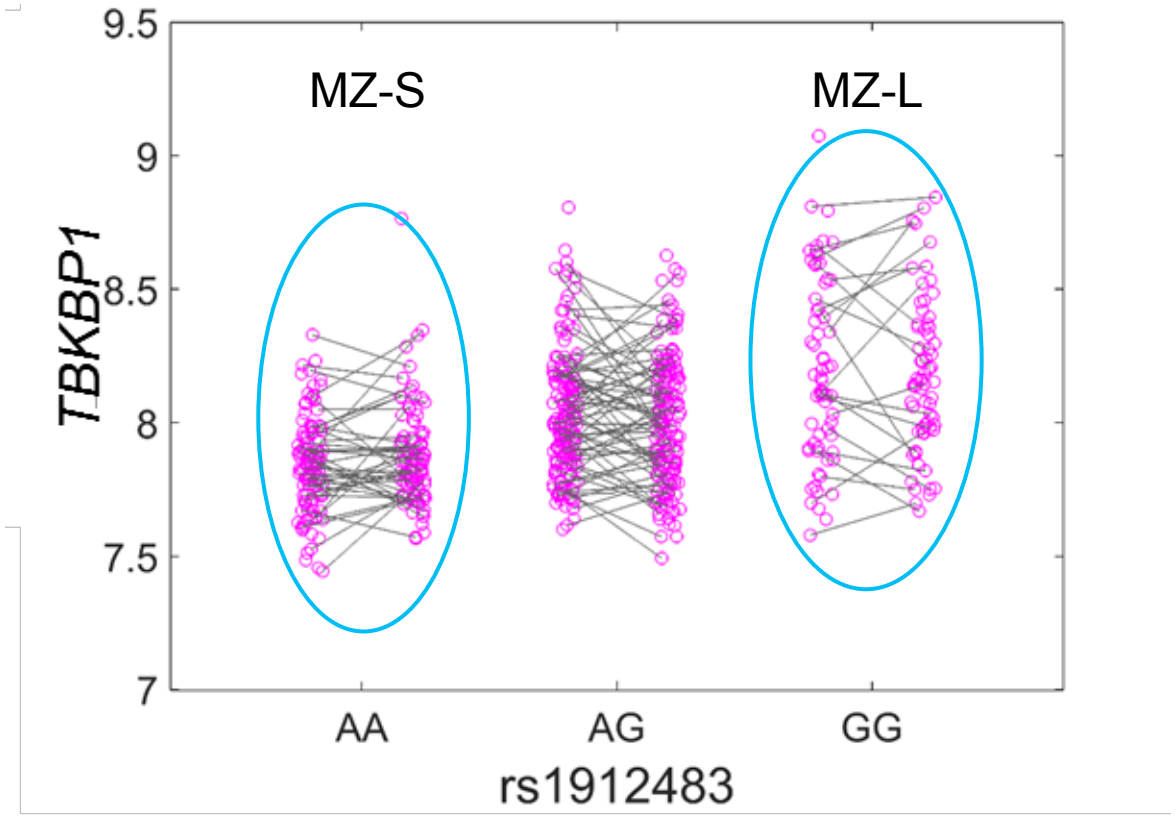


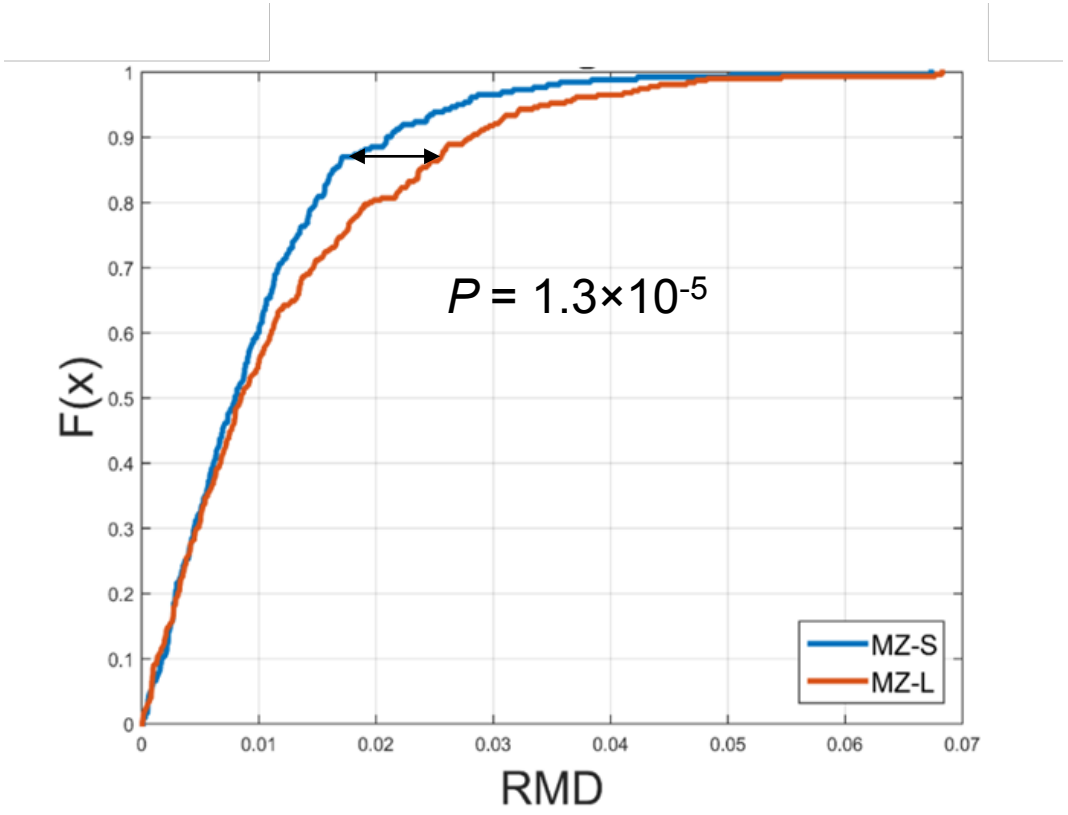



Yang *et al.* (Cai) *Hum Mol Genet* 2016

# GxE (**destabilization**) model – discordant expression between monozygotic (MZ) twins




**Discordant Expression btw MZ Twin Pairs**

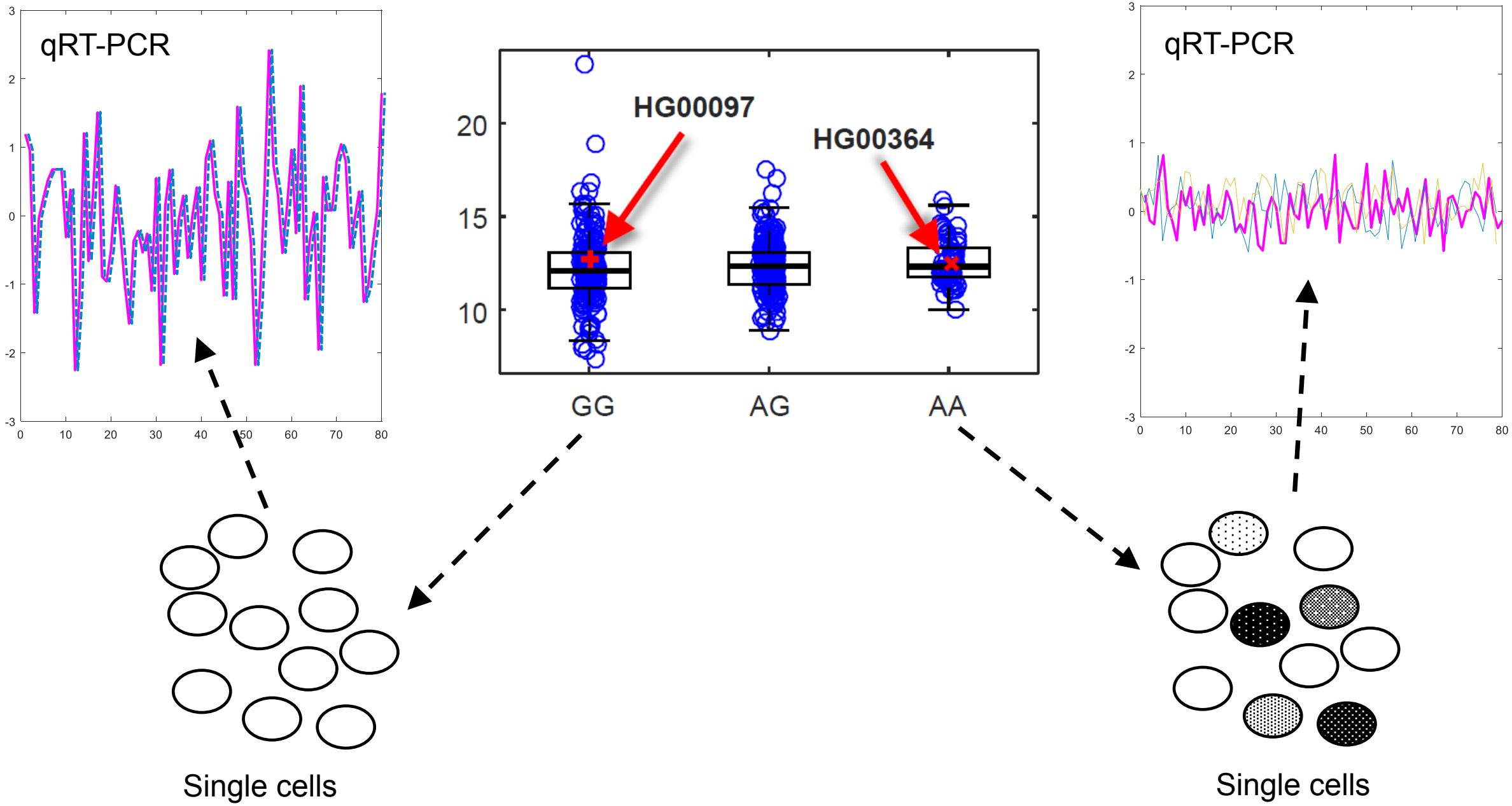
Yang *et al.* (Cai) *Hum Mol Genet* 2016

# Future plans

**Circadian rhythm** gene expression analysis (D. Earnest)

**Single-cell** gene expression analysis (A. Raj)

**CRISPR/Cas9**-based gene editing (D. Segal)

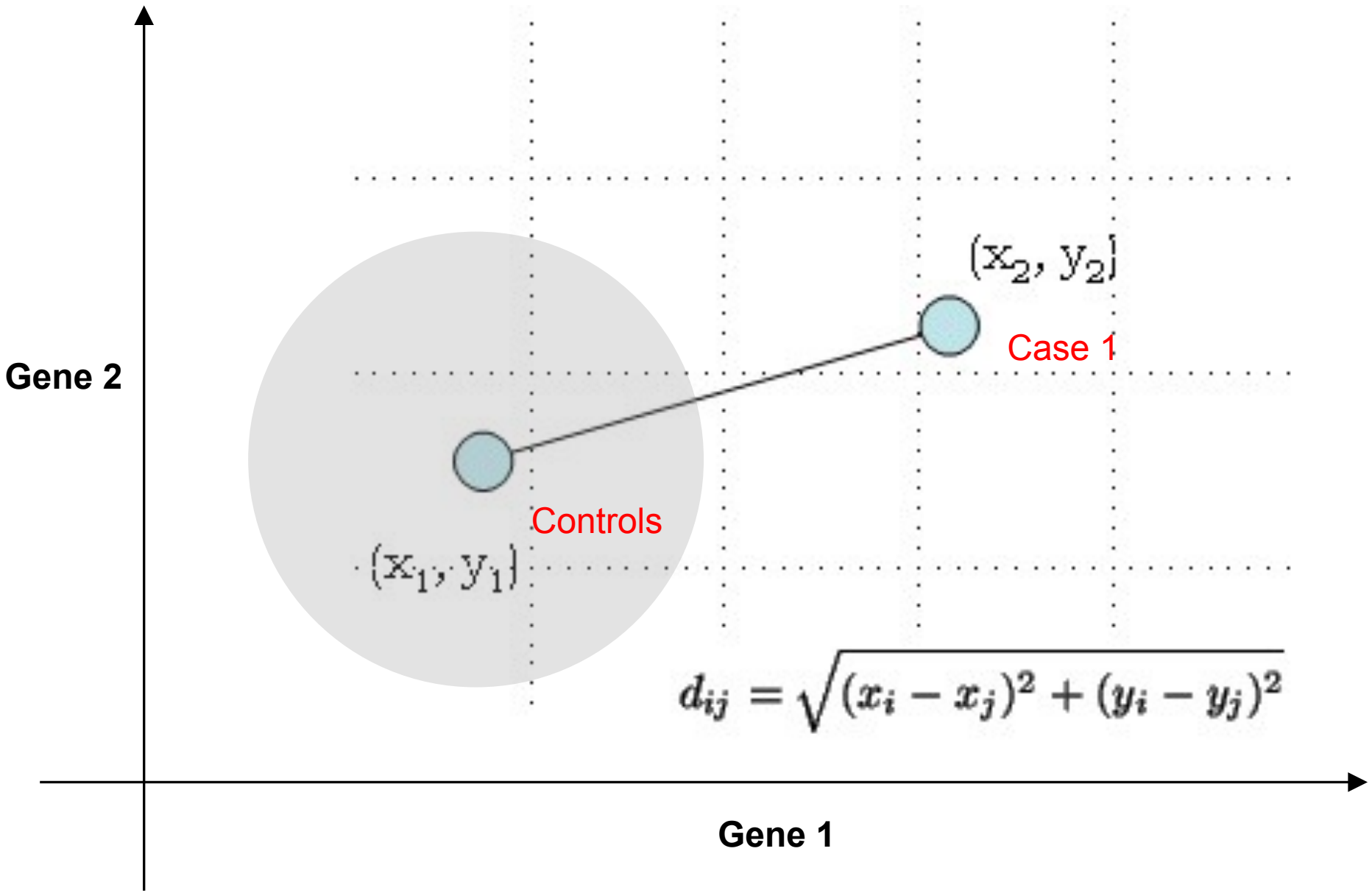

qRT-PCR
HG00097
HG00364
GG
AG
AA
qRT-PCR
Single cells
Single cells

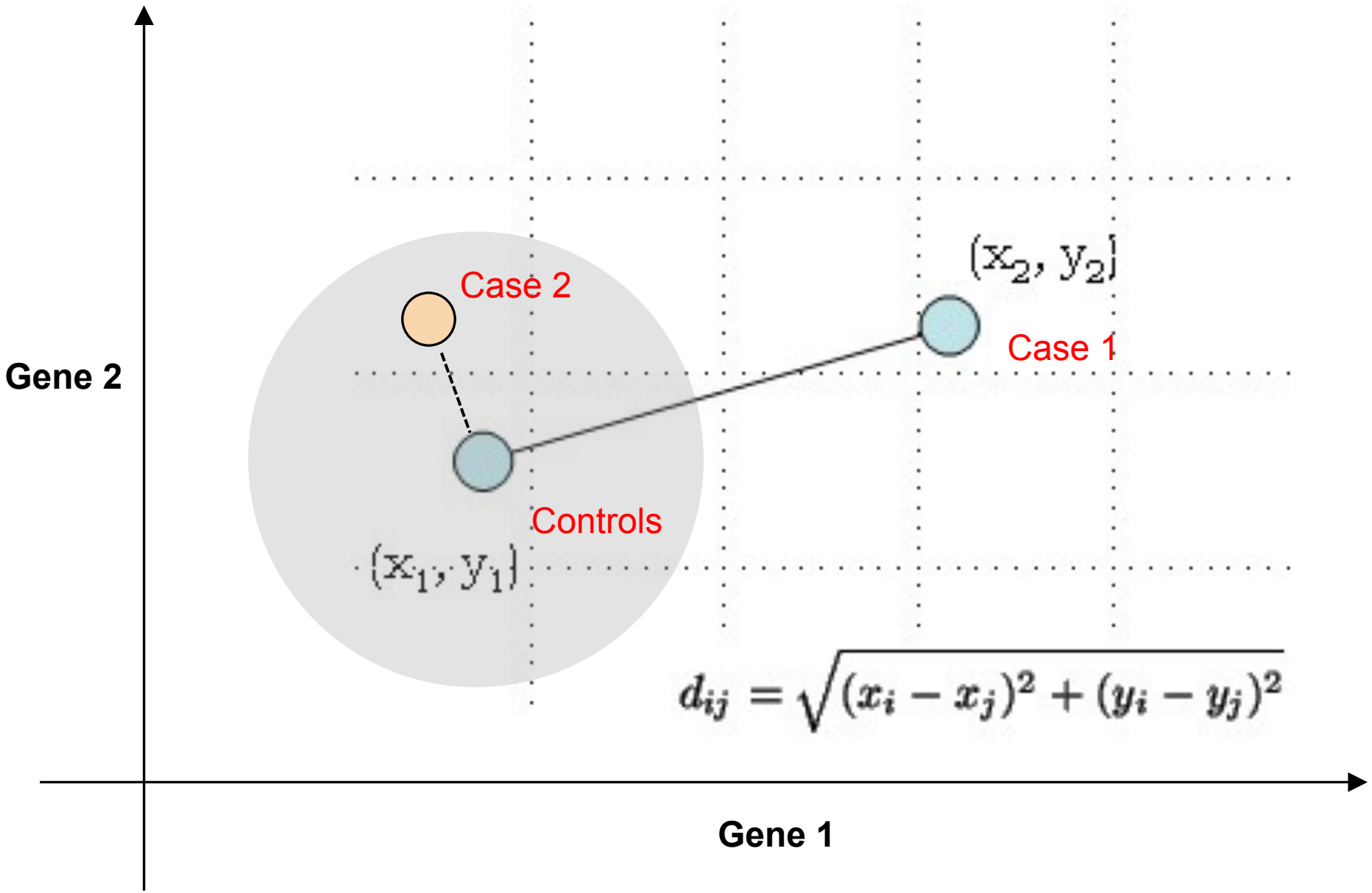
# Summary

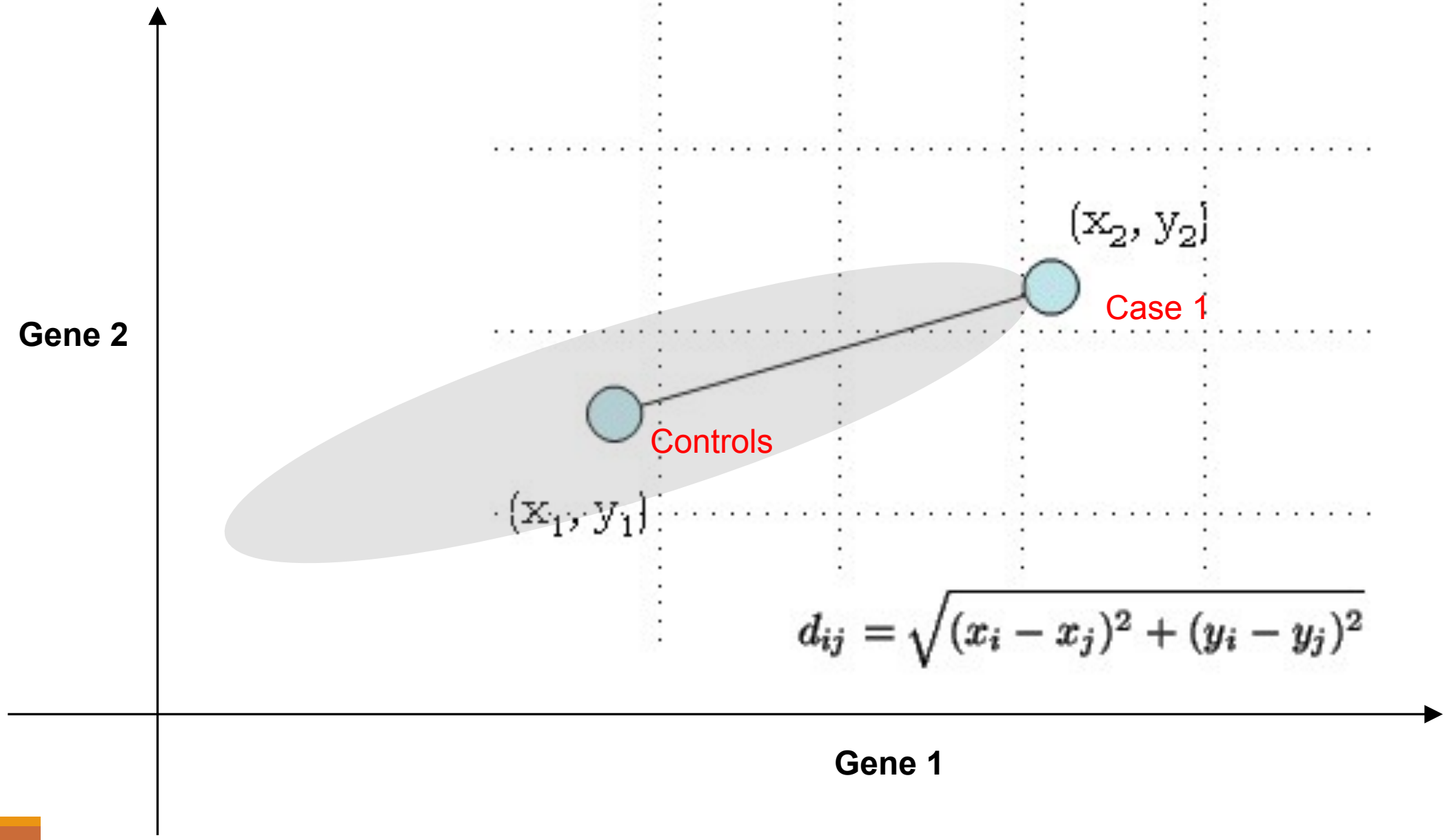
- Two distinct modes of action — **epistasis** and **destabilization**.
  - Genetic variants work either interactively (GxG) or independently (GxE) to influence gene expression variance.

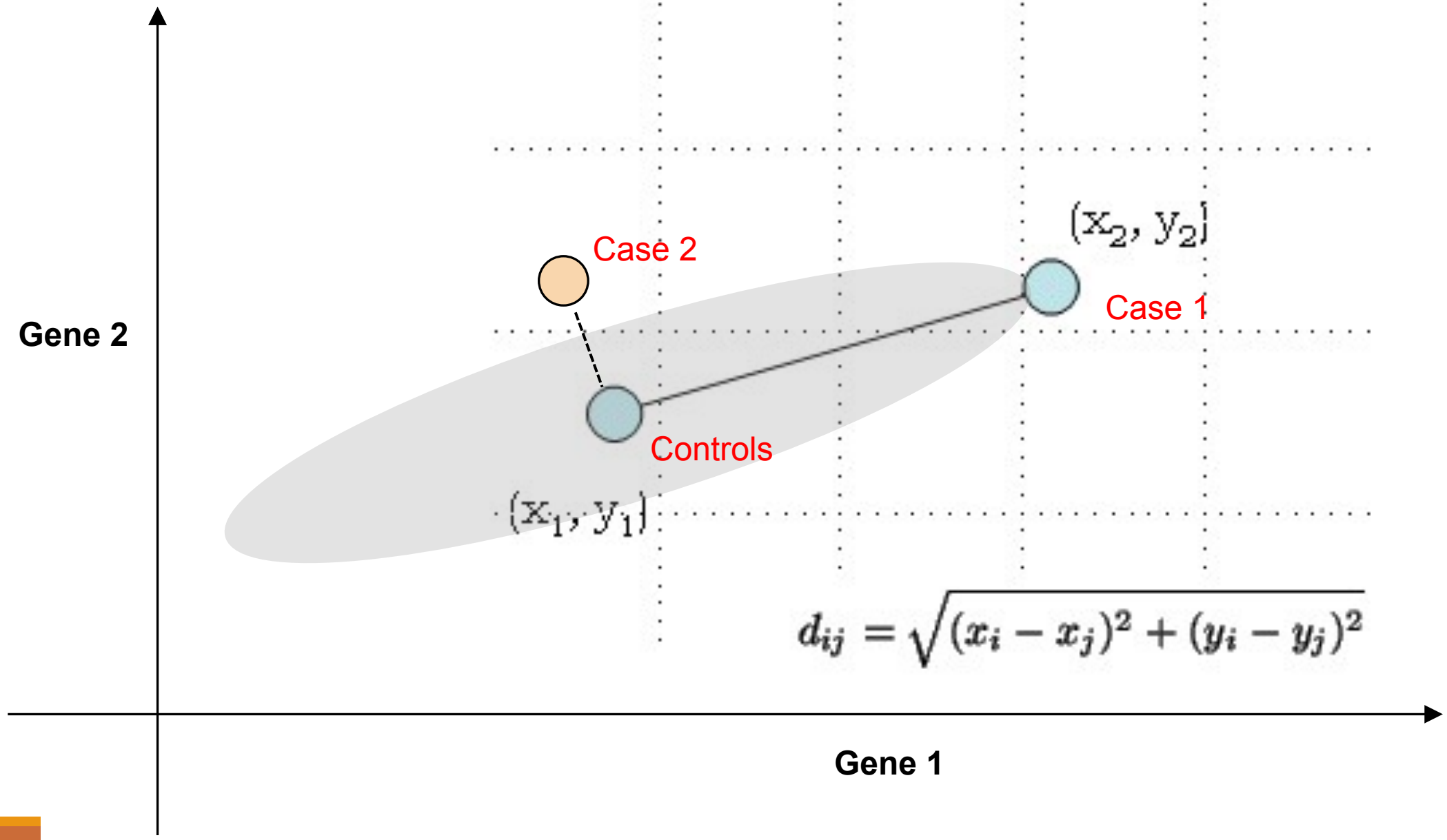
# Exploiting **aberrant gene expression** in autism for discovery and diagnosis

Effect of **rare** genetic variants on gene expression variability


Gene 2
Gene 1
$(x_2, y_2)$
Case 1
Controls
$(x_1, y_1)$
$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$

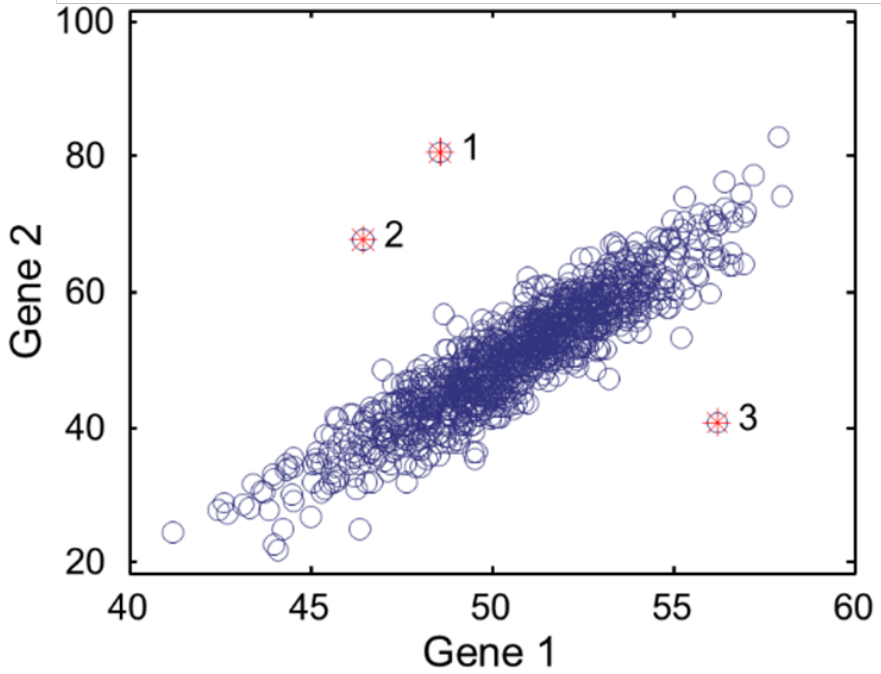

Gene 2
Case 2
$(x_2, y_2)$
Case 1
Controls
$(x_1, y_1)$
$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$
Gene 1

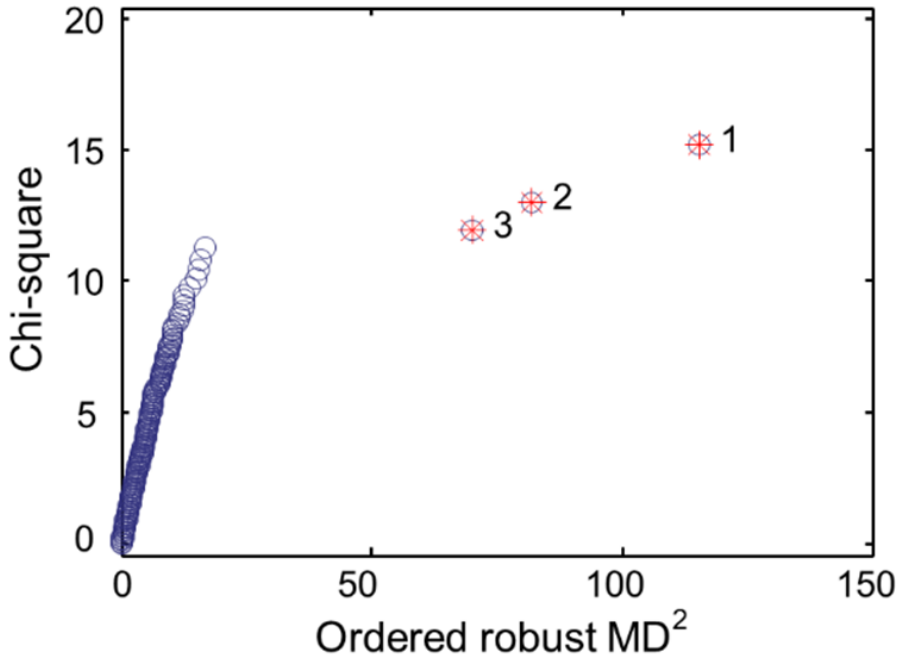

Gene 2
$(x_2, y_2)$
Case 1
Controls
$(x_1, y_1)$
$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$
Gene 1


Gene 2
Case 2
$(x_2, y_2)$
Case 1
Controls
$(x_1, y_1)$
$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$
Gene 1

# **Mahalanobis distance (MD)** is used to detect outliers

1893 – 1972

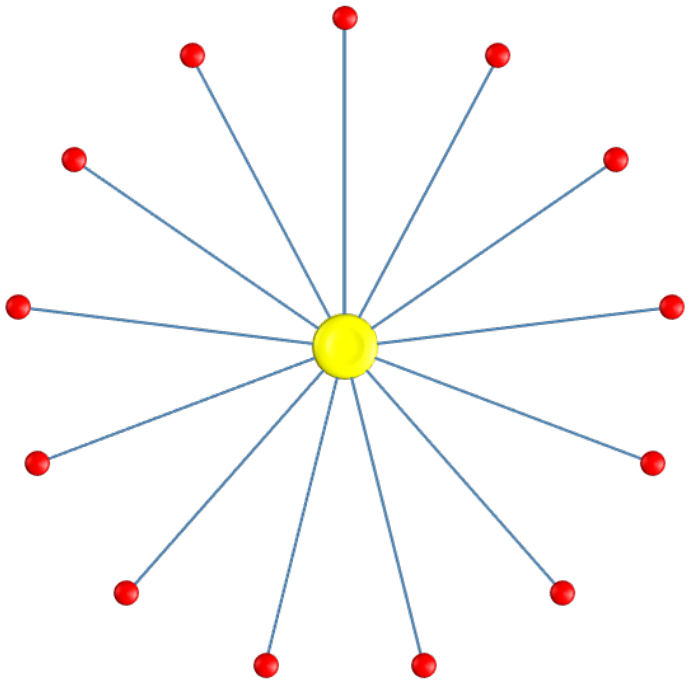



$$MD_i = \sqrt{(E_{i\cdot} - \vec{\mu})^T Cov^{-1} (E_{i\cdot} - \vec{\mu})}$$

# MD measures the level **gene expression dispersion** for a population

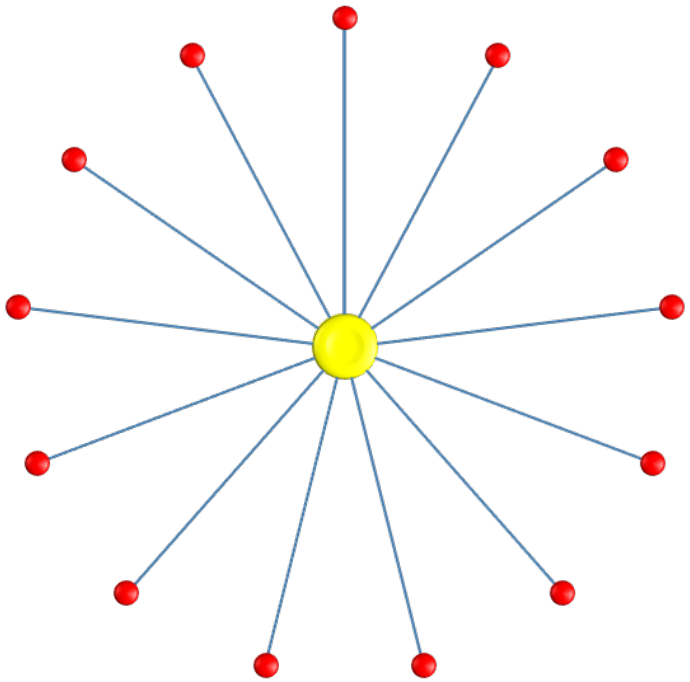
GENE SET 1



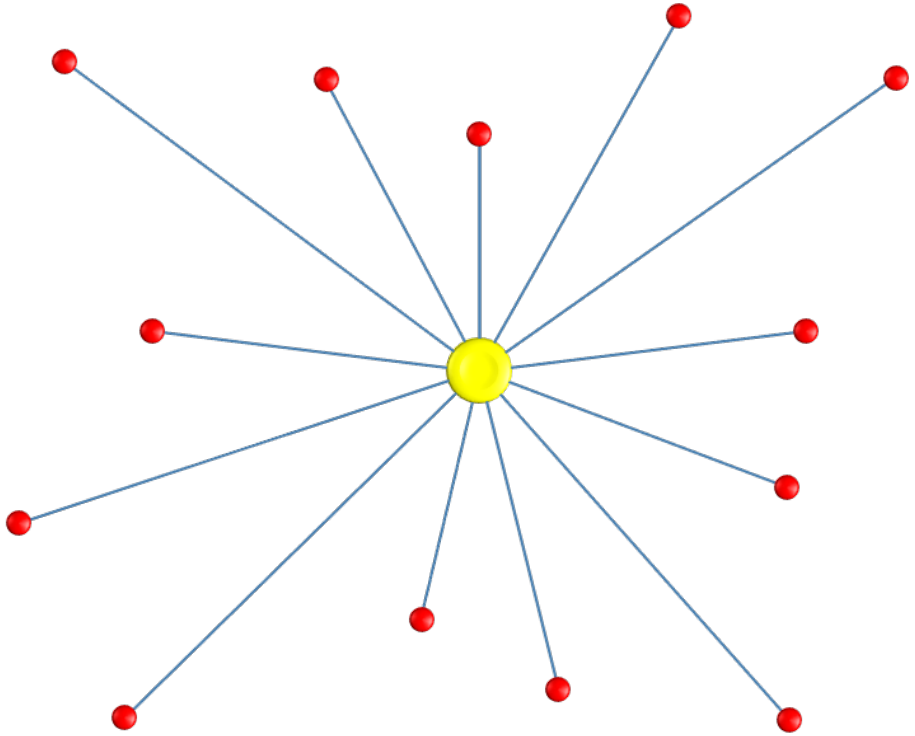
Zeng *et al.* (Cai) *PLoS Genet* 2015

# MD measures the level gene **expression dispersion** for a population

GENE SET 1



GENE SET 2



Zeng *et al.* (Cai) *PLoS Genet* 2015

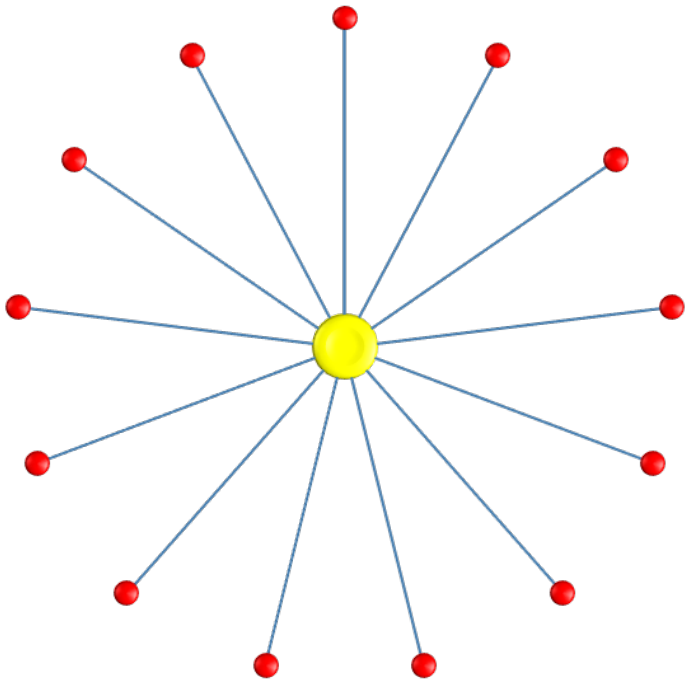
# Sum of squared MD (**SSMD**) – **Overall dispersion** level of a **gene set**

$$SSMD = \sum_{i=1}^{M} MD_i^2$$

# SSMD – overall dispersion level of a gene set

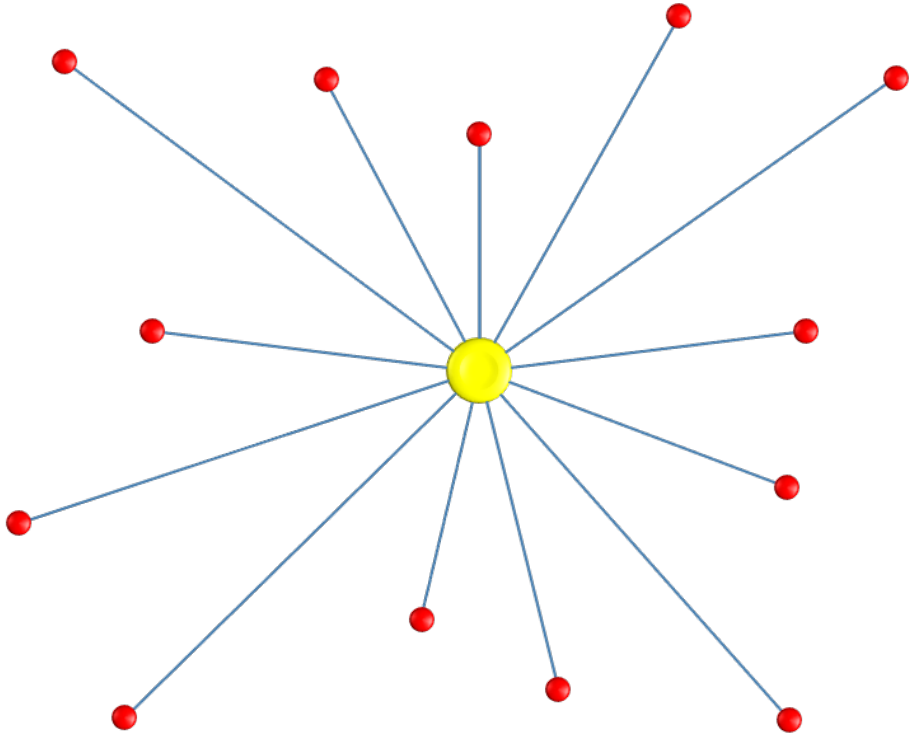
GENE SET 1

SSMD ↓↓



GENE SET 2

SSMD ↑↑



# Gene sets (L-SSMD) that tend to be aberrantly expressed

MSigDB: molecular signatures database from the Broad Institute
31 gene sets

- **G-protein coupled receptor activity**
- **Transmission of nerve impulse**
- **Ligand-gated ion channel transportation**
- **Cyclic guanosine monophosphate (cGMP) effects**

**Regulation of cellular processes and modulation of signal transduction**

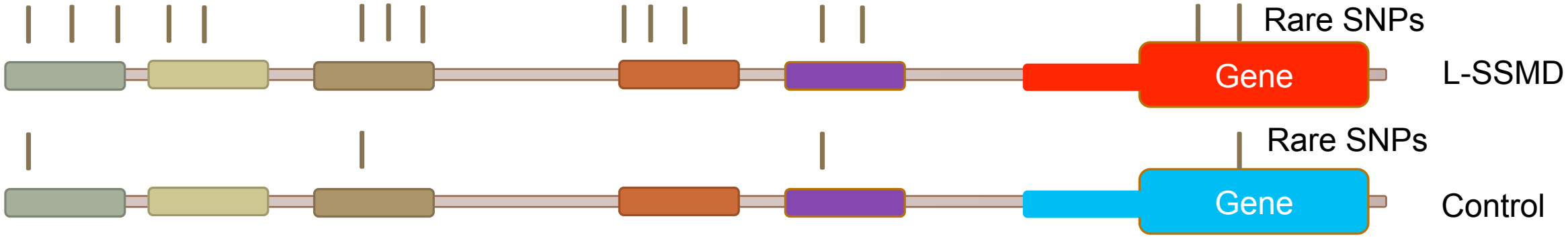
# Gene sets (S-SSMD) that tend not to be aberrantly expressed

MSigDB: molecular signatures database from the Broad Institute
13 gene sets

- **Homologous recombination repair of replication-independent double-strand breaks**
- **Transfer of a phosphate group to a carbohydrate substrate**
- **Cell cycle control**

**Fundamental molecular functions and metabolic pathways**

# SNP density in regulatory regions of L-SSMD genes in outlier individuals

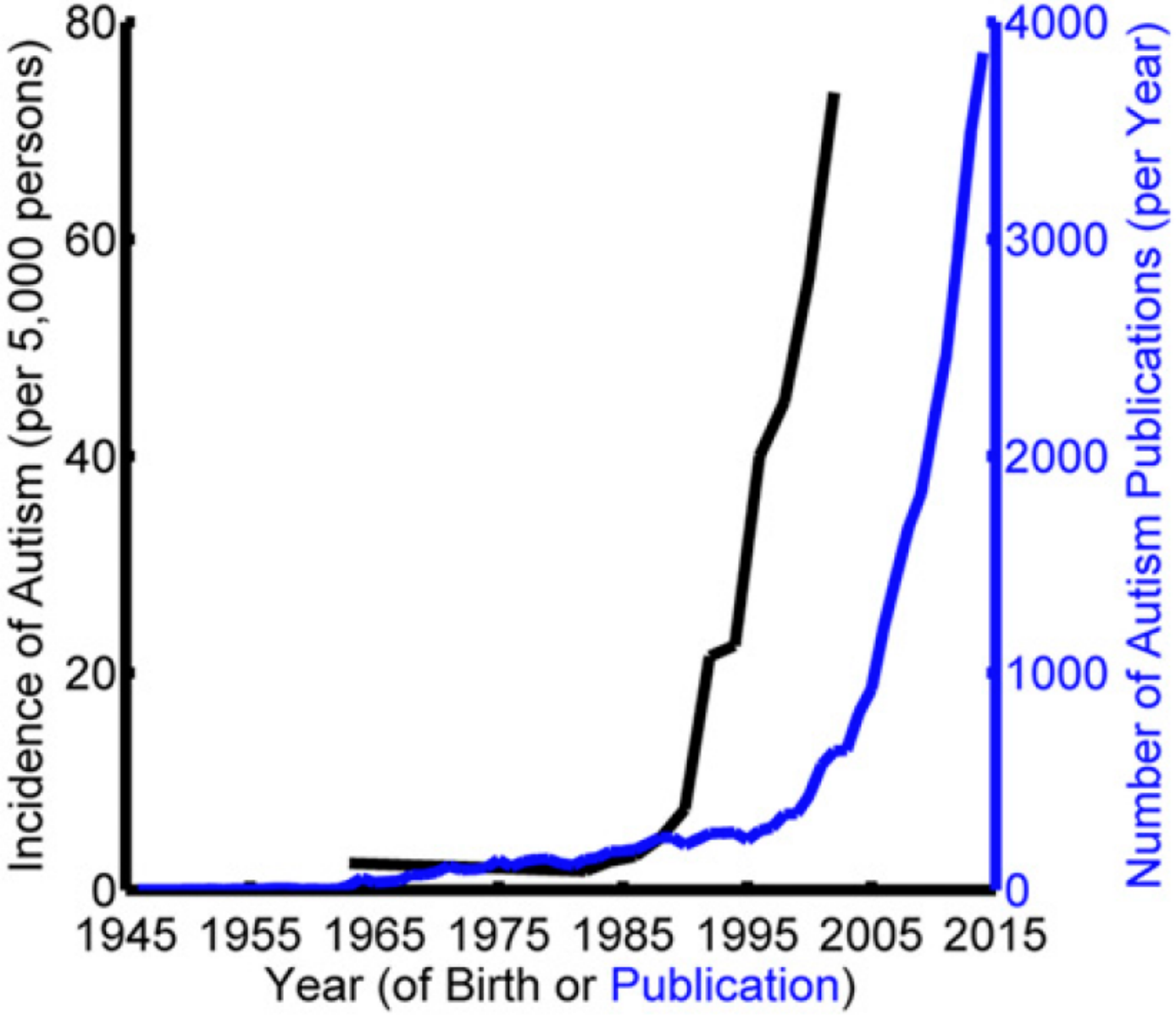


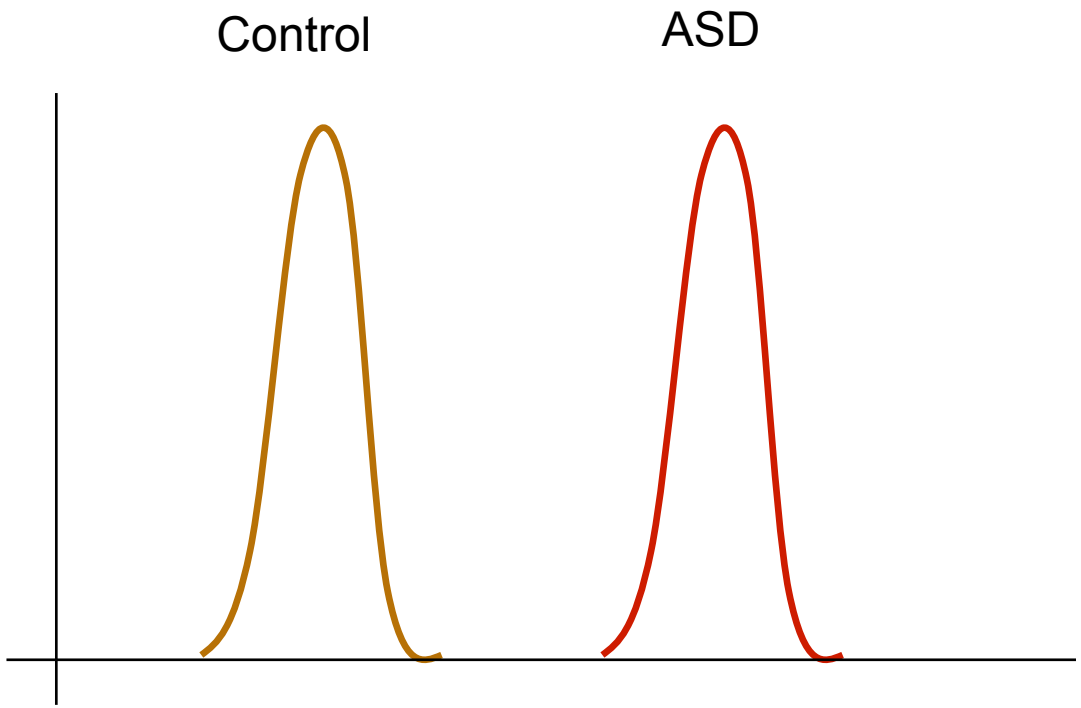

ENCODE regulatory regions

- E: enhancer
- TSS: transcription start site
- T: transcribed region
- PF: predicted promoter flanking region
- CTCT: CTCF-enriched element
- R: repressed or low-activity region
- WE: weak enhancer or open chromatin cis-regulatory element

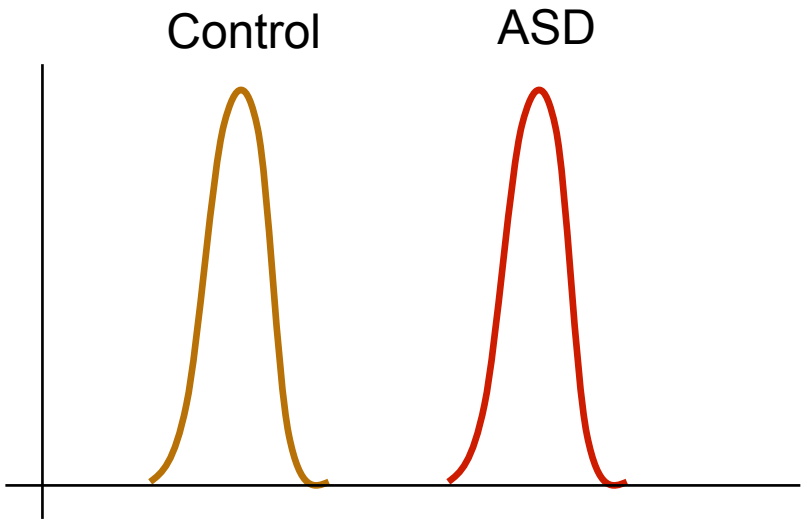
Zeng *et al.* (Cai) *PLoS Genet* 2015

# Autism Spectrum Disorder (ASD)

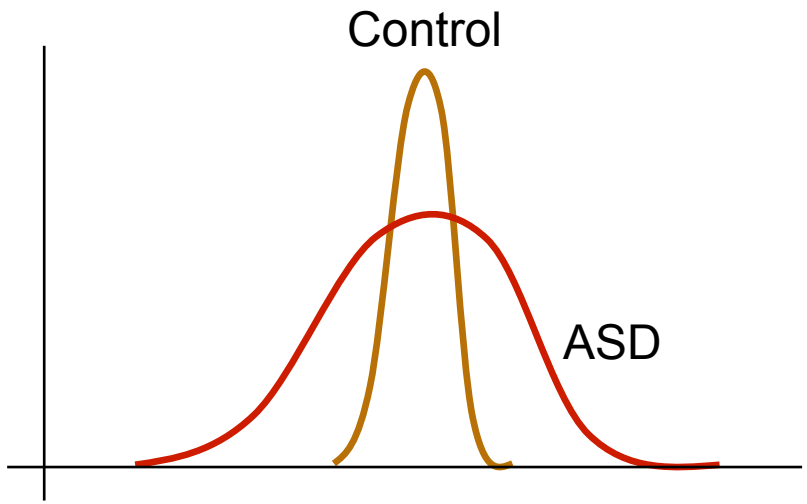





# DE




**DE**

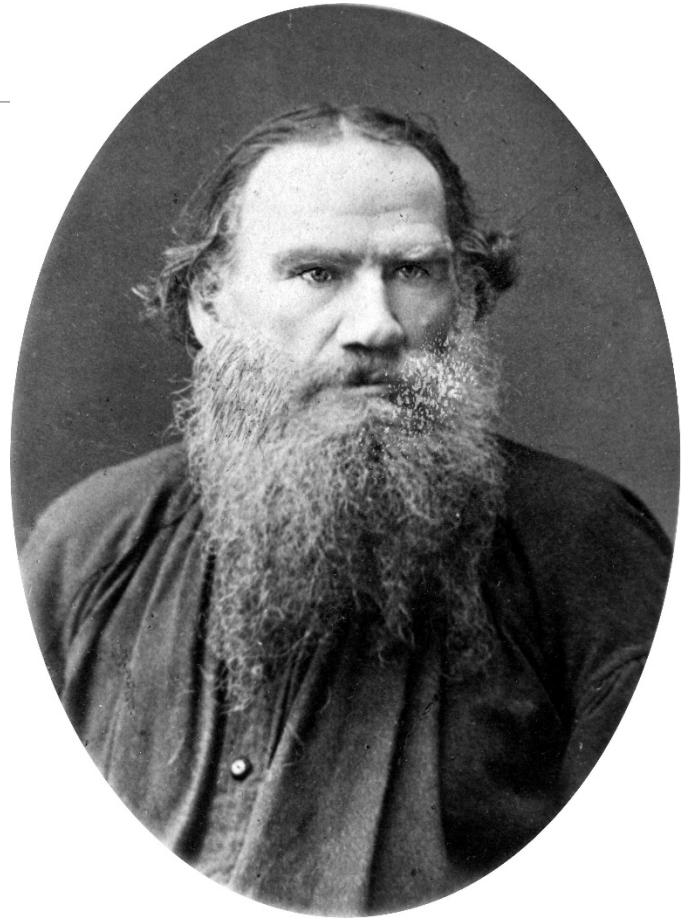



**DV**

# Anna Karenina Principle

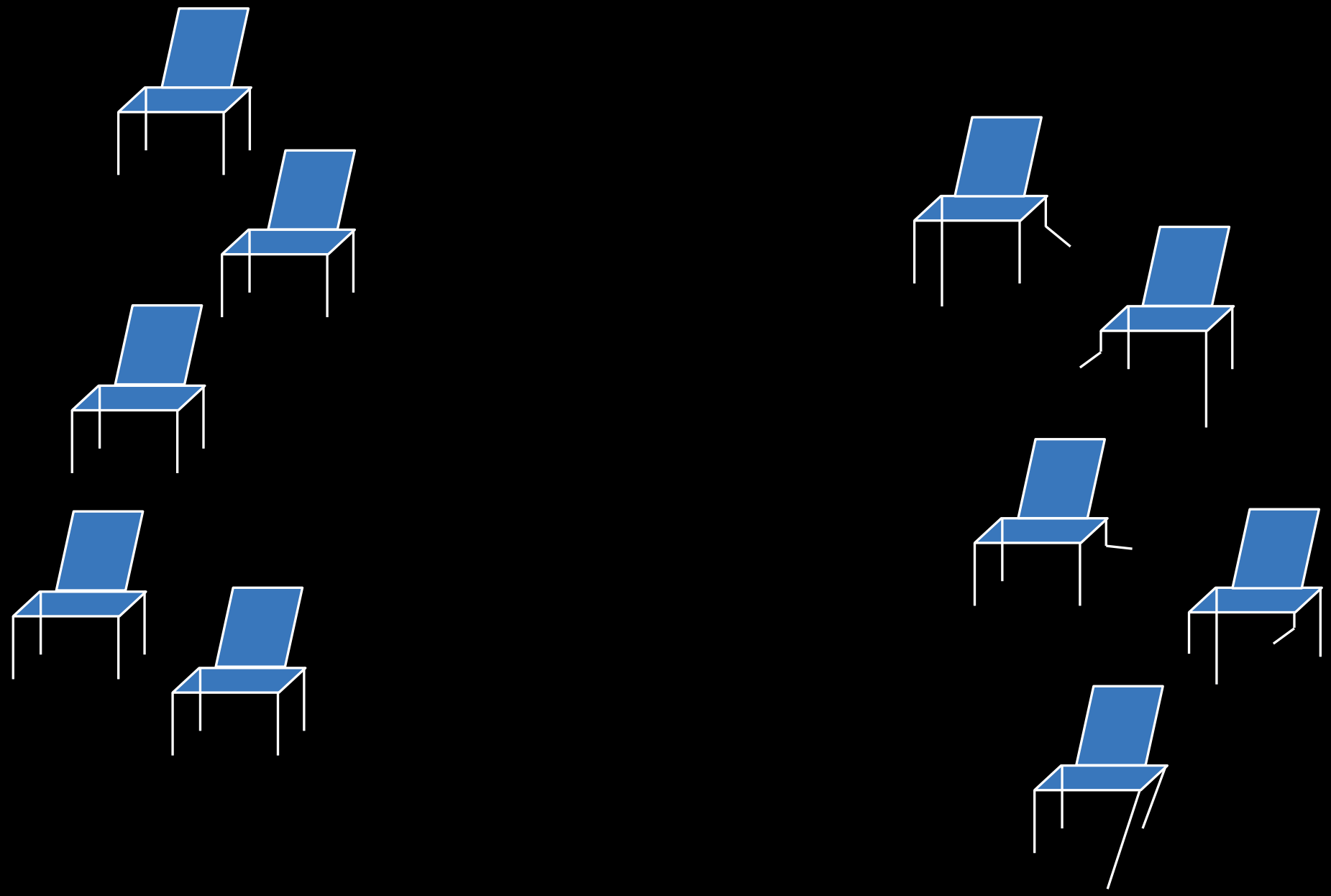
*"Happy families are all alike; every unhappy family is unhappy in its own way."*

***All healthy people are alike; each sick person is sick in his or her own way.***



Leo Tolstoy
1828 – 1910

# Chair Model



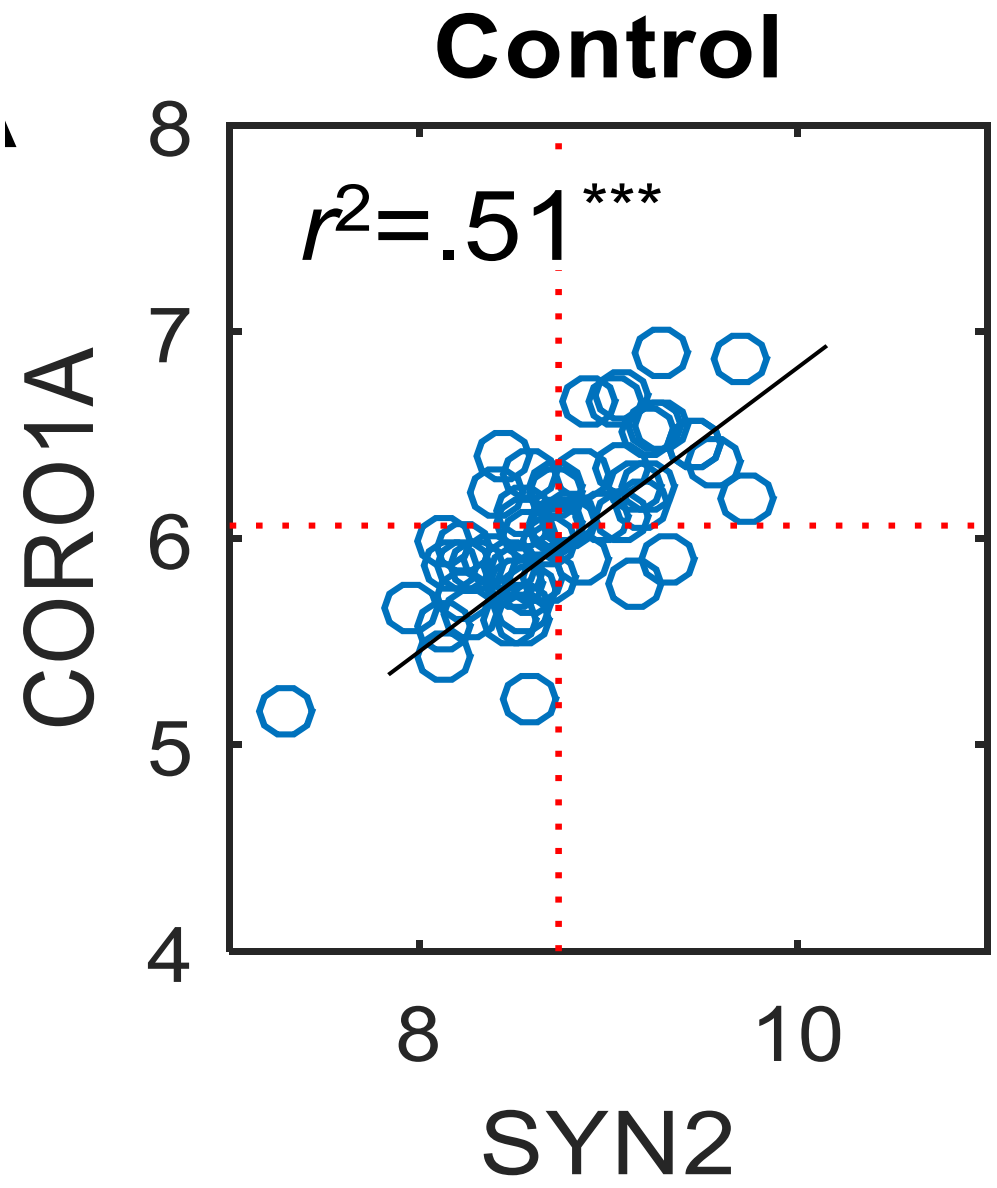
Brain RNA-seq:

- 47 ASD
- 57 controls

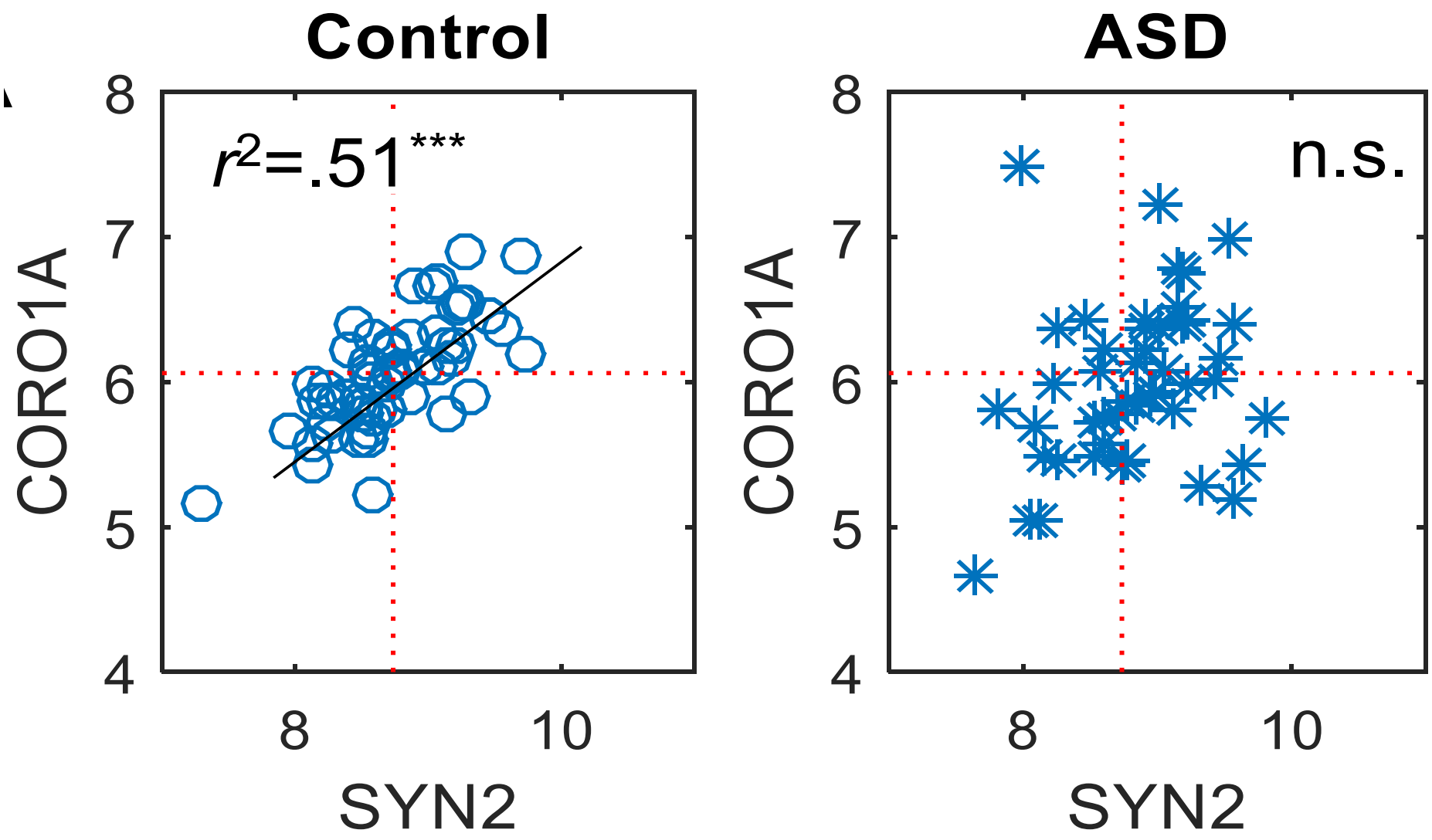
Gupta et al. (2014) *Nat Commun* 5:5748.

Coronin 1A facilitates formation of heterotrimeric or multiprotein complexes.

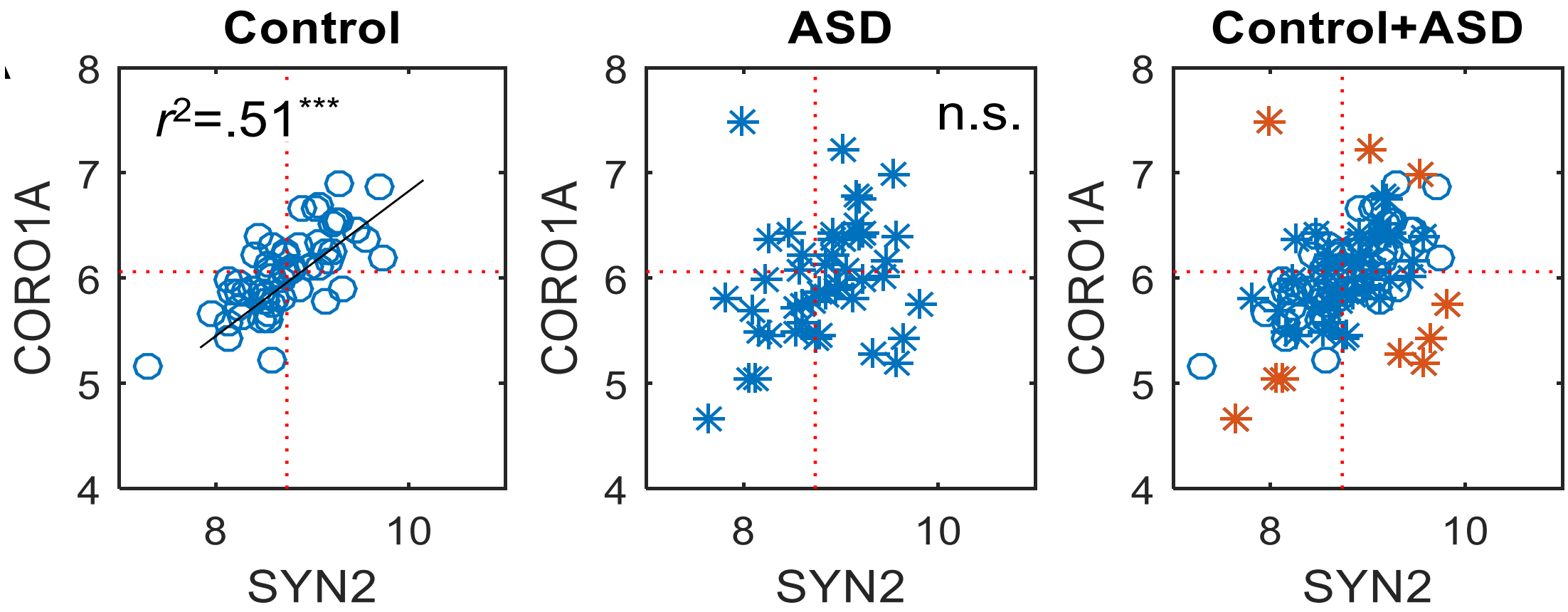
Synapsin II encodes neuronal phosphoprotein associated with the cytoplasmic surface of synaptic vesicles.



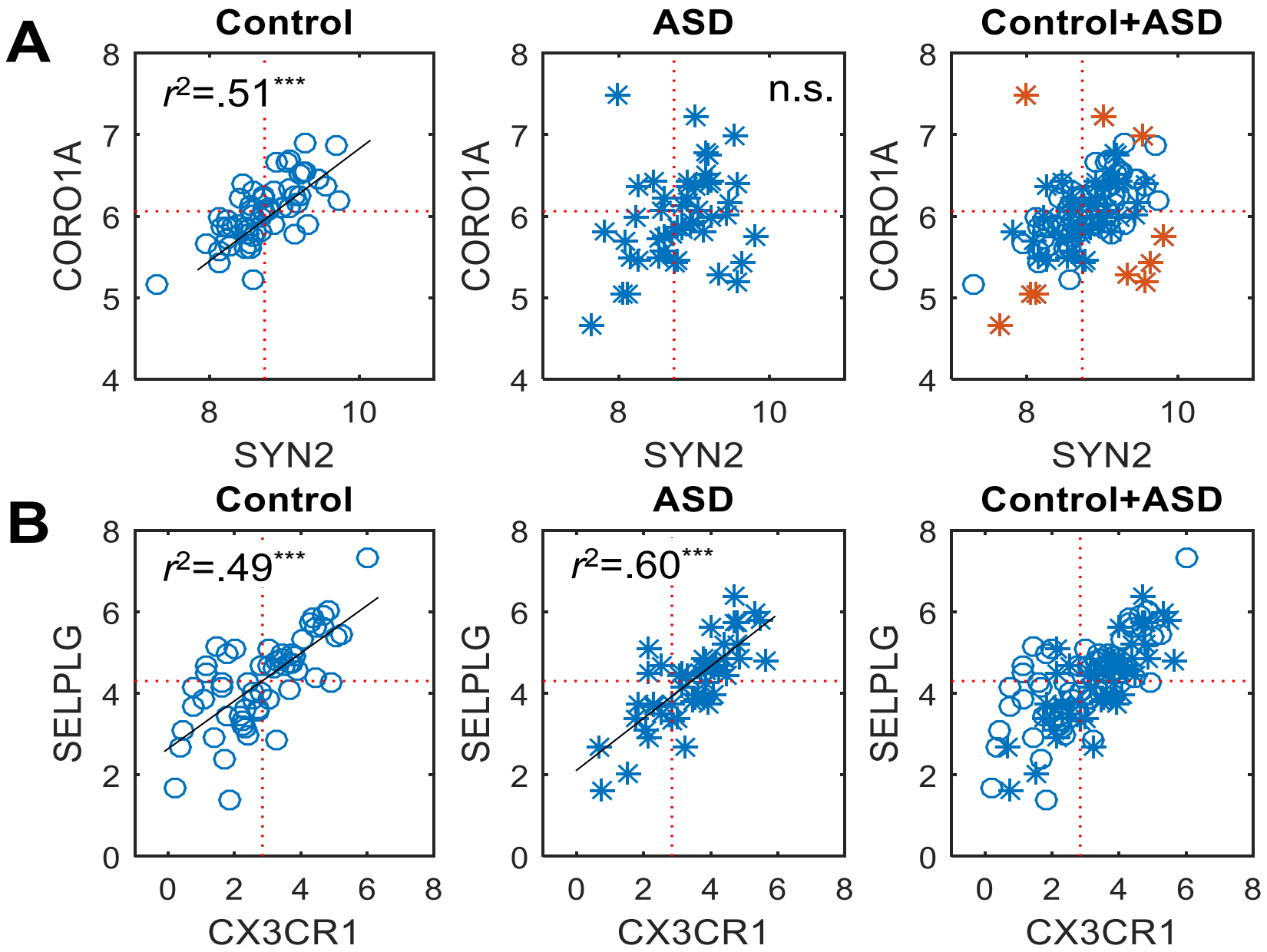

Guan *et al.* (Cai) *Hum Genet* 2016



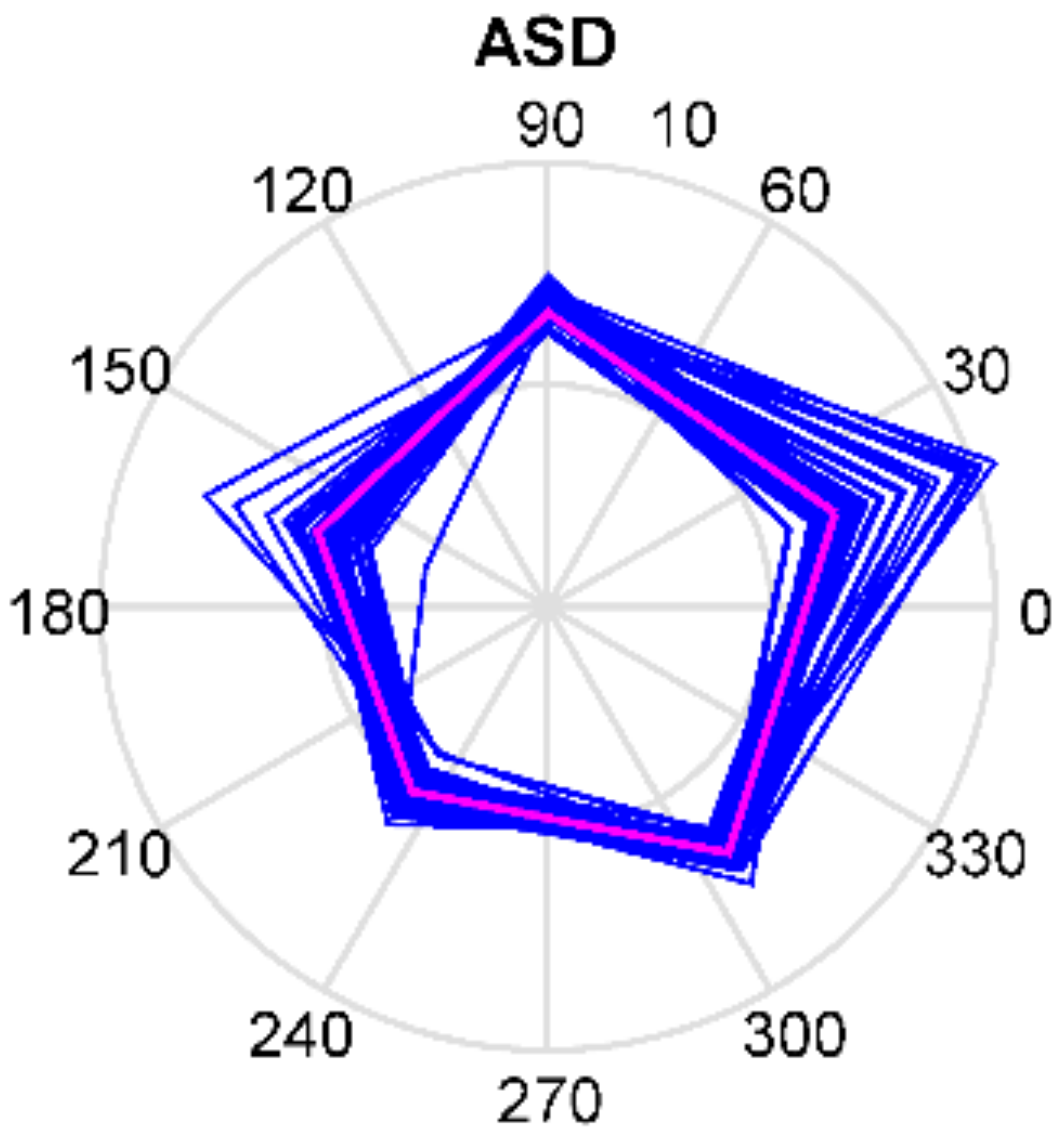

Guan *et al.* (Cai) *Hum Genet* 2016

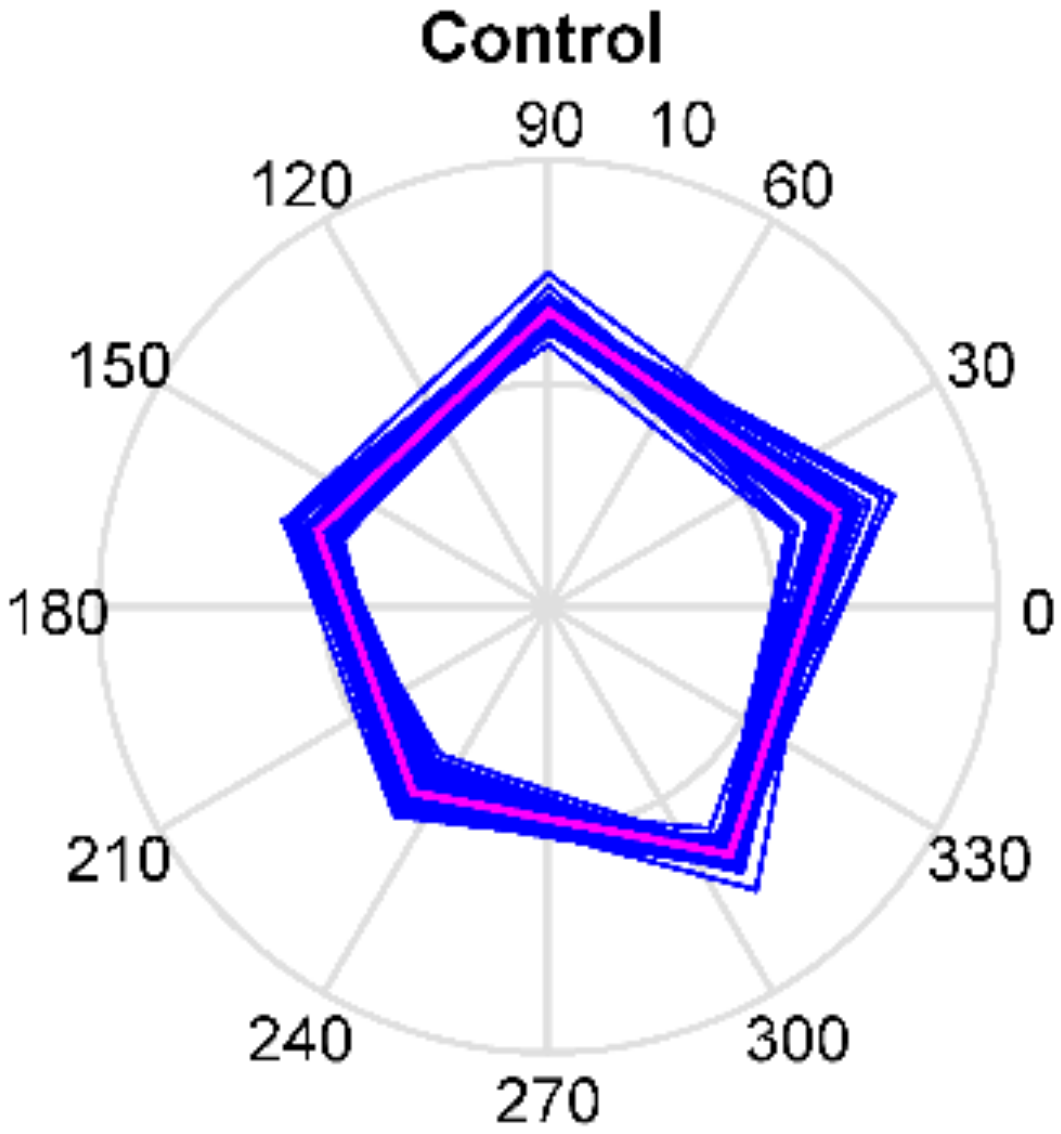



Guan *et al.* (Cai) *Hum Genet* 2016



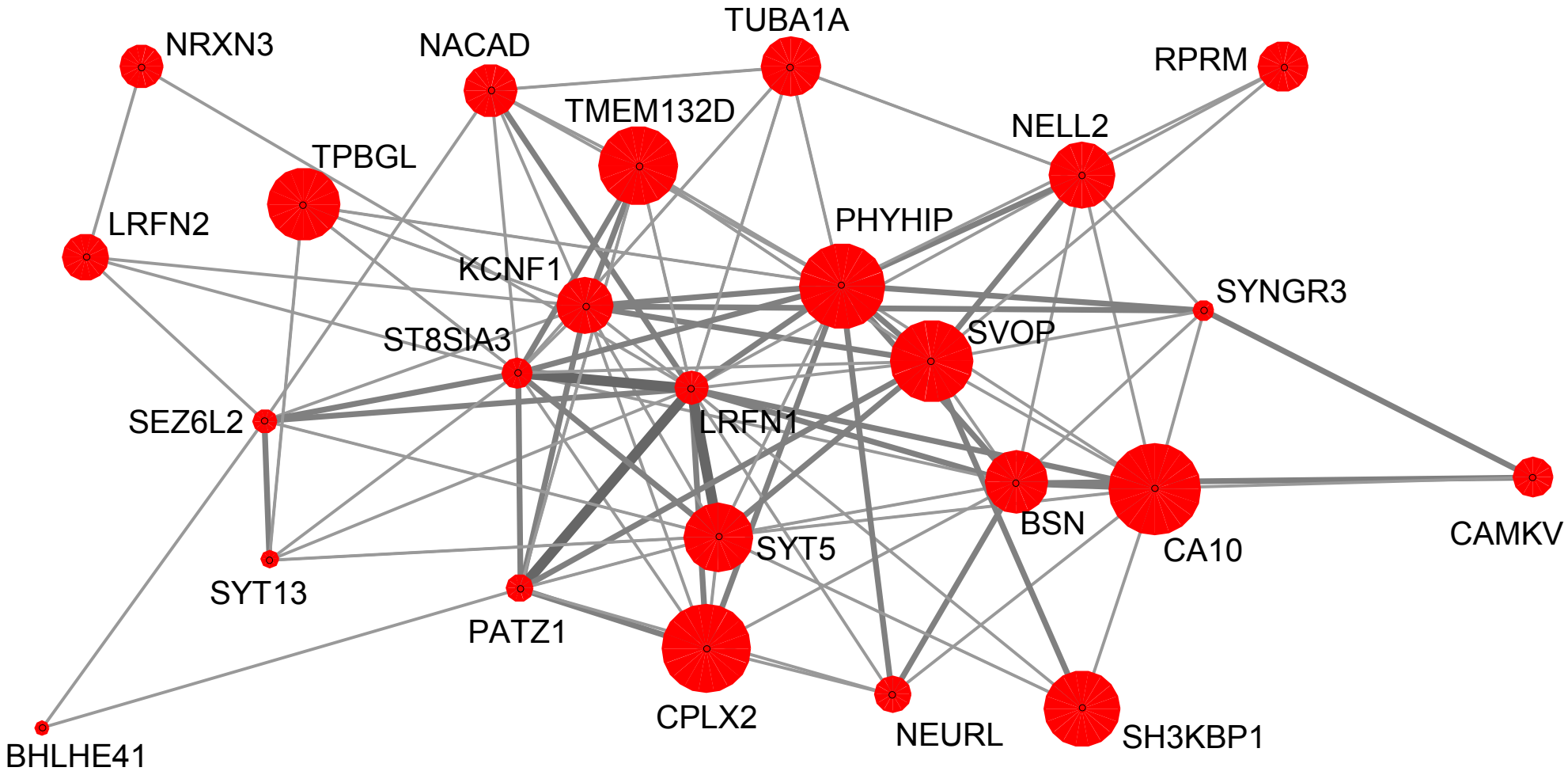

Guan *et al.* (Cai) *Hum Genet* 2016


ASD
90
10
60
30
0
330
300
270
240
210
180
150
120

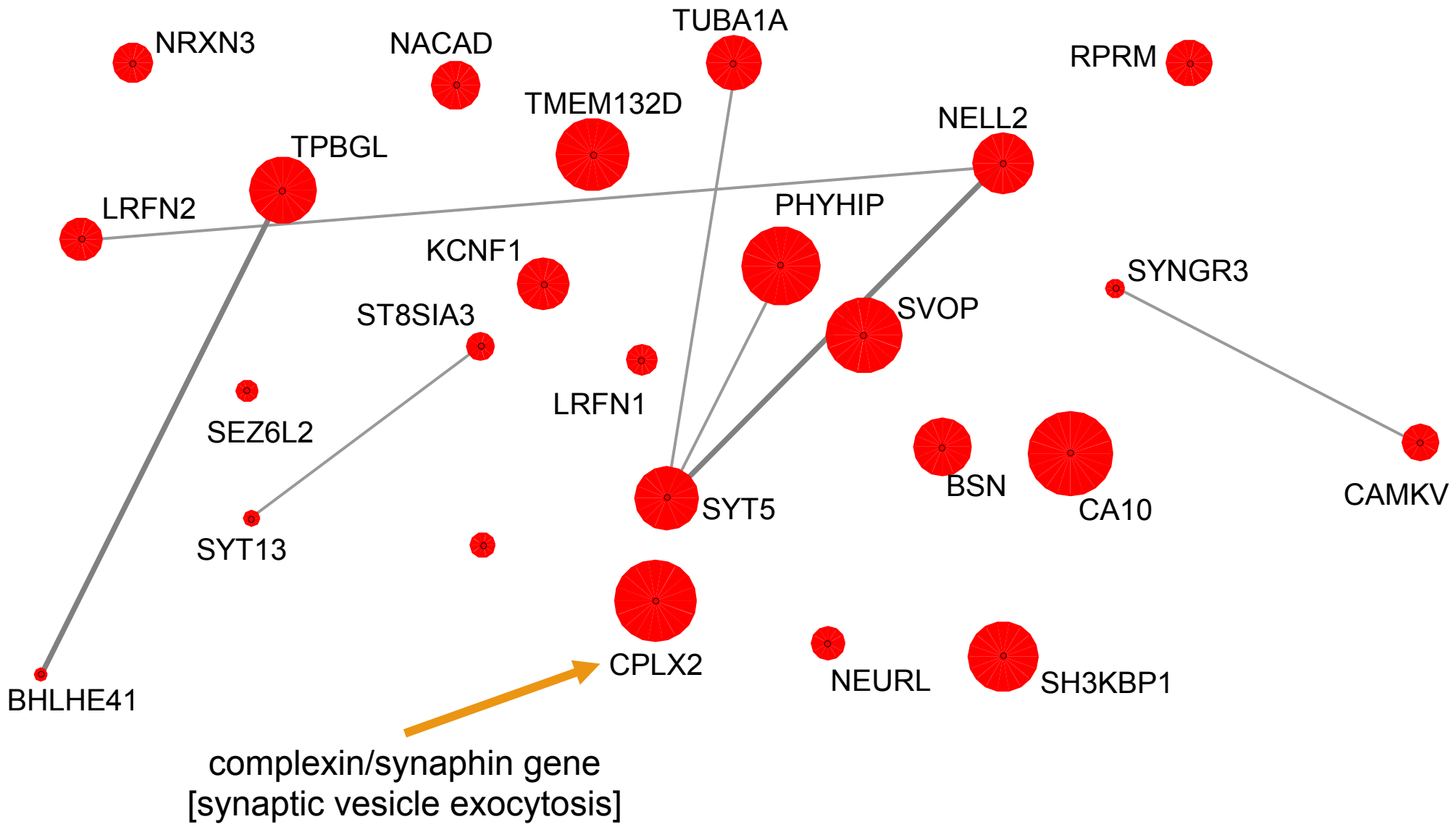

Control
90
10
60
30
0
330
300
270
240
210
180
150
120

| | GSEA gene set | # of genes* | Top ΔSSMD gene |
|---|---|---|---|
| **Metabolism and biosynthesis** | | | |
| | KEGG_PENTOSE_PHOSPHATE_PATHWAY | 19/27 | H6PD, PRPS2, PFKP |
| | KEGG_STEROID_BIOSYNTHESIS | 14/17 | SC5DL, NSDHL, DHCR7 |
| | REACTOME_CHOLESTEROL_BIOSYNTHESIS | 20/24 | SQLE, HSD17B7, HMGCR |
| | REACTOME_BRANCHED_CHAIN_AMINO_ACID_ CATABOLISM | 16/17 | DLD, HIBADH, MCCC2 |
| **Immune/Inflammatory response** | | | |
| | BIOCARTA_LAIR_PATHWAY | 4/17 | SELPLG, C3, ITGB1 |
| | BIOCARTA_41BB_PATHWAY | 12/17 | MAPK8, ATF2, MAPK14 |
| | REACTOME_IL1_SIGNALING | 25/39 | CHUK, RBX1, BTRC |
| | REACTOME_REGULATION_OF_IFNA_SIGNALING | 6/24 | STAT1, PTPN1, JAK1 |
| **Signaling pathway** | | | |
| | BIOCARTA_IGF1_PATHWAY | 20/21 | JUN, CSNK2A1, ELK1 |
| | PID_S1P_S1P2_PATHWAY | 21/24 | MAPK8, MAPK14, JUN |
| | PID_HNF3APATHWAY (FOXA1/HNF3A TF network) | 22/44 | NDUFV3, PISD, FOS |
| | REACTOME_ENERGY_DEPENDENT_REGULATION_ OF_MTOR_BY_LKB1_AMPK | 15/18 | PRKAA1, CAB39, TSC1 |
| **Vitamins and supplements** | | | |
| | BIOCARTA_VITCB_PATHWAY | 6/11 | SLC2A3, COL4A2, SLC2A1 |
| | REACTOME_TETRAHYDROBIOPTERIN_BH4_SYNTHESIS_ | 9/13 | GCHFR, PTS, AKT1 |

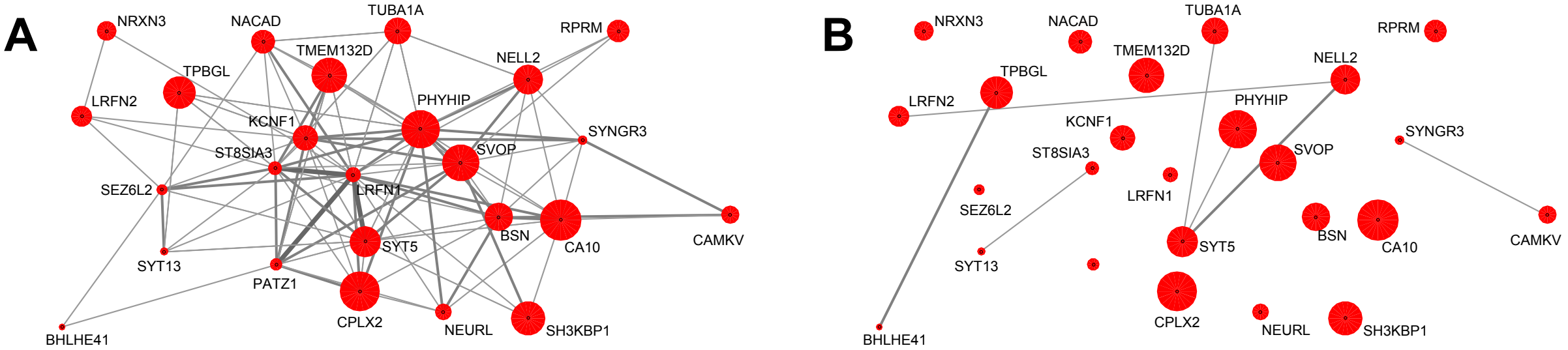
| | | | |
|---|---|---|---|
| | OF_MTOR_BY_LKB1_AMPK | | |
| **Vitamins and supplements** | | | |
| | BIOCARTA_VITCB_PATHWAY | 6/11 | SLC2A3, COL4A2, SLC2A1 |
| | REACTOME_TETRAHYDROBIOPTERIN_BH4_SYNTHESIS_ RECYCLING_SALVAGE_AND_REGULATION | 9/13 | GCHFR, PTS, AKT1 |
| **Miscellaneous** | | | |
| | REACTOME_ACTIVATED_POINT_MUTANTS_OF_FGFR2 | 4/16 | FGF9, FGFR2, FGF1 |
| | REACTOME_ACTIVATION_OF_THE_AP1_FAMILY_OF_ TRANSCRIPTION_FACTORS | 10/10 | MAPK14, MAPK3, ATF2 |
| | REACTOME_INWARDLY_RECTIFYING_K_CHANNELS | 20/31 | KCNJ10, KCNJ4, GNG4 |
| | REACTOME_G2_M_CHECKPOINTS | 22/45 | MCM2, RFC5, RPA2 |




Guan *et al.* (Cai) *Hum Genet* 2016

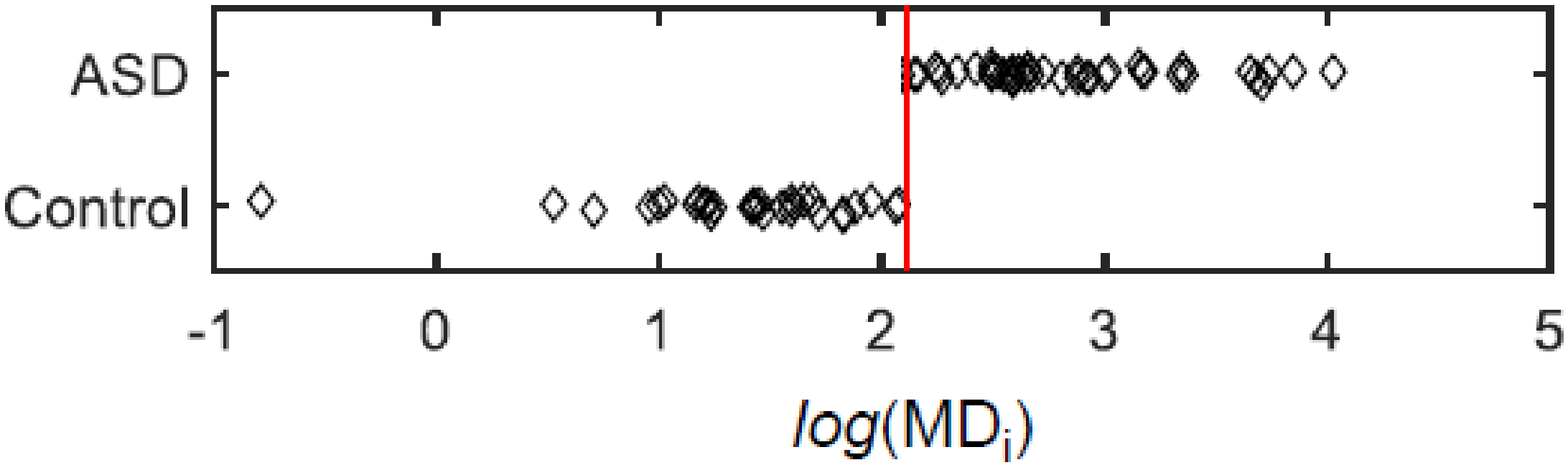



Guan *et al.* (Cai) *Hum Genet* 2016




Guan *et al.* (Cai) *Hum Genet* 2016

# Search for gene expression markers for **early** diagnosis



# Search for gene expression markers for **early** diagnosis

$(\blacksquare N@2\ )=1.0000494500$

$(\blacksquare N@3\ )=2.2177e+11$

$(\blacksquare N@4\ )=6.0971e+14$

$(\blacksquare N@5\ )=1.3409e+18$

# **Genetic Algorithm** Optimization

| Fitness | Initial Population |
|---|---|
| 22 | 1010101001111101 01 |
| 9 | 110011010101011100 |
| 8 | 111110101111010101 |
| 70 | 111001111100001001 |
| 19 | 110011010101011100 |
| 48 | 101110101111001001 |
| 23 | 110011010101011100 |
| 38 | 111001111100001001 |

Selection

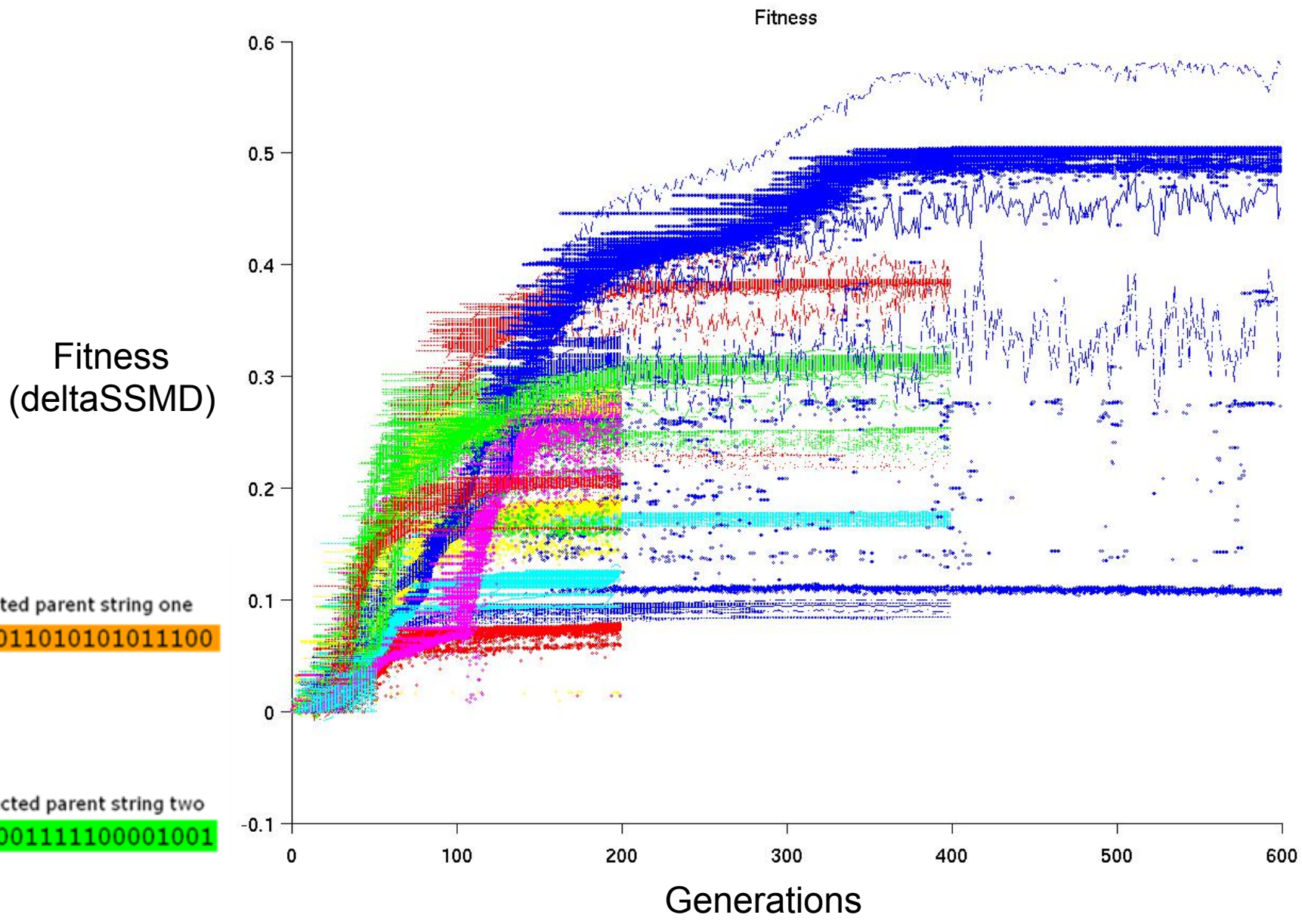
Selected parent string one

110011010101011100

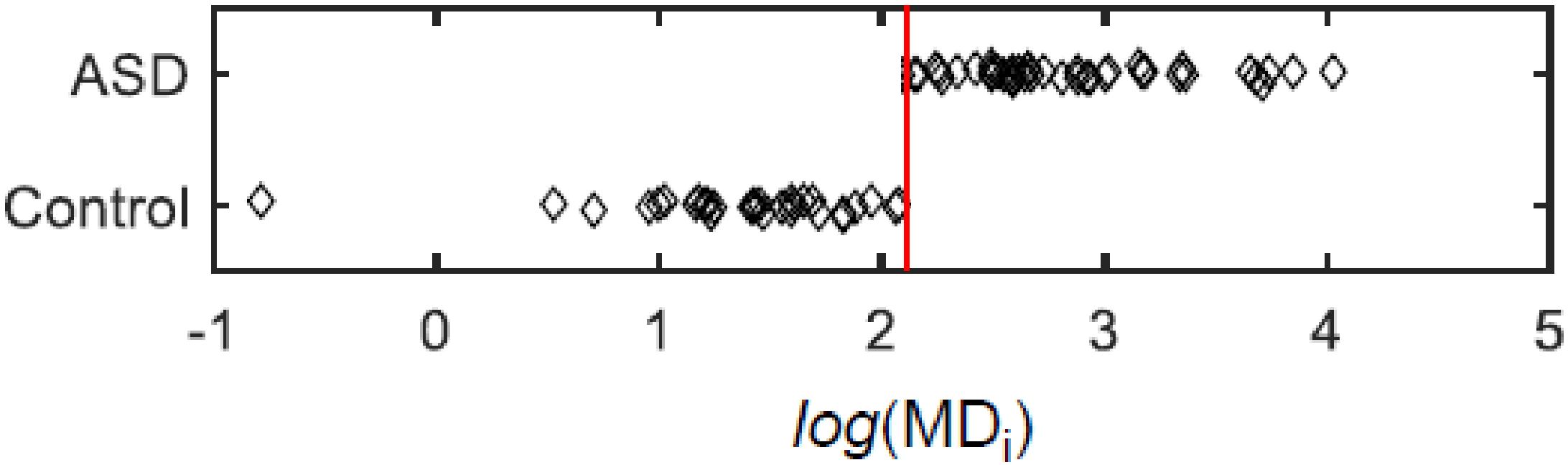
Selection

Selected parent string two

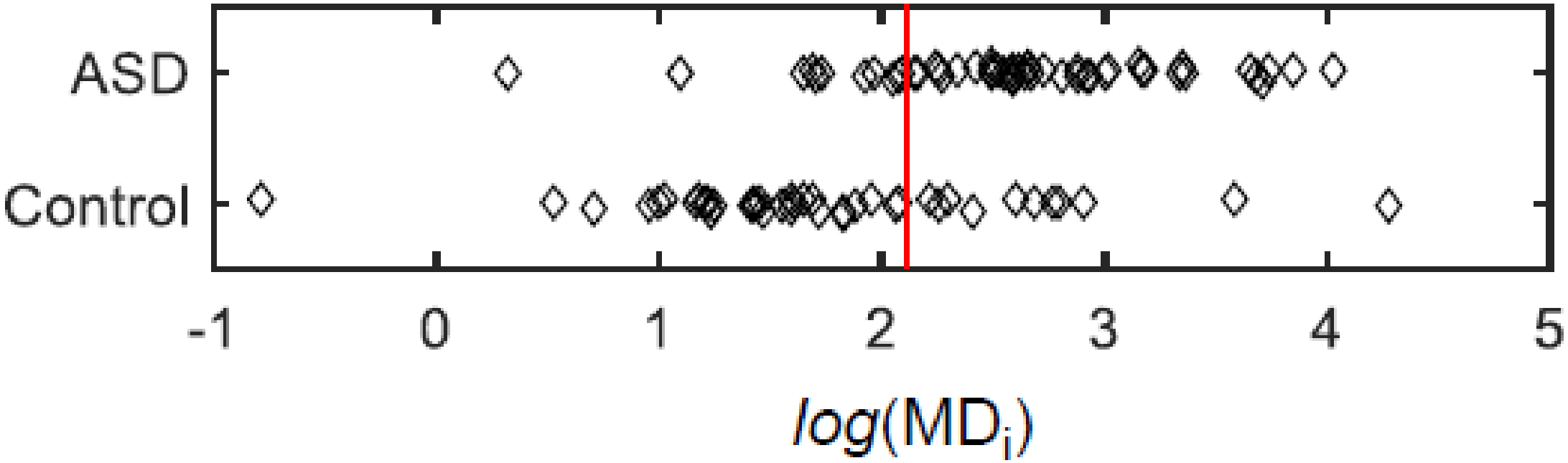
111001111100001001



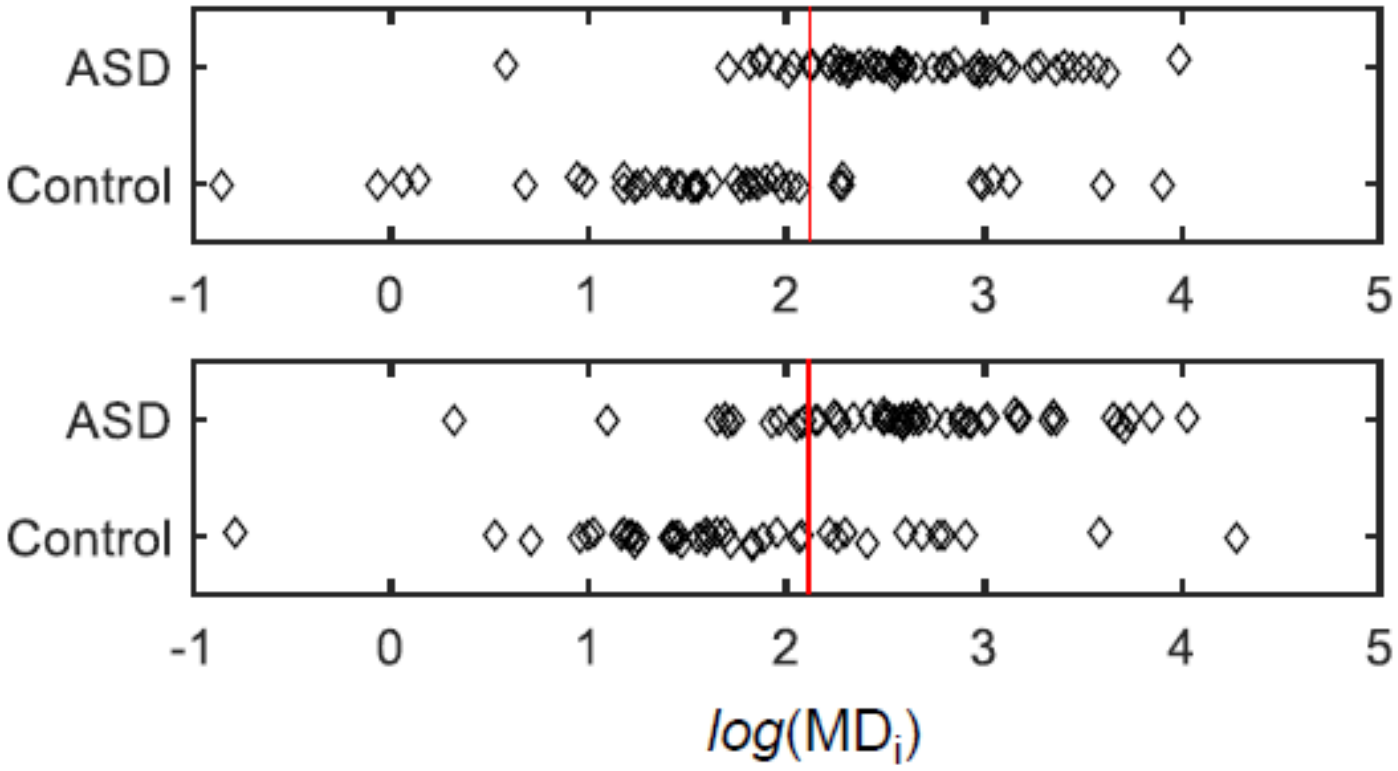
# Search for gene expression markers for **early** diagnosis



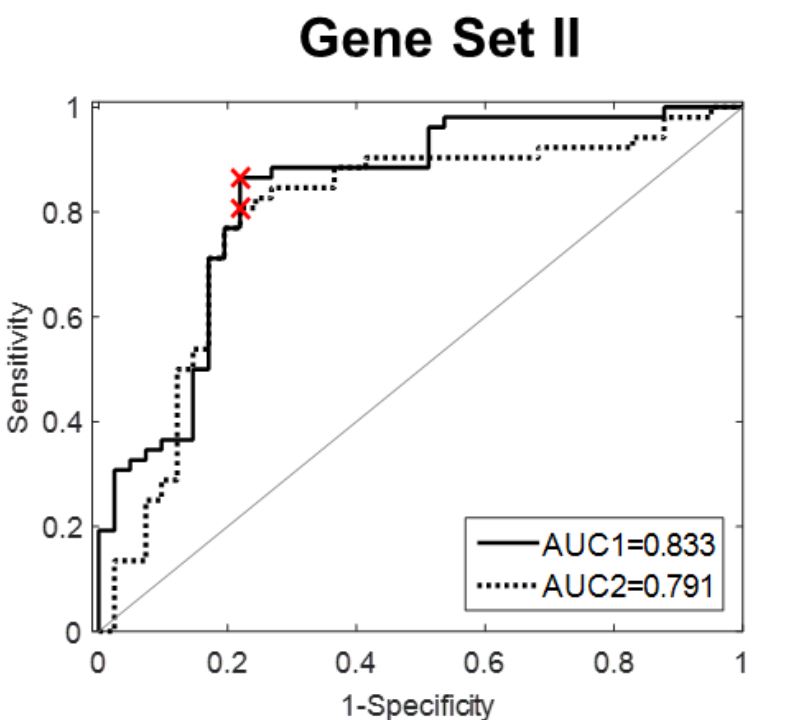
# {***EVI2B***, ***MYLIP***, *OR11G2*, *TSPAN16*, ***ZNF594***}



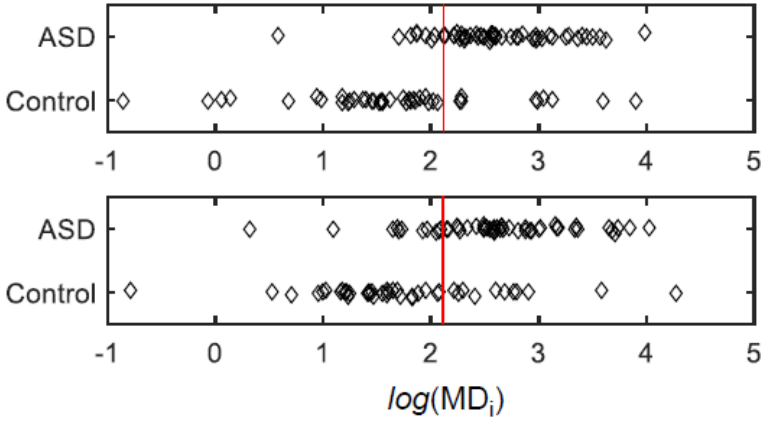
# {***EVI2B***, ***MYLIP***, *OR11G2*, *TSPAN16*, ***ZNF594***}



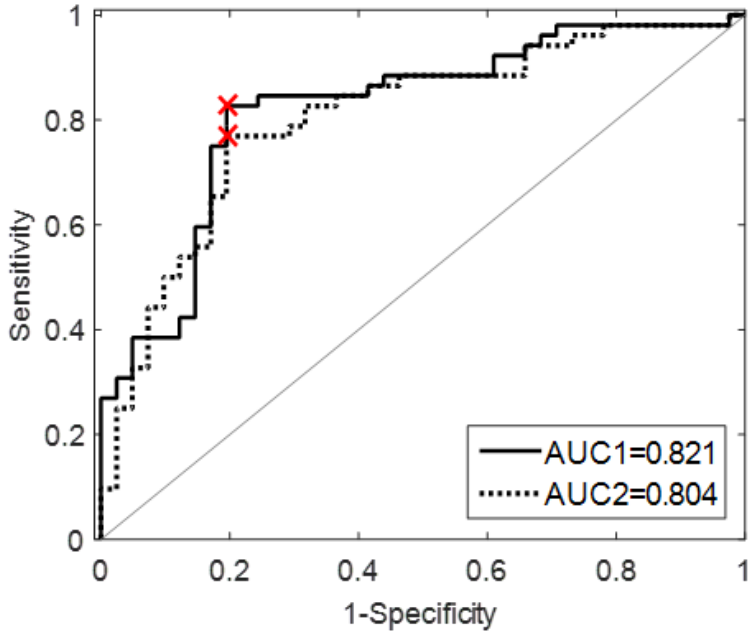
# Receiver Operating Characteristic (ROC) Curve




{**EVI2B**,**MYLIP**,OR11G2,TSPAN16,**ZNF594**}

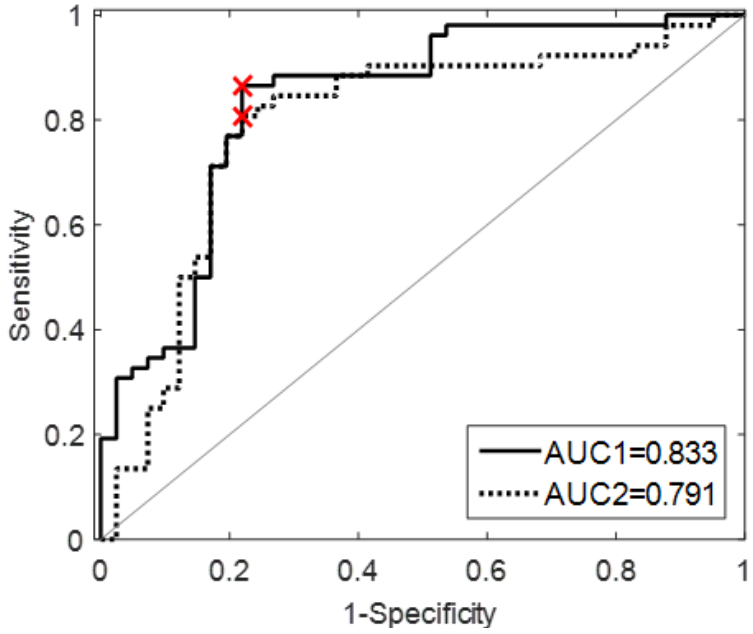


## Gene Set I



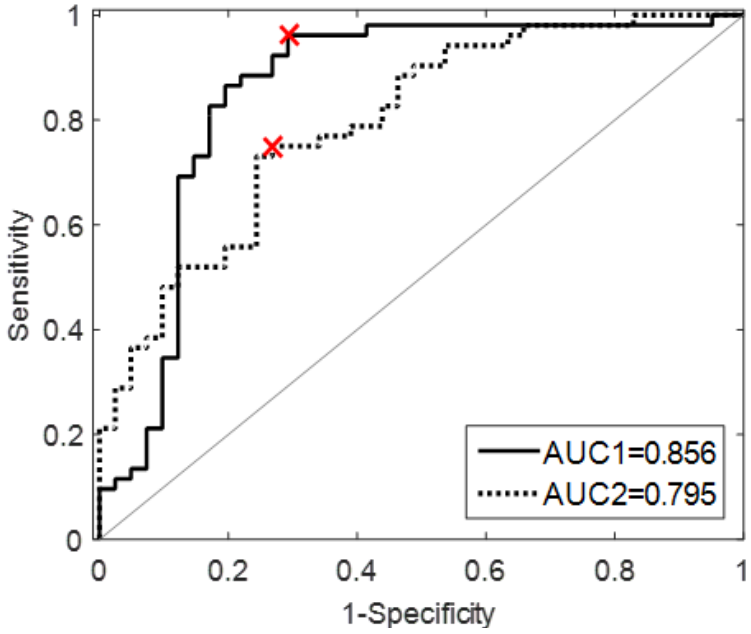

{FAM120A,HDC,**OR13C8**,PSAP,**RFX8**}

## Gene Set II



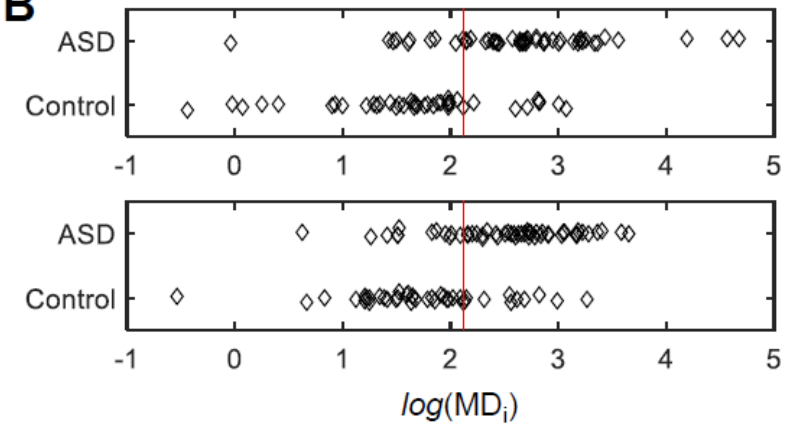

{**EVI2B**,**MYLIP**,OR11G2,TSPAN16,**ZNF594**}

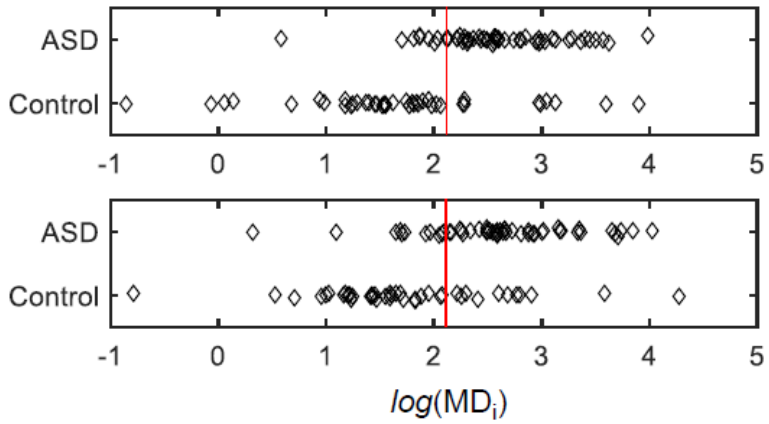
## Gene Set III

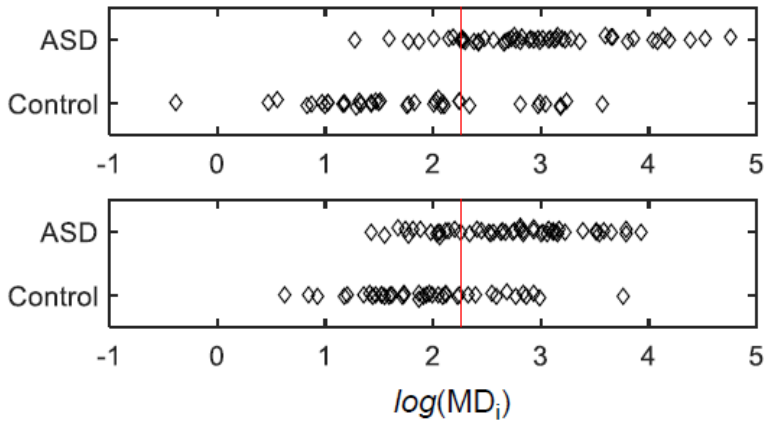



{**BCL11A**,**DST**,ORM2,RBM14,**SERAC1**}

**B**



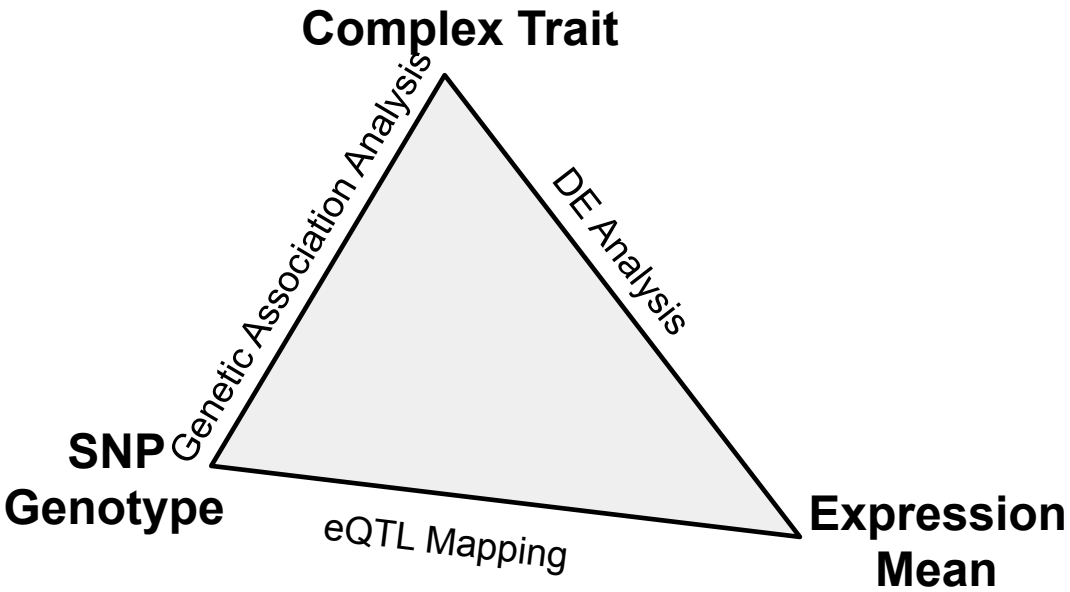



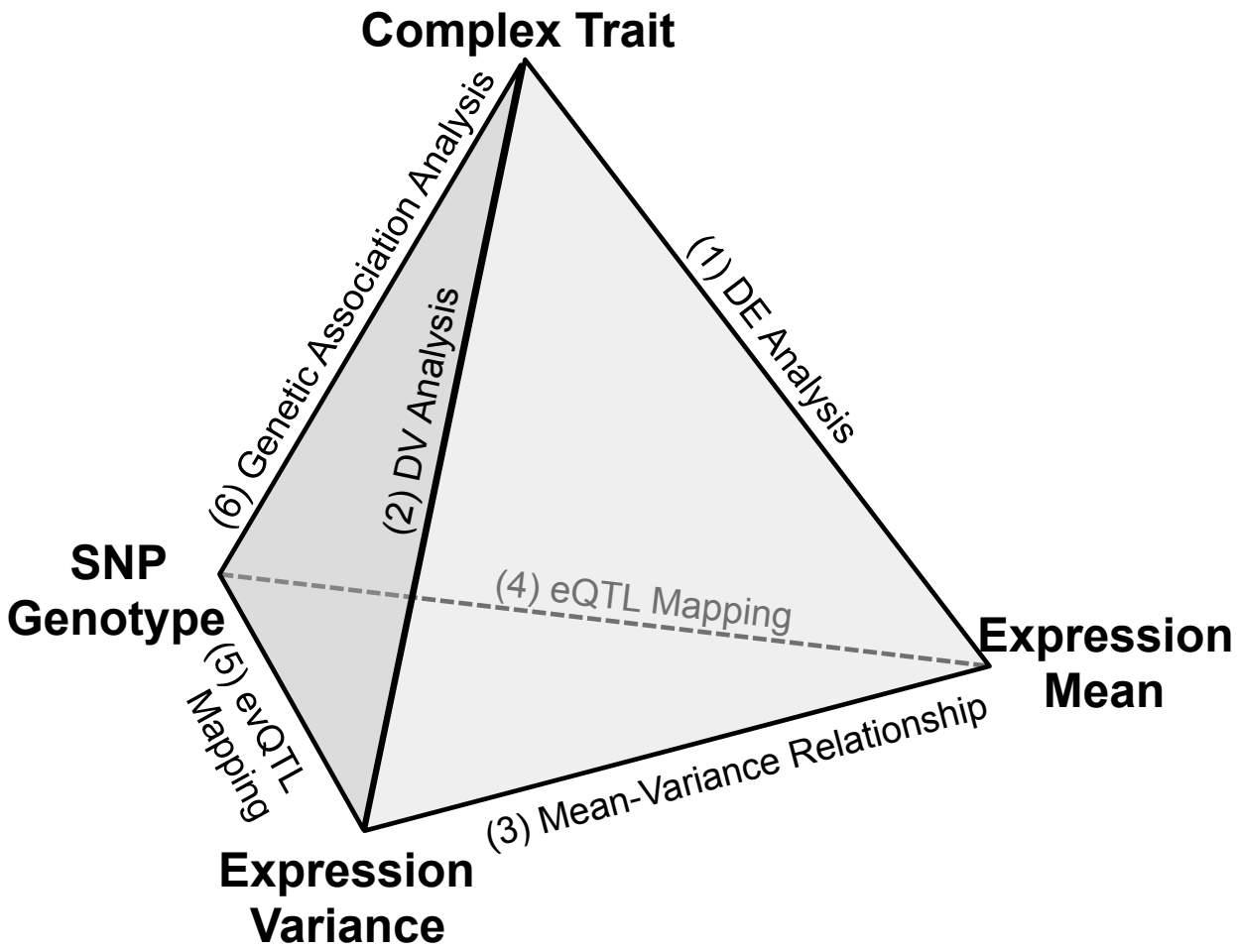
# Common dysregulated gene sets between **AUT**, **SCZ**, and **BPD**


Guan *et al.* (Cai) *Unpublished*

# Summary

- Detecting **aberrant gene expression** and identifying underlying genes and mutations represent a new discovery and diagnostic strategy for *genetically heterogeneous* disorders such as autism.


Complex Trait
Genetic Association Analysis
DE Analysis
SNP Genotype
eQTL Mapping
Expression Mean


Complex Trait
(6) Genetic Association Analysis
(2) DV Analysis
(1) DE Analysis
SNP Genotype
(4) eQTL Mapping
Expression Mean
(5) evQTL Mapping
(3) Mean-Variance Relationship
Expression Variance